BEYOND SIGNIFICANCE TESTING

STATISTICS REFORM IN THE BEHAVIORAL SCIENCES

SECOND EDITION

REX B. KLINE

BEYOND SIGNIFICANCE TESTING

BEYOND SIGNIFICANCE TESTING

STATISTICS REFORM IN THE BEHAVIORAL SCIENCES

SECOND EDITION

REX B. KLINE

American Psychological Association
Washington, DC

Published by
American Psychological Association
750 First Street, NE
Washington, DC 20002
www.apa.org

To order
APA Order Department
P.O. Box 92984
Washington, DC 20090-2984
Tel: (800) 374-2721; Direct: (202) 336-5510
Fax: (202) 336-5502; TDD/TTY: (202) 336-6123
Online: www.apa.org/books/
E-mail: order@apa.org

In the U.K., Europe, Africa, and the Middle East, copies may be ordered from
American Psychological Association
3 Henrietta Street
Covent Garden, London
WC2E 8LU England

Typeset in Goudy by Circle Graphics, Inc., Columbia, MD

Printer: United Book Press Inc., Baltimore, MD
Cover Designer: Naylor Design, Washington, DC

Library of Congress Cataloging-in-Publication Data

Kline, Rex B.
Beyond significance testing : statistics reform in the behavioral sciences / Rex B. Kline.
p. cm.
"Second edition"—Introduction.
Rev ed. of: Beyond significance testing : reforming data analysis methods in behavioral research. c2004.
Includes bibliographical references and index.
ISBN-13: 978-1-4338-1278-1
ISBN-10: 1-4338-1278-9
1. Psychometrics. I. Title.

BF39.K59 2013
150.72'4—dc23

2012035086

British Library Cataloguing-in-Publication Data

A CIP record is available from the British Library.

Printed in the United States of America
Second Edition

DOI: 10.1037/14136-000

For my family,
Joanna, Julia Anne, and Luke Christopher,
and
my brother,
Don Neil Justin Dwayne Foxworth (1961–2011),
fellow author

And so it is with us: we face change, much of it hard,
whether we like it or not.
But it is in the hard times especially that we grow,
that we become transformed.
—*Patrick Doyle*

CONTENTS

ACKNOWLEDGMENTS

It was a privilege to work once again with the APA Books staff, including Linda Malnasi McCarter, who helped to plan the project; Beth Hatch, who worked with the initial draft and offered helpful suggestions; Dan Brachtesende, who shepherded the book through the various production stages; Ron Teeter, who oversaw copyediting and organized the book's design; and Robin Easson, who copyedited the technically complex manuscript while helping to improve the presentation. Bruce Thompson reviewed the complete first draft and gave many helpful suggestions. Any remaining shortcomings in the presentation are solely my own. My loving family was again at my side the whole time. Thanks Joanna, Julia, and Luke.

BEYOND SIGNIFICANCE TESTING

INTRODUCTION

The goals of this second edition are basically the same as those of the original. This book introduces readers to the principles and practice of statistics reform in the behavioral sciences. It (a) reviews the now even larger literature about shortcomings of significance testing; (b) explains why these criticisms have sufficient merit to justify major changes in the ways researchers analyze their data and report the results; (c) helps readers acquire new skills concerning interval estimation and effect size estimation; and (d) reviews alternative ways to test hypotheses, including Bayesian estimation. I aim to change how readers think about data analysis, especially among those with traditional backgrounds in statistics where significance testing was presented as basically the only way to test hypotheses. I want all readers to know that there is a bigger picture concerning the analysis that blind reliance on significance testing misses.

I wrote this book for researchers and students in psychology and other behavioral sciences who do not have strong quantitative backgrounds. I

DOI: 10.1037/14136-011
Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, Second Edition, by R. B. Kline

assume that the reader has had undergraduate courses in statistics that covered at least the basics of regression and factorial analysis of variance. Each substantive chapter emphasizes fundamental statistical concepts but does not get into the minutiae of statistical theory. Works that do so are cited throughout the text, and readers can consult such works when they are ready. I emphasize instead both the sense and the nonsense of common data analysis practices while pointing out alternatives that I believe are more scientifically strong. I do not shield readers from complex topics, but I try to describe such topics using clear, accessible language backed up by numerous examples. This book is suitable as a textbook for an introductory course in behavioral science statistics at the graduate level. It can also be used in undergraduate-level courses for advanced students, such as honors program students, about modern methods of data analysis. Especially useful for all readers are Chapters 3 and 4, which respectively consider the logic and illogic of significance testing and misinterpretations about the outcomes of statistical tests. These misinterpretations are so widespread among researchers and students alike that one can argue that data analysis practices in the behavioral sciences are based more on myth than fact.

That the first edition of this book was so well reviewed and widely cited was very satisfying. I also had the chance to correspond with hundreds of readers from many different backgrounds where statistics reform is increasingly important. We share a common sense that the behavioral sciences should be doing better than they really are concerning the impact and relevance of research. Oh, yes, the research literature is very large, but quantity does not in this case indicate quality, and many of us know that most published studies in the behavioral studies have very little impact. Indeed, most publications are never cited again by authors other than those of the original works, and part of the problem has been our collective failure to modernize our methods of data analysis and describe our findings in ways relevant to target audiences.

New to this edition is coverage of robust statistical methods for parameter estimation, effect size estimation, and interval estimation. Most data sets in real studies do not respect the distributional assumptions of parametric statistical tests, so the use of robust statistics can lend a more realistic tenor to the analysis. Robust methods are described over three chapters (2, 3, and 5), but such methods do not remedy the major shortcomings of significance testing. There is a new chapter (3) about the logic and illogic of significance testing that deals with issues students rarely encounter in traditional statistics courses. There is expanded coverage of interval estimation in all chapters and also of Bayesian estimation as an increasingly viable alternative to traditional significance testing. Exercises are included for chapters that deal with fundamental topics (2–8). A new section in the last chapter summarizes best practice recommendations.

PLAN OF THE BOOK

Part I is concerned with fundamental concepts and summarizes the significance testing controversy. Outlined in Chapter 1 is the rationale of statistics reform. The history of the controversy about significance testing in psychology and other disciplines is recounted in this chapter. Principles of sampling and estimation that underlie confidence intervals and statistical tests are reviewed in Chapter 2. The logic and illogic of significance testing is considered in Chapter 3, and misunderstandings about *p* values are elaborated in Chapter 4. The purpose of Chapters 3–4 is to help you to understand critical weaknesses of statistical tests.

Part II comprises four chapters about effect size estimation in **comparative studies**, where at least two different groups or conditions are contrasted. In Chapter 5, the rationale of effect size estimation is outlined and basic effect sizes for continuous outcomes are introduced. The problem of evaluating substantive significance is also considered. Effect sizes for categorical outcomes, such as relapsed versus not relapsed, are covered in Chapter 6. Chapters 7 and 8 concern effect size estimation in, respectively, single-factor designs with at least three conditions and factorial designs with two or more factors and continuous outcomes. Many empirical examples are offered in Part II. There are exercises for Chapters 2–8 and suggested answers are available on the book's website.

Part III includes two chapters that cover alternatives to significance testing. Chapter 9 deals with replication and meta-analysis. The main points of this chapter are that a larger role for replication will require a cultural change in the behavioral sciences and that meta-analysis is an important tool for research synthesis but is no substitute for explicit replication. Bayesian estimation is the subject of Chapter 10. Bayesian statistics are overlooked in psychology research, but this approach offers an inference framework consistent with many goals of statistics reform. Best practice recommendations are also summarized in this chapter.

This book has a compendium website, where readers will find sample answers to the chapter exercises, downloadable raw data files for many research examples, and links to other useful websites. The URL for this book's website is http://forms.apa.org/books/supp/kline

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank

1

CHANGING TIMES

> It is simply that the things that appear to be permanent and dominant at any given moment in history can change with stunning rapidity. Eras come and go.
>
> —George Friedman (2009, p. 3)

This chapter explains the basic rationale of the movement for statistics reform in the behavioral sciences. It also identifies critical limitations of traditional significance testing that are elaborated throughout the book and reviews the controversy about significance testing in psychology and other disciplines. I argue that overreliance on significance testing as basically the sole way to evaluate hypotheses has damaged the research literature and impeded the development of psychology and other areas as empirical sciences. Alternatives are introduced that include using interval estimation of effect sizes, taking replication seriously, and focusing on the substantive significance of research results instead of just on whether or not they are statistically significant. Prospects for further reform of data analysis methods are also considered.

DOI: 10.1037/14136-001
Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, Second Edition, by R. B. Kline

Depending on your background, some of these points may seem shocking, even radical, but they are becoming part of mainstream thinking in many disciplines. **Statistics reform** is the effort to improve quantitative literacy in psychology and other behavioral sciences among students, researchers, and university faculty not formally trained in statistics (i.e., most of us). The basic aims are to help researchers better understand their own results, communicate more clearly about those findings, and improve the quality of published studies. Reform advocates challenge conventional wisdom and practices that impede these goals and emphasize more scientifically defensible alternatives.

Reformers also point out uncomfortable truths, one of which is that much of our thinking about data analysis is stuck in the 1940s (if not earlier). A sign of arrested development is our harmful overreliance on significance testing. Other symptoms include the failure to report effect sizes or consider whether results have scientific merit, both of which have nothing to do with statistical significance. In studies of intervention outcomes, a statistically significant difference between treated and untreated cases also has nothing to do with whether treatment leads to any tangible benefits in the real world. In the context of diagnostic criteria, **clinical significance** concerns whether treated cases can no longer be distinguished from control cases not meeting the same criteria. For example, does treatment typically prompt a return to normal levels of functioning? A treatment effect can be statistically significant yet trivial in terms of its clinical significance, and clinically meaningful results are not always statistically significant. Accordingly, the proper response to claims of statistical significance in any context should be "so what?"—or, more pointedly, "who cares?"—without more information.

Cognitive Errors

Another embarrassing truth is that so many cognitive errors are associated with significance testing that some authors describe a kind of **trained incapacity** that prevents researchers from understanding their own results; others describe a major educational failure (Hubbard & Armstrong, 2006; Ziliak & McCloskey, 2008). These misinterpretations are widespread among students, researchers, and university professors, some of whom teach statistics courses. So students learn false beliefs from people who should know better, but do not, in an ongoing cycle of misinformation. Ziliak and McCloskey (2008) put it this way:

> The textbooks are wrong. The teaching is wrong. The seminar you just attended is wrong. The most prestigious journal in your scientific field is wrong. (p. 250)

Most cognitive errors involve exaggerating what can be inferred from the outcomes of statistical tests, or *p* values (probabilities), listed in computer output. Common misunderstandings include the belief that *p* measures the likelihood that a result is due to sampling error (chance) or the probability that the null hypothesis is true. These and other false beliefs make researchers overconfident about their findings and excessively lax in some critical practices. One is the lip service paid to replication. Although I would wager that just as many behavioral scientists as their natural science colleagues would endorse replication as important, replication is given scant attention in the behavioral sciences. This woeful practice is supported by false beliefs.

Costs of Significance Testing

Summarized next are additional ways in which relying too much on significance testing has damaged our research literature. Nearly all published studies feature statistical significance, but studies without significant results are far less likely to be published or even submitted to journals (Kupfersmid & Fiala, 1991). This **publication bias for significance** suggests that the actual rate among published studies of Type I error, or incorrect rejection of the null hypothesis, is higher than indicated by conventional levels of statistical significance, such as .05. Ellis (2010) noted that because researchers find it difficult to get negative results published, Type I errors, once made, are hard to correct. Longford (2005) warned that the uncritical use of significance testing would lead to a "junkyard of unsubstantiated confidence," and Simmons, Nelson, and Simonsohn (2011) used the phrase "false-positive psychology" to describe the same problem.

Publication bias for significance also implies that the likelihood of Type II error, or failure to reject the null hypothesis when it is false in the population, is basically zero. In a less biased literature, though, information about the power, or the probability of finding statistical significance (rejecting the null hypothesis) when there is a real effect, would be more relevant. There are free computer tools for estimating power, but most researchers—probably at least 80% (e.g., Ellis, 2010)—ignore the power of their analyses. This is contrary to advice in the *Publication Manual* of the American Psychological Association (APA) that researchers should "routinely provide evidence that the study has sufficient power to detect effects of substantive interest" (APA, 2010, p. 30).

Ignoring power is regrettable because the median power of published nonexperimental studies is only about .50 (e.g., Maxwell, 2004). This implies a 50% chance of correctly rejecting the null hypothesis based on the data. In this case the researcher may as well not collect any data but instead just toss a coin to decide whether or not to reject the null hypothesis. This simpler,

cheaper method has the same chance of making correct decisions in the long run (F. L. Schmidt & Hunter, 1997).

A consequence of low power is that the research literature is often difficult to interpret. Specifically, if there is a real effect but power is only .50, about half the studies will yield statistically significant results and the rest will yield no statistically significant findings. If all these studies were somehow published, the number of positive and negative results would be roughly equal. In an old-fashioned, narrative review, the research literature would appear to be ambiguous, given this balance. It may be concluded that "more research is needed," but any new results will just reinforce the original ambiguity, if power remains low.

Confusing statistical significance with scientific relevance unwittingly legitimizes fad topics that clutter the literature but have low substantive value. With little thought about a broader rationale, one can collect data and then apply statistical tests. Even if the numbers are random, some of the results are expected to be statistically significant, especially in large samples. The objective appearance of significance testing can lend an air of credibility to studies with otherwise weak conceptual foundations. This is especially true in "soft" research areas where theories are neither convincingly supported nor discredited but simply fade away as researchers lose interest (Meehl, 1990). This lack of cumulativeness led Lykken (1991) to declare that psychology researchers mainly build castles in the sand.

Statistical tests of a treatment effect that is actually clinically significant may fail to reject the null hypothesis of no difference when power is low. If the researcher in this case ignored whether the observed effect size is clinically significant, a potentially beneficial treatment may be overlooked. This is exactly what was found by Freiman, Chalmers, Smith, and Kuebler (1978), who reviewed 71 randomized clinical trials of mainly heart- and cancer-related treatments with "negative" results (i.e., not statistically significant). They found that if the authors of 50 of the 71 trials had considered the power of their tests along with the observed effect sizes, those authors should have concluded just the opposite, or that the treatments resulted in clinically meaningful improvements.

If researchers become too preoccupied with statistical significance, they may lose sight of other, more important aspects of their data, such as whether the variables are properly defined and measured and whether the data respect test assumptions. There are clear problems in both of these areas. One is the **measurement crisis**, which refers to a substantial decline in the quality of instruction about measurement in psychology over the last 30 years or so. Psychometrics courses have disappeared from many psychology undergraduate programs, and about one third of psychology doctoral programs in North America offer no formal training in this area at all (Aiken et al., 1990;

Friederich, Buday, & Kerr, 2000). There is also evidence of widespread poor practices. For example, Vacha-Haase and Thompson (2011) found that about 55% of authors did not even mention score reliability in over 13,000 primary studies from a total of 47 meta-analyses of reliability generalization in the behavioral sciences. Authors mentioned reliability in about 16% of the studies, but they merely inducted values reported in other sources, such as test manuals, as if these applied to their data. Such **reliability induction** requires explicit justification, but researchers rarely compared characteristics of their samples with those from cited studies of score reliability.

A related problem is the **reporting crisis**, which refers to the fact that researchers infrequently present evidence that their data respect distributional or other assumptions of statistical tests (e.g., Keselman et al., 1998). The false belief that statistical tests are robust against violations of their assumptions in data sets of the type analyzed in actual studies may explain this flawed practice. Other aspects of the reporting crisis include the common failure to describe the nature and extent of missing data, steps taken to deal with the problem, and whether selection among alternatives could appreciably affect the results (e.g., Sterner, 2011). Readers of many journal articles are given little if any reassurance that the results are trustworthy.

Even if researchers avoided the kinds of mistakes just described, there are grounds to suspect that *p* values from statistical tests are simply incorrect in most studies:

1. They (*p* values) are estimated in theoretical sampling distributions that assume random sampling from known populations. Very few samples in behavioral research are random samples. Instead, most are convenience samples collected under conditions that have little resemblance to true random sampling. Lunneborg (2001) described this problem as a mismatch between design and analysis.
2. Results of more quantitative reviews suggest that, due to assumptions violations, there are few actual data sets in which significance testing gives accurate results (e.g., Lix, Keselman, & Keseleman, 1996). These observations suggest that *p* values listed in computer output are usually suspect. For example, this result for an independent samples *t* test calculated in SPSS looks impressively precise,

$$t(27) = 2.373,\ p = .025000184017821007$$

but its accuracy is dubious, given the issues just raised. If *p* values are generally wrong, so too are decisions based on them.

3. Probabilities from statistical tests (p values) generally assume that all other sources of error besides sampling error are nil. This includes measurement error; that is, it is assumed that $r_{XX} = 1.00$, where r_{XX} is a score reliability coefficient. Other sources of error arise from failure to control for extraneous sources of variance or from flawed operational definitions of hypothetical constructs. It is absurd to assume in most studies that there is no error variance besides sampling error. Instead it is more practical to expect that sampling error makes up the small part of all possible kinds of error when the number of cases is reasonably large (Ziliak & McCloskey, 2008).

The p values from statistical tests do not tell researchers what they want to know, which often concerns whether the data support a particular hypothesis. This is because p values merely estimate the conditional probability of the data under a *statistical* hypothesis—the null hypothesis—that in most studies is an implausible, straw man argument. In fact, p values do not directly "test" any hypothesis at all, but they are often misinterpreted as though they describe hypotheses instead of data.

Although p values ultimately provide a yes-or-no answer (i.e., reject or fail to reject the null hypothesis), the question—$p < \alpha$?, where α is the criterion level of statistical significance, usually .05 or .01—is typically uninteresting. The yes-or-no answer to this question says nothing about scientific relevance, clinical significance, or effect size. This is why Armstrong (2007) remarked that significance tests do not aid scientific progress even when they are properly done and interpreted.

New Statistics, New Thinking

Cumming (2012) recommended that researchers pay less attention to p values. Instead, researchers should be more concerned with sample results Cumming (2012) referred to as the **new statistics**. He acknowledged that the "new" statistics are not really new at all. What should be new instead is a greater role afforded them in describing the results. The new statistics consist mainly of effect sizes and confidence intervals. The *Publication Manual* is clear about effect size: "For the reader to appreciate the magnitude or importance of a study's findings, it is almost always necessary to include some measure of effect size" (APA, 2010, p. 34). The qualifier "almost always" refers to the possibility that, depending on the study, it may be difficult to compute effect sizes, such as when the scores are ranks or are presented in complex hierarchically structured designs. But it is possible to calculate effect sizes in most studies, and the effect size void for some kinds of designs is being filled by ongoing research.

Significance tests do not directly indicate effect size, and a common mistake is to answer the question $p < \alpha$? but fail to report and interpret effect sizes. Because effect sizes are sample statistics, or **point estimates**, that approximate population effect sizes, they are subject to sampling error. A confidence interval, or **interval estimate**, on a point estimate explicitly indicates the degree of sampling error associated with that statistic. Although sampling error is estimated in significance testing, that estimate winds up "hidden" in the calculation of *p*. But the amount of sampling error is made explicit by the lower and upper bounds of a confidence interval. Reporting confidence intervals reflects **estimation thinking** (Cumming, 2012), which deals with the questions "how much?" (point estimate) and "how precise?" (margin of error). The *Publication Manual* offers this advice: "Whenever possible, base discussion and interpretation of results on point and interval estimates" (APA, 2010, p. 34).

Estimation thinking is subsumed under **meta-analytic thinking**, which is fundamentally concerned with the accumulation of evidence over studies. Its basic aspects are listed next:

1. An accurate appreciation of the results of previous studies is seen as essential.
2. A researcher should view his or her own study as making a modest contribution to the literature. Hunter, Schmidt, and Jackson (1982) put it this way: "Scientists have known for centuries that a single study will not resolve a major issue. Indeed, a small sample study will not even resolve a minor issue" (p. 10).
3. A researcher should report results so that they can be easily incorporated into a future meta-analysis.
4. Retrospective interpretation of new results, once collected, is called for via direct comparison with previous effect sizes.

Thinking meta-analytically is incompatible with using statistical tests as the sole inference tool. This is because the typical meta-analysis estimates the central tendency and variability of effect sizes across sets of related primary studies. The focus on effect size and not statistical significance in individual studies also encourages readers of meta-analytic articles to think outside the limitations of the latter. There are statistical tests in meta-analysis, but the main focus is on whether a particular set of effect sizes is estimating the same population effect size and also on the magnitude and precision of mean effect sizes.

The new statistics cannot solve all that ails significance testing; no such alternative exists (see Cohen, 1994). For example, the probabilities associated with confidence intervals also assume that all other sources of imprecision besides sampling error are zero. There are ways to correct some effect sizes for measurement error, though, so this assumption is not always so strict. Abelson

(1997a) referred to the **law of the diffusion of idiocy**, which says that every foolish practice of significance testing will beget a corresponding misstep with confidence intervals. This law applies to effect sizes, too. But misinterpretation of the new statistics is less likely to occur if researchers can refrain from applying the same old, dichotomous thinking from significance testing. Thinking meta-analytically can also help to prevent misunderstanding.

You should know that measuring effect size in treatment outcome studies is insufficient to determine clinical significance, especially when outcomes have **arbitrary (uncalibrated) metrics** with no obvious connection to real-world status. An example is a 7-point Likert scale for an item on a self-report measure. This scale is arbitrary because its points could be represented with different sets of numbers, such as 1 through 7 versus −3 through 3 in whole-number increments, among other possibilities. The total score over a set of such items is arbitrary, too. It is generally unknown for arbitrary metrics (a) how a 1-point difference reflects the magnitude of change on the underlying construct and (b) exactly at what absolute points along the latent dimension observed scores fall. As Andersen (2007) noted, "Reporting effect sizes on arbitrary metrics alone with no reference to real-world behaviors, however, is no more meaningful or interpretable than reporting *p* values" (p. 669). So, determining clinical significance is not just a matter of statistics; it also requires strong knowledge about the subject matter.

These points highlight the idea that the evaluation of the clinical, practical, theoretical, or, more generally, **substantive significance** of observed effect sizes is a qualitative judgment. This judgment should be informed and open to scrutiny, but it will also reflect personal values and societal concerns. This is not unscientific because the assessment of all results in science involves judgment (Kirk, 1996). It is better to be open about this fact than to base decisions solely on "objective," mechanically applied statistical rituals that do not address substantive significance. Ritual is no substitute for critical thinking.

RETROSPECTIVE

Behavioral scientists did not always use statistical tests, so it helps to understand a little history behind the significance testing controversy; see Oakes (1986), Nickerson (2000), and Ziliak and McCloskey (2008) for more information.

Hybrid Logic of Statistical Tests (1920–1960)

Logical elements of significance testing were present in scientific papers as early as the 1700s (Stigler, 1986), but those basics were not organized into a

systematic method until the early 1900s. Today's significance testing is actually a hybrid of two schools of thought, one from the 1920s associated with Ronald Fisher (e.g., 1925) and another from the 1930s called the Neyman–Pearson approach, after Jerzy Neyman and Egon S. Pearson (e.g., 1933). Other individuals, such as William Gosset and Karl Pearson, contributed to these schools, but the work of the three principals listed first forms the genesis of significance testing (Ziliak & McCloskey, 2008, elaborate on Gosset's role).

Briefly, the Neyman–Pearson model is an extension of the Fisher model, which featured only a null hypothesis and estimation with statistical tests of the conditional probability of the data, or *p* values. There was no alternative hypothesis in Fisher's model. The conventional levels of statistical significance used today, .05 and .01, are correctly attributed to Fisher, but he did *not* advocate that they be blindly applied across all studies. Doing so, wrote Fisher (1956, p. 42), would be "absurdly academic" because no fixed level of significance could apply across all studies. This view is very different from today's practice, where $p < .05$ and $p < .01$ are treated as golden rules. For its focus on *p* values under the null hypothesis, Fisher's model has been called the ***p* value approach** (Huberty, 1993). The addition of the alternative hypothesis to the basic Fisher model, the attendant specification of one- or two-tailed regions of rejection, and the a priori specification of fixed levels of α across all studies characterize the Neyman–Pearson model, also called the **fixed α approach** (Huberty, 1993). This model also brought with it the conceptual framework of power and related decision errors, Type I and Type II.

To say that advocates of the Fisher model and the Neyman–Pearson model exchanged few kind words about each other's ideas is an understatement. Their long-running debate was acrimonious and included attempts by Fisher to block faculty appointments for Neyman. Nevertheless, the integration of the two models by other statisticians into what makes up contemporary significance testing took place roughly between 1935 and 1950. Gigerenzer (1993) referred to this integrated model as the **hybrid logic of scientific inference**, and Dixon and O'Reilly (1999) called it the **Intro Stats method**. Many authors have noted that (a) this hybrid model would have been rejected by Fisher, Neyman, and Pearson, although for different reasons, and (b) its composite nature is a source of confusion among students and researchers.

Rise of the Intro Stats Method, Testimation, and Sizeless Science (1940–1960)

Before 1940, statistical tests were rarely used in psychology research. Authors of works from the time instead applied in nonstandard ways a variety of descriptive statistics or rudimentary test statistics, such as the **critical ratio** of a sample statistic over its standard error (now called *z* or *t* when assuming

normality). An older term for the standard error—actually two times the square root of the standard error—is the **modulus**, described in 1885 by the economist Francis Ysidro Edgeworth (Stigler, 1978) to whom the term *statistical significance* is attributed. From about 1940–1960, during what Gigerenzer and Murray (1987) called the **inference revolution**, the Intro Stats method was widely adopted in psychology textbooks and journal editorial practice as *the* method to test hypotheses. The move away from the study of single cases (e.g., operant conditioning studies) to the study of groups over roughly 1920–1950 contributed to this shift. Another factor is what Gigerenzer (1993) called the **probabilistic revolution**, which introduced indeterminism as a major theoretical concept in areas such as quantum mechanics in order to better understand the subject matter. In psychology, though, it was used to mechanize the inference process, a critical difference, as it turns out.

After the widespread adoption of the Intro Stats method, there was an increase in the reporting of statistical tests in journal articles in psychology. This trend is obvious in Figure 1.1, reproduced from Hubbard and Ryan (2000). They sampled about 8,000 articles published during 1911–1998 in randomly selected issues of 12 different APA journals. Summarized in the figure are percentages of articles in which results of statistical tests were reported.

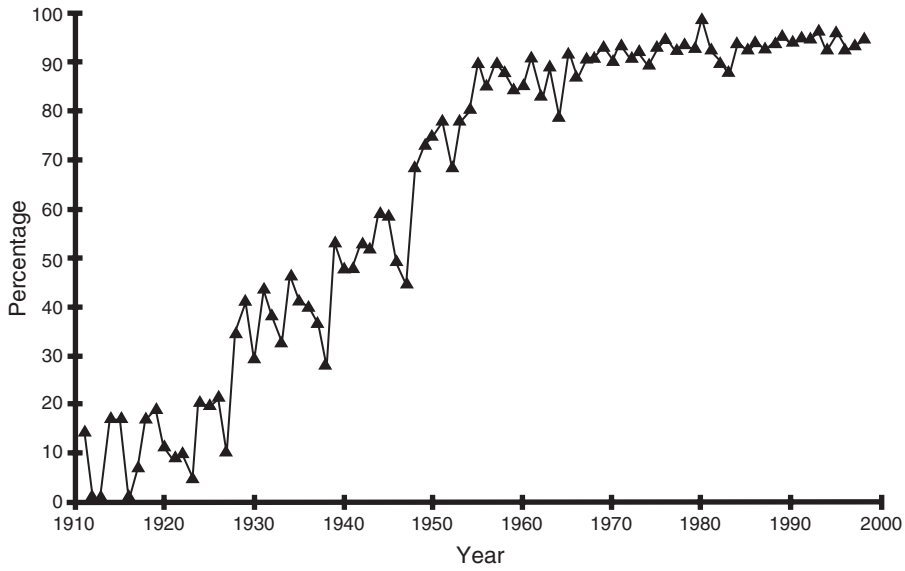

Figure 1.1. Percentage of articles reporting results of statistical tests in 12 journals of the American Psychological Association from 1911 to 1988. From "The Historical Growth of Statistical Significance Testing in Psychology—And Its Future Prospects," by R. Hubbard and P. A. Ryan, 2000, *Educational and Psychological Measurement, 60,* p. 665. Copyright 2001 by Sage Publications. Reprinted with permission.

This percentage is about 17% from 1911 to 1929. It increases to around 50% in 1940, continues to rise to about 85% by 1960, and has exceeded 90% since the 1970s. The time period 1940–1960 corresponds to the inference revolution.

Although the 1990s is the most recent decade represented in Figure 1.1, there is no doubt about the continuing, near-universal reporting of statistical tests in journals. Hoekstra, Finch, Kiers, and Johnson (2006) examined a total of 266 articles published in *Psychonomic Bulletin & Review* during 2002–2004. Results of significance tests were reported in about 97% of the articles, but confidence intervals were reported in only about 6%. Sadly, *p* values were misinterpreted in about 60% of surveyed articles. Fidler, Burgman, Cumming, Buttrose, and Thomason (2006) sampled 200 articles published in two different biology journals. Results of significance testing were reported in 92% of articles published during 2001–2002, but this rate dropped to 78% in 2005. There were also corresponding increases in the reporting of confidence intervals, but power was estimated in only 8% and *p* values were misinterpreted in 63%.

Some advantages to the institutionalization of the Intro Stats method were noted by Gigerenzer (1993). Journal editors could use significance test outcomes to decide which studies to publish or reject, respectively, those with or without statistically significant results, among other considerations. The method of significance testing is mechanically applied and thus seems to eliminate subjective judgment. That this objectivity is illusory is another matter. Significance testing gave researchers a common language and perhaps identity as members of the same grand research enterprise. It also distinguished them from their natural science colleagues, who may use statistical tests to detect outliers but not typically to test hypotheses (Gigerenzer, 1993).

The combination of significance testing and a related cognitive error is **testimation** (Ziliak & McCloskey, 2008). It involves exclusive focus on the question $p < \alpha$? If the answer is "yes," the results are automatically taken to be scientifically relevant, but issues of effect size and precision are ignored. Testimators also commit the **inverse probability error** (Cohen, 1994) by falsely believing that *p* values indicate the probability that the null hypothesis is true. Under this fallacy, the result $p = .025$, for example, is taken to mean that there is only a 2.5% chance that the null hypothesis is true. A researcher who mistakenly believes that low *p* values make the null hypothesis unlikely may become overly confident in the results.

Presented next is hypothetical text that illustrates the language of testimation:

> A 2 × 2 × 2 (Instructions × Incentive × Goals) factorial ANOVA was conducted with the number of correct items as the dependent variable. The 3-way interaction was significant, $F(1, 72) = 5.20$, $p < .05$, as were all 2-way

> interactions, Instructions × Incentive, $F(1, 72) = 11.95$, $p < .001$; Instructions × Goals, $F(1, 72) = 25.40$, $p < .01$; Incentive × Goals, $F(1, 72) = 9.25$, $p < .01$, and two of three of the main effects, Instructions, $F(1, 72) = 11.60$, $p < .01$; Goals, $F(1, 72) = 6.25$, $p < .05$.

This text chockablock with numbers—which is poor writing style—says nothing about the magnitudes of all those "significant" effects. If later in the hypothetical article the reader is still given no information about effect sizes, that is **sizeless science**. Getting excited about "significant" results while knowing nothing about their magnitudes is like ordering from a restaurant menu with no prices: You may get a surprise (good or bad) when the bill (statement of effect size) comes.

Increasing Criticism of Statistical Tests (1940–Present)

There has been controversy about statistical tests for more than 80 years, or as long as they have been around. Boring (1919), Berkson (1942), and Rozeboom (1960) are among earlier works critical of significance testing. Numbers of published articles critical of significance testing have increased exponentially since the 1940s. For example, Anderson, Burnham, and Thompson (2000) found less than 100 such works published during the 1940s–1970s in ecology, medicine, business, economics, or the behavioral sciences, but about 200 critical articles were published in the 1990s. W. L. Thompson (2001) listed a total of 401 references for works critical of significance testing, and Ziliak and McCloskey (2008, pp. 57–58) cited 125 such works in psychology, education, business, epidemiology, and medicine, among other areas.

Proposals to Ban Significance Testing (1990s–Present)

The significance testing controversy escalated to the point where, by the 1990s, some authors called for a ban in research journals. A ban was discussed in special sections or issues of *Journal of Experimental Education* (B. Thompson, 1993), *Psychological Science* (Shrout, 1997), and *Research in the Schools* (McLean & Kaufman, 1998) and in an edited book by Harlow, Mulaik, and Steiger (1997), the title of which asks "What if there were no significance tests?" Armstrong (2007) offered this more recent advice:

> When writing for books and research reports, researchers should omit mention of tests of statistical significance. When writing for journals, researchers should seek ways to reduce the potential harm of reporting significance tests. They should also omit the word significance because findings that reject the null hypothesis are not significant in the everyday use of the term, and those that [fail to] reject it are not insignificant. (p. 326)

In 1996, the Board of Scientific Affairs of the APA convened the Task Force on Statistical Inference (TFSI) to respond to the ongoing significance testing controversy and elucidate alternatives. The report of the TFSI (Wilkinson & the TFSI, 1999) dealt with many issues and offered suggestions for the then-upcoming fifth edition of the *Publication Manual:*

1. Use minimally sufficient analyses (simpler is better).
2. Do not report results from computer output without knowing what they mean. This includes *p* values from statistical tests.
3. Document assumptions about population effect sizes, sample sizes, or measurement behind a priori estimates of statistical power. Use confidence intervals about observed results instead of estimating observed (post hoc) power.
4. Report effect sizes and confidence intervals for primary outcomes or whenever *p* values are reported.
5. Give assurances to a reasonable degree that the data meet statistical assumptions.

The TFSI decided in the end not to recommend a ban on statistical tests. In its view, such a ban would be a too extreme way to curb abuses.

Fifth and Sixth Editions of the APA's *Publication Manual* (2001–2010)

The fifth edition of the *Publication Manual* (APA, 2001) took a stand similar to that of the TFSI regarding significance testing. That is, it acknowledged the controversy about statistical tests but stated that resolving this issue was not a proper role of the *Publication Manual*. The fifth edition went on to recommend the following:

1. Report adequate descriptive statistics, such as means, variances, and sizes of each group and a pooled within-groups variance–covariance matrix in a comparative study. This information is necessary for later meta-analyses or secondary analyses by others.
2. Effect sizes should "almost always" be reported, and the absence of effect sizes was cited as an example of a study defect.
3. The use of confidence intervals was "strongly recommended" but not required.

The sixth edition of the *Publication Manual* (APA, 2010) used similar language when recommending the reporting of effect sizes and confidence intervals. Predictably, not everyone is happy with the report of the TFSI or the wording of the *Publication Manual*. B. Thompson (1999) noted that only encouraging the reporting of effect sizes or confidence intervals presents a self-canceling mixed message. Ziliak and McCloskey (2008, p. 125) chastised

the *Publication Manual* for "retaining the magical incantations of $p < .05$ and $p < .01$." S. Finch, Cumming, and Thomason (2001) contrasted the recommendations about statistical analyses in the *Publication Manual* with the more straightforward guidelines in the *Uniform Requirements for Manuscripts Submitted to Biomedical Journals*, recently revised (International Committee of Medical Journal Editors, 2010). Kirk (2001) urged that the then-future sixth edition of the *Publication Manual* should give more detail than the fifth edition about the TFSI's recommendations. Alas, the sixth edition does not contain such information, but I aim to provide you with specific skills of this type as you read this book.

Reform-Oriented Editorial Policies and Mixed Evidence of Progress (1980s–Present)

Journal editorials and reviewers are the gatekeepers of the research literature, so editorial policies can affect the quality of what is published. Described next are three examples of efforts to change policies in reform-oriented directions with evaluations of their impact; see Fidler, Thomason, Cumming, Finch, and Leeman (2004) and Fidler et al. (2005) for more examples.

Kenneth J. Rothman was the assistant editor of the *American Journal of Public Health* (AJPH) from 1984 to 1987. In his revise-and-submit letters, Rothman urged authors to remove from their manuscripts all references to *p* values (e.g., Fidler et al., 2004, p. 120). He founded the journal *Epidemiology* in 1990 and served as its first editor until 2000. Rothman's (1998) editorial letter to potential authors was frank:

> When writing for *Epidemiology*, you can . . . enhance your prospects if you omit tests of statistical significance. . . . In *Epidemiology*, we do not publish them at all. . . . We discourage the use of this type of thinking in the data analysis. . . . We also would like to see the interpretation of a study based not on statistical significance, or lack of it . . . but rather on careful quantitative consideration of the data in light of competing explanations for the findings. (p. 334)

Fidler et al. (2004) examined 594 AJPH articles published from 1982 to 2000 and 100 articles published in *Epidemiology* between 1990 and 2000. Reporting based solely on statistical significance dropped from about 63% of the AJPH articles in 1982 to about 5% of articles in 1986–1989. But in many AJPH articles there was evidence that interpretation was based mainly on undisclosed significance test results. The percentages of articles in which confidence intervals were reported increased from about 10% to 54% over the same period. But these changes in reporting practices in AJPH articles did not generally persist past Rothman's tenure.

From 1993 to 1997, Geoffrey R. Loftus was the editor of *Memory & Cognition*. Loftus (1993) gave these guidelines to potential contributors:

> I intend to try to decrease the overwhelming reliance on hypothesis testing as the major means of transiting from data to conclusions. . . . In lieu of hypothesis testing, I will emphasize the increased use of figures depicting sample means along with standard error bars. . . . More often than not, inspection of such a figure will immediately obviate the necessity of any hypothesis testing procedures. In such situations, presentation of the usual hypothesis-testing information (*F* values, *p* values, etc.) will be discouraged. I believe . . . that . . . an overreliance on the impoverished binary conclusions yielded by the hypothesis-testing procedure has subtly seduced our discipline into insidious conceptual cul-de-sacs that have impeded our vision and stymied our potential. (p. 3)

Loftus apparently encountered considerable resistance, if not outright obstinacy, on the part of some authors. For example, Loftus calculated confidence intervals for about 100 authors who failed or even refused to do so on their own. In contrast, Rothman reported little resistance from authors who submitted works to *Epidemiology* (see Fidler et al., 2004, p. 124). S. Finch et al. (2004) examined a total of 696 articles published in *Memory & Cognition* before, during, and after Loftus's editorship. The rate of reporting of confidence intervals increased from 7% from before Loftus's tenure to 41%, but the rate dropped to 24% just after Loftus departed. But these confidence intervals were seldom interpreted; instead, authors relied mainly on statistical test outcomes to describe the results.

Another expression of statistics reform in editorial policy are the requirements of about 24 journals in psychology, education, counseling, and other areas for authors to report effect sizes.[1] Some of these are flagship journals of associations (e.g., American Counseling Association, Council for Exceptional Children), each with about 40,000–45,000 members. Included among journals that require effect sizes are three APA journals, *Health Psychology*, *Journal of Educational Psychology*, and *Journal of Experimental Psychology: Applied*. The requirement to report effect sizes sends a strong message to potential contributors that use of significance testing alone is not acceptable.

Early suggestions to report effect sizes fell mainly on deaf ears. S. Finch et al. (2001) found little evidence for effect size estimation or interval estimation in articles published in *Journal of Applied Psychology* over the 40-year period from 1940 to 1999. Vacha-Haase and Ness (1999) found the rate of effect size reporting was about 25% in *Professional Psychology: Research and Practice*, but authors did not always interpret the effect sizes they reported. Results from more recent surveys are better. Dunleavy, Barr, Glenn, and

[1] http://people.cehd.tamu.edu/~bthompson/index.htm, scroll down to hyperlinks.

Miller (2006) reviewed 736 articles published over 2002–2005 in five different applied, experimental, or personnel psychology journals. The overall rate of effect size reporting was about 62.5%. Among studies where no effect sizes were reported, use of the techniques of analysis of variance (ANOVA) and the *t* test were prevalent. Later I will show you that effect sizes are actually easy to calculate in such analyses, so there is no excuse for not reporting them. Andersen (2007) found that in a total of 54 articles published in 2005 in three different sport psychology journals, effect sizes were reported in 44 articles, or 81%. But the authors of only seven of these articles interpreted effect sizes in terms of substantive significance. Sun, Pan, and Wang (2010) reviewed a total of 1,243 works published in 14 different psychology and education journals during 2005–2007. The percentage of articles reporting effect sizes was 49%, and 57% of these authors interpreted their effect sizes.

Evidence for progress in statistics reform is thus mixed. Researchers seem to report effect sizes more often, but improvement in reporting confidence intervals may lag behind. Too many authors do not interpret the effect sizes they report, which avoids dealing with the question of why does an effect of this size matter. It is poor practice to compute effect sizes only for statistically significant results. Doing so amounts to business as usual where the significance test is still at center stage (Sohn, 2000). Real reform means that effect sizes are interpreted for their substantive significance, not just reported.

OBSTACLES TO REFORM

There are two great obstacles to continued reform. The first is inertia: It is human nature to resist change, and it is hard to give up familiar routines. Belasco and Stayer (1993) put it like this: "Most of us overestimate the value of what we currently have, and have to give up, and underestimate the value of what we may gain" (p. 312). But science demands that researchers train the lens of skepticism on their own assumptions and methods. Such self-criticism and intellectual honesty do not come easy, and not all researchers are up for the task. Defense attorney Gerry Spence (1995) wrote, "I would rather have a mind opened by wonder than one closed by belief" (p. 98). This conviction identifies a scientist's special burden.

The other big obstacle is vested interest, which is in part economic. I am speaking mainly about applying for research grants. Most of us know that grant monies are allocated in part on the assurance of statistical significance. Many of us also know how to play the **significance game**, which goes like this: Write application. Promise significance. Get money. Collect data until significance is found, which is virtually guaranteed because any effect that is not zero needs only a large enough sample in order to be significant.

Report results but mistakenly confuse statistical significance with scientific relevance. Sound trumpets about our awesomeness, move on to a different kind of study (do not replicate). Ziliak and McCloskey (2008) were even more candid:

> Significance unfortunately is a useful means toward personal ends in the advance of science—status and widely distributed publications, a big laboratory, a staff of research assistants, a reduction in teaching load, a better salary, the finer wines of Bordeaux. Precision, knowledge, and control. In a narrow and cynical sense statistical significance is the way to achieve these. Design experiment. Then calculate statistical significance. Publish articles showing "significant" results. Enjoy promotion. But it is not science, and it will not last. (p. 32)

Maybe I am a naive optimist, but I believe there is enough talent and commitment to improving research practices among too many behavioral scientists to worry about unheeded calls for reform. But such changes do not happen overnight. Recall that it took about 20 years for researchers to widely use statistical tests (see Figure 1.1), and sometimes shifts in scientific mentality await generational change. Younger researchers may be less set in their ways than the older generation and thus more open to change. But some journal editors—who are typically accomplished and experienced researchers—are taking the lead in reform. So are the authors of many of the works cited throughout this book.

Students are promising prospects for reform because they are, in my experience and that of others (Hyde, 2001), eager to learn about the significance testing controversy. They can also understand ideas such as effect size and interval estimation even in introductory courses. In fact, I find it is easier to teach undergraduates these concepts than the convoluted logic of significance testing. Other reform basics are even easier to convey (e.g., replicate—do not just talk about it.)

PROSPECTIVE

I have no crystal ball, but I believe that I can reasonably speculate about three anticipated developments in light of the events just described:

1. The role of significance testing will continue to get smaller and smaller to the point where researchers must defend its use. This justification should involve explanation of why the narrow assumptions about sampling and score characteristics in significance testing are not unreasonable in a particular study. Estimation of a priori power will also be required whenever statistical

tests are used. I and others (e.g., Kirk, 1996) envision that the behavioral sciences will become more like the natural sciences. That is, we will report the directions, magnitudes, and precisions of our effects; determine whether they replicate; and evaluate them for their substantive significance, not simply their statistical significance.

2. I expect that the best behavioral science journals will require evidence for replication. This requirement would send the strong message that replication is standard procedure. It would also reduce the number of published studies, which may actually improve quality by reducing noise (one-shot studies, unsubstantiated claims) while boosting signal (replicated results).
3. I concur with Rodgers (2010) that a "quiet methodological revolution" is happening that is also part of statistics reform. This revolution concerns the shift from testing individual hypotheses for statistical significance to the evaluation of entire mathematical and statistical models. There is a limited role for significance tests in statistical modeling techniques such as structural equation modeling (e.g., Kline, 2010, Chapter 8), but it requires that researchers avoid making the kinds of decision errors often associated with such tests.

CONCLUSION

Basic tenets of statistics reform emphasize the need to (a) decrease the role of significance testing and thus also reduce the damaging impact of related cognitive distortions; (b) shift attention to other kinds of statistics, such as effect sizes and confidence intervals; (c) reestablish the role of informed judgment and downplay mere statistical rituals; and (d) elevate replication. The context for reform goes back many decades, and the significance testing controversy has now spread across many disciplines. Progress toward reform has been slow, but the events just summarized indicate that continued use of significance testing as the only way to evaluate hypotheses is unlikely. The points raised set the stage for review in the next chapter of fundamental concepts about sampling and estimation from a reform perspective.

LEARN MORE

Listed next are three works about the significance testing controversy from fields other than psychology, including Armstrong (2007) in forecasting;

Guthery, Lusk, and Peterson (2001) in wildlife management; and McCloskey and Ziliak (2009) in medicine.

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting, 23*, 321–327. doi:10.1016/j.ijforecast.2007.03.004

Guthery, F. S., Lusk, J. J., & Peterson, M. J. (2001). The fall of the null hypothesis: Liabilities and opportunities. *Journal of Wildlife Management, 65*, 379–384. doi:10.2307/3803089

McCloskey, D. N., & Ziliak, S. T. (2009). The unreasonable ineffectiveness of Fisherian "tests" in biology, and especially in medicine. *Biological Theory, 4*, 44–53. doi:10.1162/biot.2009.4.1.44

This page intentionally left blank

2

SAMPLING AND ESTIMATION

> In times of change, learners inherit the Earth, while the learned find themselves beautifully equipped to deal with a world that no longer exists.
> —Eric Hoffer (1973, p. 22)

Fundamental concepts of sampling and estimation are the subject of this chapter. You will learn that (a) sampling error affects virtually all sample statistics, (b) interval estimation approximates margins of error associated with statistics, but (c) there are other sources of error variance that should not be ignored. You will also learn about central versus noncentral test statistics, the role of bootstrapping in interval estimation, and the basics of robust estimation. Entire books are devoted to some of these topics, so it is impossible in a single chapter to describe all of them in detail. Instead, the goal is to make you aware of concepts that underlie key aspects of statistics reform.

SAMPLING AND ERROR

A basic distinction in the behavioral sciences is that between populations and samples. It is rare that entire populations are studied. If a population is large, vast resources may be needed. For instance, the budget for

DOI: 10.1037/14136-002
Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, Second Edition, by R. B. Kline

the 2010 Census in the United States was $13 billion, and about 635,000 temporary workers were hired for it (U.S. Census Bureau, 2010). It may be practically impossible to study even much smaller populations. The base rate of schizophrenia, for example, is about 1%. But if persons with schizophrenia are dispersed over a large geographic area, studying all of them is probably impracticable.

Types of Samples

Behavioral scientists usually study samples, of which there are four basic kinds: random, systematic, ad hoc, and purposive. **Random (probability) samples** are selected by a chance-based method that gives all observations an equal likelihood of appearing in the sample. Variations on simple random sampling include stratified sampling and cluster sampling. In both, the population is divided into smaller groups that are mutually exclusive and collectively exhaustive. In **stratified sampling**, these groups are referred to as strata, and they are formed on the basis of shared characteristics. Strata may have quite different means on variables of interest. A random sample is taken from each stratum in proportion to its relative size in the population, and these subsamples are then pooled to form the total sample. Normative samples of psychological tests are often stratified on the basis of combinations of variables such as age, gender, or other demographic characteristics.

Partitions of the population are called clusters in cluster sampling. Each cluster should be generally representative of the whole population, which implies that clusters should also be reasonably similar on average. That is, most of the variation should be within clusters, not between them. In **single-stage cluster sampling**, random sampling is used to select the particular clusters to study. Next, all elements from the selected clusters contribute to the total sample, but no observations from the unselected clusters are included. In **two-stage cluster sampling**, elements from within each selected cluster are randomly sampled. One benefit of cluster sampling is that costs are reduced by studying some but not all clusters. When clusters are geographic areas, cases in the final sample are from the selected regions only.

Random sampling implies independent observations, which means that the score of one case does not influence the score of any other. If couples complete a relationship satisfaction questionnaire in the presence of each other, their responses may not be independent. The independence assumption is critical in many types of statistical techniques. Scores from repeated measurement of the same case are probably not independent, but techniques for such data estimate the degree of dependence in the scores and thus control for it. If scores are really not independent, results of analyses that assume independence could be biased. There is no magic statistical fix for

lack of independence. Therefore, the independence requirement is usually met through design, measurement, and use of statistical techniques that take explicit account of score dependence, such as designs with repeated measures.

The discussion that follows assumes that random samples are not extraordinarily small, such as $N = 2$. More sophisticated ways to estimate minimum sample sizes are considered later, but for now let us assume more reasonable sample sizes of, say, $N = 50$ or so. There are misconceptions about random sampling. Suppose that a simple random sample is selected. What can be said about the characteristics of the observations in that sample? A common but incorrect response is to say that the observations are representative of the population. But this may not be true, because there is no guarantee that the characteristics of any particular random sample will match those in the population. People in a random sample could be older, more likely to be women, or wealthier compared with the general population. A stratified random sample may be representative in terms of the strata on which it is based (e.g., gender), but results on other, nonstrata variables are not guaranteed to be representative. It is only across replications, or in the long run, that characteristics of observations in random samples reflect those in the population. That is, random sampling generates representative samples on average over replications. This property explains the role of random sampling in the **population inference model**, which is concerned with generalizability of sample results (external validity).

There is a related misunderstanding about **randomization**, or random assignment of cases to conditions (e.g., treatment vs. control). A particular randomization is not guaranteed to result in equivalent groups such that there are no initial group differences confounded with the treatment effect. Randomization results in equivalent groups only on average. Sometimes it happens that randomly formed groups are clearly not equal on some characteristic. The expression "failure of random assignment" is used to describe this situation, but it is a misnomer because it assumes that randomization should guarantee equivalence every time it is used. Random assignment is part of the **randomization model**, which deals with the correctness of causal inference that treatment is responsible for changes among treated cases (internal validity).

The use of random sampling and randomization together—the **statistician's two-step**—guarantees that the average effect observed over replications of treatment–control comparisons will converge on the value of the population treatment effect. But this ideal is almost never achieved in real studies. This is because random sampling requires a list of all observations in the population, but such lists rarely exist. Randomization is widely used in experimental studies but usually with nonrandom samples. Many more studies are based on the randomization model than on the population inference

model, but it is the latter that is assumed by the probabilities, or *p* values, generated by statistical tests and used in confidence intervals.

Observations in **systematic samples** are selected according to an orderly sampling plan that may yield a representative sample, but this is not certain. Suppose that an alphabetical list of every household is available for some area. A random number between 10 and 20 is generated and turns out to be 17. Every 17th household on the list is contacted for an interview, which yields a 6% (1/17) sample in that area. Systematic samples are relatively rare in the behavioral sciences.

Most samples are neither random nor systematic but rather are **ad hoc samples**, also known as **convenience samples**, **accidental samples**, or **locally available samples**. Cases in such samples are selected because they happen to be available. Whether ad hoc samples are representative is often a concern. Volunteers differ from nonvolunteers, for example, and patients seen in one clinic may differ from those treated in others. One way to mitigate bias is to measure a posteriori a variety of sample characteristics and report them. This allows others to compare the sample with those in related studies. Another option is to compare the sample profile with that of the population (if such a profile exists) in order to show that an ad hoc sample is not grossly unrepresentative.

The cases in a **purposive sample** are intentionally selected from defined groups or dimensions in ways linked to hypotheses. A researcher who wishes to evaluate whether the effectiveness of a drug varies by gender would intentionally select both women and men. After the data are collected, gender would be represented as a factor in the analysis, which may facilitate generalization of the results to both genders. A purposive sample is usually a convenience sample, and dividing cases by gender or some other variable does not change this fact.

Sampling Error

This discussion assumes a population size that is very large and assumes that the size of each sample is a relatively small proportion of the total population size. There are some special corrections if the population size is small, such as less than 5,000 cases, or if the sample size exceeds 20% or so of the population size that are not covered here (see S. K. Thompson [2012] for more information).

Values of population parameters, such as means (μ) or variances (σ^2), are usually unknown. They are instead estimated with sample statistics, such as M (means) or s^2 (variances). Statistics are subject to **sampling error**, which refers to the difference between an estimator and the corresponding parameter (e.g., $\mu - M$). These differences arise because the values of statistics from

random samples vary around that of the parameter. Some of these statistics will be too high and others too low (i.e., they over- or underestimate the parameter), and only a relatively small number will exactly equal the population value. This variability among estimators is a statistical phenomenon akin to background (natural) radiation: It is always there, sometimes more or less, fluctuating randomly from sample to sample.

The amount of sampling error is generally affected by the variability of population observations, how the samples are selected, and their size. If the population is heterogeneous, values of sample statistics may also be quite variable. Obviously, estimators from biased samples may differ substantially from those of the corresponding parameters. But assuming random sampling and constant variability in the population, sampling error varies inversely with sample size. This means that statistics in larger samples tend to be closer on average than those in smaller samples to the corresponding parameter. This property describes the **law of large numbers**, and it says that one is more likely to get more accurate estimates from larger samples than smaller samples with random sampling.

It is a myth that the larger the sample, the more closely it approximates a normal distribution. This idea probably stems from a misunderstanding of the **central limit theorem**, which applies to certain group statistics such as means. This theorem predicts that (a) distributions of random means, each based on the same number of scores, get closer to a normal distribution as the sample size increases, and (b) this happens regardless of whether the population distribution is normal or not normal. This theorem justifies approximating distributions of random means with normal curves, but it does not apply to distributions of scores in individual samples. Thus, larger samples do not generally have more normal distributions than smaller samples. If the population distribution is, say, positively skewed, this shape will tend to show up in the distributions of random samples that are either smaller or larger.

The sample mean describes the central tendency of a distribution of scores on a continuous variable. It is the balance point in a distribution, because the mean is the point from which (a) the sum of deviations from M equals zero and (b) the sum of squared deviations is as small as possible. The latter quantity is the **sum of squares** (*SS*). That is, if *X* represents individual observations, then

$$\sum(X - M) = 0 \text{ and the quantity } SS = \sum(X - M)^2 \qquad (2.1)$$

takes on the lowest value possible in a particular sample. Due to these properties, sample means are described as **least squares estimators**. The statistic M is also an **unbiased estimator** because its expected value across random samples of the same size is the population mean μ.

The sample variance s^2 is another least squares estimator. It estimates the population variance σ^2 without bias if computed as

$$s^2 = \frac{SS}{df} \tag{2.2}$$

where $df = N - 1$. But the sample variance derived as

$$S^2 = \frac{SS}{N} \tag{2.3}$$

is a **negatively biased estimator** because its values are on average less than σ^2. The reason is that squared deviations are taken from M (Equation 2.1), which is not likely to equal μ. Therefore, sample sums of squares are generally too small compared with taking squared deviations from μ. The division of *SS* by *df* instead of *N*, which makes the whole ratio larger ($s^2 > S^2$), is sufficient to render s^2 an unbiased estimator. In larger samples, though, the values of s^2 and S^2 converge, and in very large samples they are asymptotically equal. Expected values of **positively biased estimators** exceed those of the corresponding parameter.

There are ways to correct other statistics for bias. For example, although s^2 is an unbiased estimator of σ^2, the sample standard deviation *s* is a negatively biased estimator of σ. Multiplication of *s* by the correction factor in parentheses that follows

$$\hat{\sigma} = \left(1 + \frac{1}{4df}\right)s \tag{2.4}$$

yields a numerical approximation to the unbiased estimator of σ. Because the value of the correction factor in Equation 2.4 is larger than 1.00, $\hat{\sigma} > s$. There is also greater correction for negative bias in smaller samples than in larger samples. If $N = 5$, for example, the value of the correction factor is 1.0625, but for $N = 50$ it is 1.0051, which shows relatively less adjustment for bias in the larger sample. For very large samples, the value of the correction factor is essentially 1.0. This is another instance of the law of large numbers: Averages of even biased statistics from large random samples tend to closely estimate the corresponding parameter.

A **standard error** is the standard deviation in a **sampling distribution**, the probability distribution of a statistic across all random samples drawn from the same population(s) and with each sample based on the same number of cases. It estimates the amount of sampling error in standard deviation units. The square of a standard error is the error variance. Standard errors of

statistics with simple distributions can be estimated with formulas that have appeared in statistics textbooks for some time. By "simple" I mean that (a) the statistic estimates only a single parameter and (b) both the shape and variance of its sampling distribution are constant regardless of the value of that parameter. Distributions of M and s^2 are simple as just defined.

The standard error in a distribution of random means is

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \tag{2.5}$$

Because σ is not generally known, this standard error is typically estimated as

$$s_M = \frac{s}{\sqrt{N}} \tag{2.6}$$

As either sample variability decreases or the sample size increases, the value of s_M decreases. For example, given $s = 10.00$, s_M equals $10.00/25^{1/2}$, or 2.00, for $N = 25$, but for $N = 100$ the value of s_M is $10.00/100^{1/2}$, or 1.00. That is, the standard error is twice as large for $N = 25$ as it is for $N = 100$. A graphical illustration is presented in Figure 2.1. An original normal distribution is shown along with three different sampling distributions of M based on $N = 4$, 16, or 64 cases. Variability of the sampling distributions in the figure decreases as the sample size increases.

The standard error s_M, which estimates variability of the group statistic M, is often confused with the standard deviation s, which measures variability at the case level. This confusion is a source of misinterpretation of both statistical tests and confidence intervals (Streiner, 1996). Note that

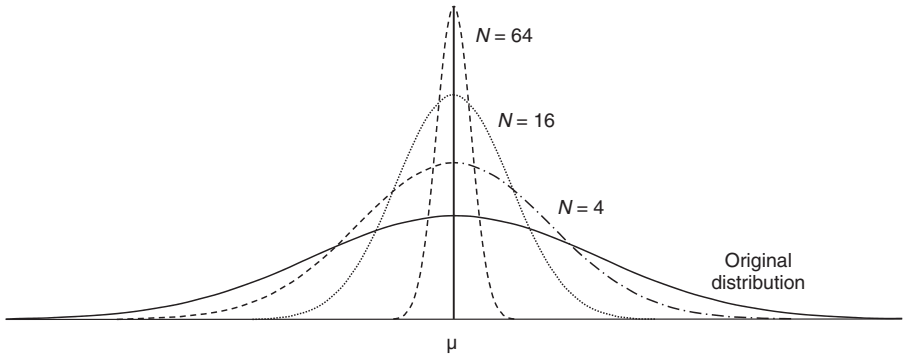

Figure 2.1. An original distribution of scores and three distributions of random sample means each based on different sample sizes, $N = 4$, $N = 16$, or $N = 64$.

the standard error s_M itself has a standard error (as do standard errors for all other kinds of statistics). This is because the value of s_M varies over random samples. This explains why one should not overinterpret a confidence interval or *p* value from a significance test based on a single sample. Exercises 1–2 concern the distinction between *s* and s_M.

Distributions of random means follow **central (Student's) *t* distributions** with degrees of freedom equal to $N - 1$ when σ is unknown. For very large samples, central *t* distributions approximate a normal curve. In **central test distributions**, the null hypothesis is assumed to be true. They are used to determine critical values of test statistics. Tables of critical values for distributions such as *t*, *F*, and χ^2 found in many statistics textbooks are based on central test distributions. There are also web calculating pages that generate critical values for central test statistics.[1] The *t* distribution originated from "Student's" (William Gosset's) attempt to approximate the distributions of means when the sample size is not large and σ is unknown. It was only later that central *t* distributions and other theoretical probability distributions were associated with the practice of significance testing.

The sample variance s^2 follows a **central χ^2 distribution** with $N - 1$ degrees of freedom. Listed next is the equation for the standard error of s^2 when the population variance is known:

$$\sigma_{s^2} = \sigma^2 \sqrt{\frac{2}{df}} \tag{2.7}$$

If σ^2 is not known, the standard error of the sample variance is estimated as

$$s_{s^2} = s^2 \sqrt{\frac{2}{df}} \tag{2.8}$$

As with *M*, the estimated standard error of s^2 becomes smaller as the sample size increases.

Other Kinds of Error

Standard errors estimate sampling error under random sampling. What they measure when sampling is not random may not be clear. The standard error in an ad hoc sample might reflect both sampling error and systematic

[1]This central *t* distributional calculator accepts either integer or noninteger *df* values: http://www.usablestats.com/calcs/tinv

selection bias that results in nonrepresentative samples. Standard errors also ignore the other sources of error described next:

1. **Measurement error** refers to the difference between an observed score X and the true score on the underlying construct. The reliability coefficient r_{XX} estimates the degree of measurement error in a particular sample. If $r_{XX} = .80$, for example, at least $1 - .80 = .20$, or 20%, of the observed variance in X is due to random error of the type estimated by that particular reliability coefficient. Measurement error reduces absolute effect sizes and the power of statistical tests. It is controlled by selecting measures that generally yield scores with good psychometric characteristics.
2. **Construct definition error** involves problems with how hypothetical constructs are defined or operationalized. Incorrect definition could include mislabeling a construct, such as when low IQ scores among minority children who do not speak English as a first language are attributed to low intelligence instead of to limited language familiarity. Error can also stem from **construct proliferation**, where a researcher postulates a new construct that is questionably different from existing constructs (F. L. Schmidt, 2010). Constructs that are theoretically distinct in the minds of researchers are not always empirically distinct.
3. **Specification error** refers to the omission from a regression equation of at least one predictor that covaries with the measured (included) predictors.[2] As covariances between omitted and included predictors increase, results based on the included predictors tend to become increasingly biased. Careful review of theory and research when planning a study is the main way to avoid a serious specification error by decreasing the potential number of left-out variables.
4. **Treatment implementation error** occurs when an intervention does not follow prescribed procedures. The failure to ensure that patients take an antibiotic medication for the prescribed duration of time is an example. Prevention includes thorough training of those who will administer the treatment and checking after the study begins whether implementation remains consistent and true.

[2]It can also refer to including irrelevant predictors, estimating linear relations only when the true relation is curvilinear, or estimating main effects only when there is true interaction.

Shadish, Cook, and Campbell (2001) described additional potential sources of error. Gosset used the term **real error** to refer all types of error besides sampling error (e.g., Student, 1927). In reasonably large samples, the impact of real error may be greater than that of sampling error. Thus, it is unwise to acknowledge sampling error only. This discussion implies that the probability that error of any kind affects sample results is virtually 1.00, and, therefore, practically all sample results are wrong (the parameter is not correctly estimated). This may be especially true when sample sizes are small, population effect sizes are not large, researchers chase statistical significance instead of substantive significance, a greater variety of methods is used across studies, and there is financial or other conflict of interest (Ioannidis, 2005).

INTERVAL ESTIMATION

Assumed next is the selection of a very large number of random samples from a very large population. The amount of sampling error associated with a statistic is explicitly indicated by a confidence interval, precisely defined by Steiger and Fouladi (1997) as follows:

1. A $1 - \alpha$ confidence interval for a parameter is a pair of statistics yielding an interval that, over many random samples, includes the parameter with the probability $1 - \alpha$. (The symbol α is the level of statistical significance.)
2. A $100\ (1 - \alpha)\%$ confidence interval for a parameter is a pair of statistics yielding an interval that, over many random samples, includes the parameter $100\ (1 - \alpha)\%$ of the time.

The value of $1 - \alpha$ is selected by the researcher to reflect the degree of statistical uncertainty due to sampling error. Because the conventional levels of statistical significance are .05 or .01, one usually sees either 95% or 99% confidence intervals, but it is possible to specify a different level, such as $\alpha = .10$ for a 90% confidence interval. Next we consider 95% confidence intervals only, but the same ideas apply to other confidence levels.

The lower bound of a confidence interval is the **lower confidence limit**, and the upper bound is the **upper confidence limit**. The *Publication Manual* (APA, 2010) recommends reporting a confidence interval in text with brackets. If 21.50 and 30.50 are, respectively, the lower and upper bounds for the 95% confidence interval based on a sample mean of 26.00, these results would be summarized as

$$M = 26.00,\ 95\%\ \text{CI}\ [21.50,\ 30.50]$$

Confidence intervals are often shown in graphics as **error bars** represented as lines that extend above and below (or to the left and right, depending on orientation) around a point that corresponds to a statistic. When the length of each error bar is one standard error ($M \pm s_M$), the interval defined by those **standard error bars** corresponds roughly to $\alpha = .32$ and a 68% confidence interval. There are also **standard deviation bars**. For example, the interval $M \pm s$ says something about the variability of scores around the mean, but it conveys no direct information about the extent of sampling error associated with that mean. Researchers do not always state what error bars represent: About 30% of articles with such figures reviewed by Cumming, Fidler, and Vaux (2007) did not provide this information.

Traditional confidence intervals are based on central test distributions, and the statistic is usually exactly between the lower and upper bounds (the interval is symmetrical about the estimator). The interval is constructed by adding and subtracting from a statistic the product of its standard error and the positive two-tailed critical value at the α level of statistical significance in a relevant central test distribution. This product is the **margin of error**. In graphical displays of confidence intervals, each of the two error bars corresponds to a margin of error.

Confidence Intervals for μ

The relevant test statistic for means when σ is unknown is central t, so the general form of a 100 $(1 - \alpha)$% confidence interval for μ based on a single observed mean is

$$M \pm s_M \left[t_{2\text{-tail},\,\alpha} (N-1) \right] \tag{2.9}$$

where the term in brackets is the positive two-tailed critical value in a central t distribution with $N - 1$ degrees of freedom at the α level of statistical significance. Suppose that

$$M = 100.00,\ s = 9.00, \text{ and } N = 25$$

The standard error is

$$s_M = \frac{9.00}{\sqrt{25}} = 1.800$$

and $t_{2\text{-tail},\,.05}\,(24) = 2.064$. The 95% confidence interval for μ is thus

$$100.00 \pm 1.800\,(2.064), \text{ or } 100.00 \pm 3.72$$

which defines the interval [96.28, 103.72]. Exercise 3 asks you to verify that the 99% confidence interval is wider than the 95% confidence interval based on the same data. Cumming (2012) described how to construct one-sided confidence intervals that are counterparts to statistical tests of null hypothesis versus directional (one-tailed) alternative hypotheses, such as H_1: $\mu > 130.00$.

Let us consider how to interpret the specific 95% confidence interval for μ just derived:

1. The interval [96.28, 103.72] defines a range of values considered equivalent within the limits of sampling error at the 95% confidence level. But equivalent within the bounds of sampling error does not imply equivalent in a scientific sense. This is especially true when the range of values included in the confidence interval indicates very different outcomes, such as when the upper confidence limit for the average blood concentration of a drug exceeds a lethal dosage.
2. It also provides a reasonable estimate of the population mean. That is, μ could be as low as 96.28 or μ could be as high as 103.72, again at the 95% confidence level.
3. There is no guarantee that μ is actually included in the confidence interval. We could construct the 95% confidence interval based on the mean in a different sample, but the center or endpoints of this new interval will probably be different. This is because confidence intervals are subject to sampling error, too.
4. If 95% confidence intervals are constructed around the means of very many random samples drawn from the same very large population, a total of 95% of them will contain μ.

The last point gives a more precise definition of "95% confident" from a **frequentist** or **long-run relative-frequency** view of probability as the likelihood of an outcome over repeatable events under constant conditions except for random error. A frequentist view assumes that probability is a property of nature that is independent of what the researcher believes. In contrast, a **subjectivist** or **subjective degree-of-belief** view defines probability as a personal belief that is independent of nature. The same view also does not distinguish between repeatable and unique events (Oakes, 1986). Although researchers in their daily lives probably take a subjective view of probabilities, it is the frequentist definition that generally underlies sampling theory.

A researcher is probably more interested in knowing the probability that a specific 95% confidence interval contains μ than in knowing that

95% of all such intervals do. From a frequentist perspective, this probability for any specific interval is either 0 or 1.00; that is, either the interval contains the parameter or it does not. Thus, it is generally incorrect to say that a specific 95% confidence interval has a 95% likelihood of including the corresponding parameter. Reichardt and Gollob (1997) noted that this kind of **specific probability inference** is permitted only in the circumstance that every possible value of the parameter is considered equally likely before the data are collected. In Bayesian estimation, the same circumstance is described by the **principle of indifference**, but it is rare when a researcher truly has absolutely no information about plausible values for a parameter.

There is language that splits the difference between frequentist and subjectivist perspectives. Applied to our example, it goes like this: The interval [96.28, 103.72] estimates μ, with 95% confidence. This statement is not quite a specific probability inference, and it also gives a nod to the subjectivist view because it associates a degree of belief with a unique interval. Like other compromises, however, it may not please purists who hold one view of probability or the other. But this wording does avoid the blatant error of claiming that a specific 95% confidence interval contains the parameter with the probability .95.

Another interpretation concerns the **capture percentage** of random means from replications that fall within the bounds of a specific 95% confidence interval for μ. Most researchers surveyed by Cumming, Williams, and Fidler (2004) mistakenly endorsed the **confidence-level misconception** that the capture percentage for a specific 95% confidence interval is also 95%. This fallacy for our example would be stated as follows: The interval [96.28, 103.72] contains 95% of all replication means. This statement would be true for this interval only if the values of $\mu - M$ and $\sigma - s$ were both about zero; otherwise, capture percentages drop off quickly as the absolute distance between μ and M increases. Cumming and Maillardert (2006) estimated that the average capture percentage across random 95% confidence intervals for μ is about 85% assuming normality and $N \geq 20$, but percentages for more extreme samples are much lower (e.g., < 50%).

These results suggest that researchers underestimate the impact of sampling error on means. Additional evidence described in the next chapter says that researchers fail to appreciate that sampling error affects p values from statistical tests, too. It seems that many researchers believe that results from small samples behave like those from large samples; that is, they believe that results from small samples are likely to replicate. Tversky and Kahneman (1971) labeled such errors **the law of small numbers**, an ironic twist on the law of large numbers, which (correctly) says that there is greater variation across results from small samples than from large samples.

Confidence Intervals for $\mu_1 - \mu_2$

Next we assume a design with two independent samples. The standard error in a distribution of contrasts between pairs of means randomly selected from different populations is

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{2.10}$$

where σ_1^2 and σ_2^2 are the population variances and n_1 and n_2 are the sizes of each group. If we assume homogeneity of population variance or **homoscedasticity** (i.e., $\sigma_1^2 = \sigma_2^2$), the expression for the standard error reduces to

$$\sigma_{M_1-M_2} = \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{2.11}$$

where σ^2 is the common population variance. This parameter is usually unknown, so the standard error of mean differences is estimated by

$$s_{M_1-M_2} = \sqrt{s_{\text{pool}}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{2.12}$$

where s_{pool}^2 is the weighted average of the within-groups variances. Its equation is

$$s_{\text{pool}}^2 = \frac{df_1(s_1^2) + df_2(s_2^2)}{df_1 + df_2} = \frac{SS_W}{df_W} \tag{2.13}$$

where s_1^2 and s_2^2 are the group variances, $df_1 = n_1 - 1$, $df_2 = n_2 - 1$, and SS_W and df_W are, respectively, the pooled within-groups sum of squares and the degrees of freedom. The latter can also be expressed as $df_W = N - 2$. Only when the group sizes are equal can s_{pool}^2 also be calculated as the simple average of the two group variances, or $(s_1^2 + s_2^2)/2$.

The general form of a $100\,(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ based on the difference between two independent means is

$$(M_1 - M_2) \pm s_{M_1-M_2}\left[t_{2\text{-tail},\,\alpha}(N-2)\right] \tag{2.14}$$

Suppose in a design with $n = 10$ cases in each group we observe

$$M_1 = 13.00,\ s_1^2 = 7.50 \quad \text{and} \quad M_2 = 11.00,\ s_2^2 = 5.00$$

which implies $M_1 - M_2 = 2.00$ and $s^2_{pool} = (7.50 + 5.00)/2 = 6.25$. The estimated standard error is

$$s_{M_1 - M_2} = \sqrt{6.25\left(\frac{1}{10} + \frac{1}{10}\right)} = 1.118$$

and $t_{2\text{-tail}, .05}\,(18) = 2.101$. The 95% confidence interval for $\mu_1 - \mu_2$ is

$$2.00 \pm 1.118\,(2.101)$$

which defines the interval [–.35, 4.35]. On the basis of these results, we can say that $\mu_1 - \mu_2$ could be as low as –.35 or as high as 4.35, with 95% confidence.

The specific interval [–.35, 4.35] includes zero as an estimate of $\mu_1 - \mu_2$. This fact is subject to misinterpretation. For example, it may be incorrectly concluded that $\mu_1 = \mu_2$ because zero falls within the interval. But zero is only one value within a range of estimates of $\mu_1 - \mu_2$, so it has no special status in interval estimation. Confidence intervals are subject to sampling error, so zero may not be included within the 95% confidence interval in a replication. Confidence intervals also assume that other sources of error are nil. All these caveats should reduce the temptation to fixate on a particular value (here, zero) in a confidence interval.

There is special relation between a confidence interval for $\mu_1 - \mu_2$ and the outcome of the independent samples *t* test based on the same data: Whether a 100 $(1 - \alpha)$% confidence interval for $\mu_1 - \mu_2$ includes zero yields an outcome equivalent to either rejecting or not rejecting the corresponding null hypothesis at the α level of statistical significance for a two-tailed test. For example, the specific 95% confidence interval [–.35, 4.35] includes zero; thus, the outcome of the *t* test for these data of H_0: $\mu_1 - \mu_2 = 0$ is not statistically significant at the .05 level, or

$$t\,(18) = \frac{2.00}{1.118} = 1.789,\ p = .091$$

But if zero is not contained within a particular 95% confidence interval for $\mu_1 - \mu_2$, the outcome of the independent samples *t* test will be statistically significant at the .05 level.

Be careful not to falsely believe that confidence intervals are just statistical tests in disguise (B. Thompson, 2006a). One reason is that null hypotheses are required for statistical tests but not for confidence intervals. Another is that many null hypotheses have little if any scientific value. For example, Anderson et al. (2000) reviewed null hypotheses tested in several hundred empirical studies published from 1978 to 1998 in two environmental sciences

TABLE 2.1
Results of Six Hypothetical Replications

Study	$M_1 - M_2$	s_1^2	s_2^2	t (38)	Reject H_0?	95% CI
1	2.50	17.50	16.50	1.92	No	–.14, 5.14
2	4.00	16.00	18.00	3.07	Yes	1.36, 6.64
3	2.50	14.00	17.25	2.00	No	–.03, 5.03
4	4.50	13.00	16.00	3.74	Yes	2.06, 6.94
5	5.00	12.50	16.50	4.15	Yes	2.56, 7.44
6	2.50	15.00	17.00	1.98	No	–.06, 5.06
Average:	3.54					2.53, 4.54

Note. Independent samples assumed. For all replications, the group size is $n = 20$, $\alpha = .05$, the null hypothesis is H_0: $\mu_1 - \mu_2 = 0$, and H_1 is two-tailed. Results for the average difference are from a meta-analysis assuming a fixed effects model. CI = confidence interval.

journals. They found many implausible null hypotheses that specified things such as equal survival probabilities for juvenile and adult members of a species or that growth rates did not differ across species, among other assumptions known to be false before collecting data. I am unaware of a similar survey of null hypotheses in the behavioral sciences, but I would be surprised if the results would be very different.

Confidence intervals over replications may be less susceptible to misinterpretation than results of statistical tests. Summarized in Table 2.1 are outcomes of six hypothetical replications where the same two conditions are compared on the same outcome variable. Results of the independent samples t test lead to rejection of the null hypothesis at $p < .05$ in three out of six studies, a "tie" concerning statistical significance (3 yeas, 3 nays). More informative than the number of null hypothesis replications is the average of $M_1 - M_2$ across all six studies, 3.54. This average is from a meta-analysis of all results in the table for a **fixed effects model**, where a single population effect size is presumed to underlie the observed contrasts. (I show you how to calculate this average in Chapter 9.) The overall average of 3.54 may be a better estimate of $\mu_1 - \mu_2$ than $M_1 - M_2$ in any individual study because it is based on all available data.

The 95% confidence intervals for $\mu_1 - \mu_2$ in Table 2.1 are shown in Figure 2.2 as error bars in a **forest plot**, which displays results from replications and a meta-analytic weighted average with confidence intervals (Cumming, 2012). The 95% confidence interval based on the overall average of 3.54, or [2.53, 4.54] (see Table 2.1), is narrower than any of the intervals from the six replications (see Figure 2.2). This is because more information contributes to the confidence interval based on results averaged over all replications. For these data, $\mu_1 - \mu_2$ may be as low as 2.53 or as high as 4.54, with 95% confidence based on all available data.

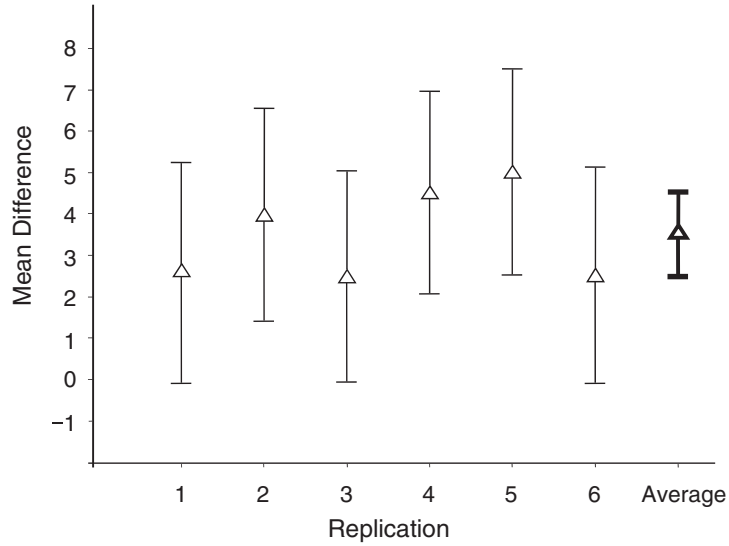

Figure 2.2. A forest plot of 95% confidence intervals for $\mu_1 - \mu_2$ based on mean differences from the six replications in Table 2.1 and the meta-analytic 95% confidence interval for $\mu_1 - \mu_2$ across all replications for a fixed effects model.

There is a widely accepted—but unfortunately incorrect—rule of thumb that the difference between two independent means is statistically significant at the α level if there is no overlap of the two 100 $(1 - \alpha)$% confidence intervals for μ (Belia, Fidler, Williams, & Cumming, 2005). It also maintains that the overlap of the two intervals indicates that the mean contrast is not statistically significant at the corresponding level of α. This rule is often applied to diagrams where confidence intervals for μ are represented as error bars that emanate outward from points that symbolize group means.

A more accurate heuristic is the **overlap rule for two independent means** (Cumming, 2012), which works best when $n \geq 10$ and the group sizes and variances are approximately equal. The overlap rule is stated next for $\alpha = .05$:

1. If there is a gap between the two 95% confidence intervals for μ (i.e., no overlap), the outcome of the independent samples *t* test of the mean difference is $p < .01$. But if the confidence intervals just touch end-to-end, *p* is approximately .01.
2. No more than moderate overlap of the 95% confidence intervals for μ implies that the *p* value for the *t* test is about .05, but less overlap indicates $p < .05$. *Moderate overlap* is about one half the length of each error bar in a graphical display.

Summarized next are the basic descriptive statistics for the example where $n_1 = n_2 = 10$:

$$M_1 = 13.00,\ s_1^2 = 7.50 \quad \text{and} \quad M_2 = 11.00,\ s_2^2 = 5.00$$

You should verify for these data the results presented next:

$$s_{M_1} = .866,\ 95\%\ \text{CI for}\ \mu_1\ [11.04, 14.96]$$

$$s_{M_2} = .707,\ 95\%\ \text{CI for}\ \mu_2\ [9.40, 12.60]$$

These confidence intervals for μ are plotted in Figure 2.3 along with the 95% confidence interval for $\mu_1 - \mu_2$ for these data [–.35, 4.35]. Group means are represented on the *y*-axis, and the mean contrast (2.00) is represented on the floating difference axis (Cumming, 2012) centered at the grand mean across both groups (12.00). The error bars of the 95% confidence intervals for μ overlap by clearly more than one half of their lengths. According to the overlap rule, this amount of overlap is more than moderate. So the mean difference should not be statistically significant at the .05 level, which is true for these data.

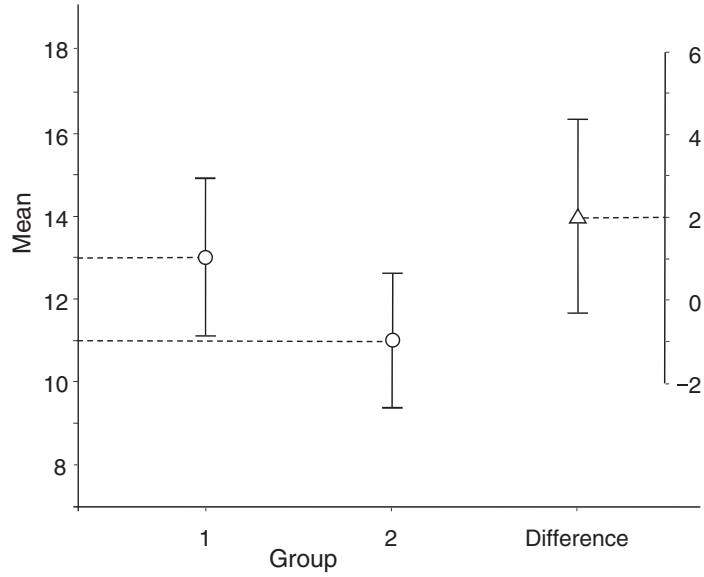

Figure 2.3. Plot of the 95% confidence interval for μ_1, 95% confidence interval for μ_2, and 95% confidence interval for $\mu_1 - \mu_2$, given $M_1 = 13.00$, $s_1^2 = 7.50$, $M_2 = 11.00$, $s_2^2 = 5.00$, and $n_1 = n_2 = 10$. Results for the mean difference are shown on a floating difference axis where zero is aligned at the grand mean across both samples (12.00).

Confidence intervals for $\mu_1 - \mu_2$ based on $s_{M_1 - M_2}$ assume homoscedasticity. In the **Welch procedure** (e.g., Welch, 1938), the standard error of a mean contrast is estimated as

$$s_{\text{Wel}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{2.15}$$

where s_1^2 estimates σ_1^2 and s_2^2 estimates σ_2^2 (i.e., heteroscedasticity is allowed). The degrees of freedom for the critical value of central t in the Welch procedure are estimated empirically as

$$df_{\text{Wel}} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2)^2}{n_1^2(n_1 - 1)} + \frac{(s_2^2)^2}{n_2^2(n_2 - 1)}} \tag{2.16}$$

Summarized next are descriptive statistics for two groups:

$$M_1 = 112.50,\ s_1^2 = 75.25,\ n_1 = 25$$
$$M_2 = 108.30,\ s_2^2 = 15.00,\ n_2 = 20$$

Variability among cases in the first group is obviously greater than that in the second group. A pooled within-groups variance would mask this discrepancy. The researcher elects to use the Welch procedure. The estimated standard error is

$$s_{\text{Wel}} = \sqrt{\frac{75.25}{25} + \frac{15.00}{20}} = 1.939$$

and the approximate degrees of freedom are

$$df_{\text{Wel}} = \frac{\left(\frac{75.25}{25} + \frac{15.00}{20}\right)^2}{\frac{75.25^2}{25^2(24)} + \frac{15.00^2}{20^2(19)}} = 34.727$$

The general form of a 100 $(1 - \alpha)$% confidence interval for $\mu_1 - \mu_2$ in the Welch procedure is

$$(M_1 - M_2) \pm s_{\text{Wel}} \left[t_{2\text{-tail},\alpha}(df_{\text{Wel}})\right] \tag{2.17}$$

Tables for critical values of central t typically list integer df values only. An alternative is to use a web distributional calculator page that accepts noninteger df (see footnote 1). Another is to use a statistical density function built into widely available software. The statistical function TINV in Microsoft Excel returns critical values of central t given values of α and df. The function Idf.T (Inverse DF) in SPSS returns the two-tailed critical value of central t given df and $1 - \alpha/2$, which is .975 for a 95% confidence interval. For this example, SPSS returned

$$t_{2\text{-tail, }.05}\,(34.727) = 2.031$$

The 95% confidence interval for $\mu_1 - \mu_2$ is

$$(112.50 - 108.30) \pm 1.939\,(2.031)$$

which defines the interval [.26, 8.14]. Thus, the value of $\mu_1 - \mu_2$ could be as low as .26 or as high as 8.14, with 95% confidence and not assuming homoscedasticity. Widths of confidence intervals in the Welch procedure tend to be narrower than intervals based on $s_{M_1 - M_2}$ for the same data when group variances are unequal. Welch intervals may less accurate when the population distributions are severely and differently nonnormal or when the group sizes are unequal and small, such as $n < 30$ (Bonett & Price, 2002); see also Grissom and Kim (2011, Chapter 2).

Confidence Intervals for μ_D

I use the symbol M_D to refer to the mean **difference (change, gain) score** when two dependent samples are compared. A difference score is computed as $D = X_1 - X_2$ for each of the n cases in a repeated measures design or for each of the n pairs of cases in a matched groups design. If $D = 0$, there is no difference; any other value indicates a higher score in one condition than in the other. The average of all difference scores equals the dependent mean contrast, or $M_D = M_1 - M_2$. Its standard error is

$$\sigma_{M_D} = \frac{\sigma_D}{\sqrt{n}} \tag{2.18}$$

where σ_D is the population standard deviation of the difference scores. The variance of the difference scores can be expressed as

$$\sigma_D^2 = 2\sigma^2\,(1 - \rho_{12}) \tag{2.19}$$

where σ^2 is the common population variance assuming homoscedasticity and ρ_{12} is the population cross-conditions correlation of the original scores.

When there is a stronger **subjects effect**—cases maintain their relative positions across the conditions—ρ_{12} approaches 1.00. This reduces the variance of the difference scores, which in turn lowers the standard error of the mean contrast (Equation 2.18). It is the subtraction of consistent individual differences from the standard error that makes confidence intervals based on dependent mean contrasts generally narrower than confidence intervals based on contrasts between unrelated means. It also explains the power advantage of the *t* test for dependent samples over the *t* test for independent samples. But these advantages are realized only if $\rho_{12} > .50$ (Equation 2.19); otherwise, confidence intervals and statistical tests may be wider and less powerful (respectively) for dependent mean contrasts.

The standard deviation σ_D is usually unknown, so the standard error of M_D is estimated as

$$s_{M_D} = \frac{s_D}{\sqrt{n}} \tag{2.20}$$

where s_D is the sample standard deviation of the D scores. The corresponding variance is

$$s_D^2 = s_1^2 + s_2^2 - 2cov_{12} \tag{2.21}$$

where cov_{12} is the cross-conditions covariance of the original scores. The latter is

$$cov_{12} = r_{12}\ s_1\ s_2 \tag{2.22}$$

where r_{12} is the sample cross-conditions correlation. (The correlation r_{12} is presumed to be zero when the samples are independent.)

The general form of a 100 $(1 - \alpha)$% confidence interval for μ_D is

$$M_D \pm s_{M_D}\left[t_{2\text{-tail},\alpha}(n-1)\right] \tag{2.23}$$

Presented in Table 2.2 are raw scores and descriptive statistics for a small data set where the mean contrast is 2.00. In a dependent samples analysis of these data, $n = 5$ and $r_{12} = .735$. The cross-conditions covariance is

$$cov_{12} = .735(2.739)(2.236) = 4.50$$

and the variance of the difference scores is

$$s_D^2 = 7.50 + 5.00 - 2(4.50) = 3.50$$

which implies that $s_D = 3.50^{1/2}$, or 1.871. The standard error of $M_D = 2.00$ is estimated as

TABLE 2.2
Raw Scores and Descriptive Statistics for Two Samples

	Sample	
	1	2
	9	8
	12	12
	13	11
	15	10
	16	14
M	13.00	11.00
s^2	7.50	5.00
s	2.739	2.236

Note. In a dependent samples analysis, $r_{12} = .735$.

$$s_{M_D} = \frac{1.871}{\sqrt{5}} = .837$$

The value of $t_{2\text{-tail}, .05}$ (4) is 2.776, so the 95% confidence interval for μ_D is

$$2.00 \pm .837(2.776)$$

which defines the interval [–.32, 4.32]. Exercise 4 asks you to verify that the 95% confidence interval for μ_D assuming a correlated design is narrower than the 95% confidence interval for $\mu_1 - \mu_2$ assuming unrelated samples for the same data (see Table 2.2), which is [–1.65, 5.65].

Confidence Intervals Based on Other Kinds of Statistics

Many statistics other than means have complex distributions. For example, distributions of the Pearson correlation *r* are symmetrical only if the population correlation is $\rho = 0$, but they are negatively skewed when $\rho > 0$ and positively skewed when $\rho < 0$. Other statistics have complex distributions, including some widely used effect sizes introduced in Chapter 5, because they estimate more than one parameter.

Until recently, confidence intervals for statistics with complex distributions were estimated with approximate methods. One method involves **confidence interval transformation** (Steiger & Fouladi, 1997), where the statistic is mathematically transformed into normally distributed units. The confidence interval is built by adding and subtracting from the transformed statistic the product of the standard error in the transformed metric and the appropriate critical value of the normal deviate *z*. The lower and upper bounds

of this interval are then transformed back into the original metric, and the resulting confidence interval may be asymmetric (unequal margins of error). **Fisher's transformation** is used to approximate construct intervals for ρ. It converts a sample correlation r with the function

$$Z_r = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \tag{2.24}$$

where ln is the natural log function to base e, which is about 2.7183. The sampling distribution of Z_r is approximately normal with the standard error

$$s_{Z_r} = \sqrt{\frac{1}{N-3}} \tag{2.25}$$

The lower and upper bounds of the 100 $(1 - \alpha)$% confidence interval based on Z_r are defined by

$$Z_r \pm s_{Z_r} \left(z_{2\text{-tail}, \alpha} \right) \tag{2.26}$$

where $z_{2\text{-tail}, \alpha}$ is the positive two-tailed critical value of the normal deviate, which is 1.96 for $\alpha = .05$ and the 95% confidence level. Next, transform both the lower and upper bounds of the confidence interval in Z_r units back to r units by applying the inverse transformation

$$r_Z = \frac{e^{2Z_r} - 1}{e^{2Z_r} + 1} \tag{2.27}$$

There are calculating web pages that automatically generate approximate 95% or 99% confidence intervals for ρ, given values of r and the sample size.[3] Four-decimal accuracy is recommended for hand calculation.

In a sample of $N = 20$ cases, $r = .6803$. Fisher's transformation and its standard error are

$$Z_r = \frac{1}{2} \ln \left(\frac{1 + .6803}{1 - .6803} \right) = .8297 \quad \text{and} \quad s_{Z_r} = \sqrt{\frac{1}{20 - 3}} = .2425$$

The approximate 95% confidence interval in Z_r units is

$$.8297 \pm .2425 (1.96)$$

[3] http://faculty.vassar.edu/lowry/rho.html

which defines the interval [.3544, 1.3051]. To convert the lower and upper bounds of this interval to *r* units, I apply the inverse transformation to each:

$$\frac{e^{2(.3544)}-1}{e^{2(.3544)}+1} = .3403 \quad \text{and} \quad \frac{e^{2(1.3051)}-1}{e^{2(1.3051)}+1} = .8630$$

In *r* units, the approximate 95% confidence interval for ρ is [.34, .86] at two-place accuracy.

Another approximate method builds confidence intervals directly around the sample statistic; thus, they are symmetrical about it. The width of the interval on either side is a product of the two-tailed critical value of a central test statistic and an estimate of the **asymptotic standard error**, which estimates what the standard error would be in a large sample (e.g., > 500). If the researcher's sample is not large, though, this estimate may not be accurate. Another drawback is that some statistics, such as R^2 in multiple regression, have distributions so complex that a computer is needed to estimate standard error. Fortunately, there are increasing numbers of computer tools for calculating confidence intervals, some of which are mentioned later.

A more precise method is **noncentrality interval estimation** (Steiger & Fouladi, 1997). It also deals with situations that cannot be handled by approximate methods. This approach is based on **noncentral test distributions** that do not assume a true null hypothesis. Some perspective is in order. Families of central distributions of *t*, *F*, and χ^2 (in which H_0 is assumed to be true) are special cases of noncentral distributions of each test statistic just mentioned. Compared to central distributions, noncentral distributions have an extra parameter called the **noncentrality parameter** that indicates the degree to which the null hypothesis is false.

Central *t* distributions are defined by a single parameter, the degrees of freedom (*df*), but noncentral *t* distributions are described by both *df* and the noncentrality parameter Δ (Greek uppercase delta). In two-group designs, the value of Δ for noncentral *t* is related to (but not exactly equal to) the true difference between the population means μ_1 and μ_2. The larger that difference, the more the noncentral *t* distribution is skewed. That is, if $\mu_1 > \mu_2$, then $\Delta > 0$ and the resulting noncentral *t* distributions are positively skewed, and if $\mu_1 < \mu_2$, then $\Delta < 0$ and the corresponding resulting noncentral *t* distributions are negatively skewed. But if $\mu_1 = \mu_2$ (i.e., there is no difference), then $\Delta = 0$ and the resulting distributions are the familiar and symmetrical central *t* distributions. Presented in Figure 2.4 are two *t* distributions where $df = 10$. For the central *t* distribution in the left part of the figure, $\Delta = 0$, but for the noncentral *t* distribution in the right side of the figure, $\Delta = 4.00$. (The meaning of a particular value for Δ is defined in Chapter 5.) Note in the figure that the distribution for noncentral *t* (10, 4.00) is positively skewed.

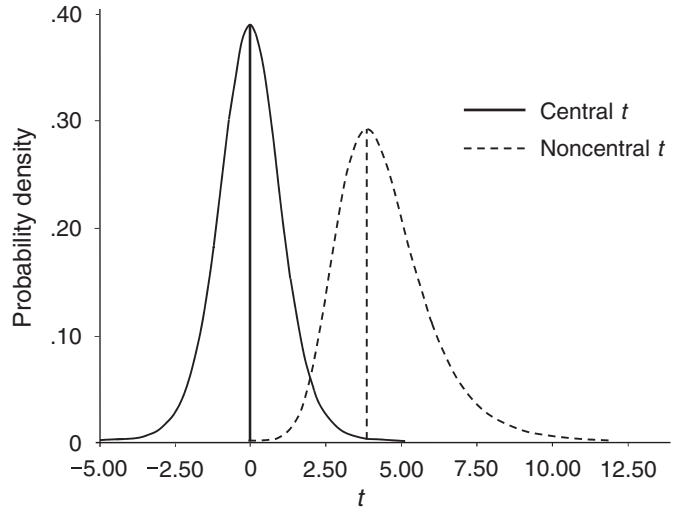

Figure 2.4. Distributions of central *t* and noncentral *t* where the degrees of freedom are $df = 10$ and where the noncentrality parameter is $\Delta = 4.00$ for noncentral *t*.

Noncentral test distributions play a role in estimating the power of statistical tests. This is because the concept of power assumes that the null hypothesis is false. Thus, computer tools for power analysis analyze noncentral test distributions. A population effect size that is not zero generally corresponds to a value of the noncentrality parameter that is also not zero. This is why some methods of interval estimation for effect sizes rely on noncentral test distributions. Noncentrality interval estimation for effect sizes is covered in Chapter 5.

Calculating noncentral confidence intervals is impractical without relatively sophisticated computer programs. Until recently, such programs were not widely available to applied researchers. An exception is Exploratory Software for Confidence Intervals (ESCI; Cumming, 2012), which runs under Microsoft Excel. It is structured as a tool for learning about confidence intervals, noncentral test distributions, power estimation, and meta-analysis. Demonstration modules for ESCI can be downloaded.[4] I used ESCI to create Figure 2.4.

Another computer tool for power estimation and noncentrality interval estimation is Steiger's Power Analysis procedure in STATISTICA 11 Advanced, an integrated program for general statistical analyses, data mining,

[4]http://www.thenewstatistics.com/

and quality control.[5] Power Analysis can automatically calculate noncentral confidence intervals based on several different types of effect sizes. Other computer tools or scripts for interval estimation with effect sizes are described in later chapters. The website for this book also has links to corresponding download pages. Considered next is bootstrapping, which can also be used for interval estimation.

BOOTSTRAPPED CONFIDENCE INTERVALS

The technique of **bootstrapping**, developed by the statistician Bradley Efron in the 1970s (e.g., 1979), is a computer-based method of **resampling** that recombines the cases in a data set in different ways to estimate statistical precision, with fewer assumptions than traditional methods about population distributions. Perhaps the best known form is **nonparametric bootstrapping**, which generally makes no assumptions other than that the distribution in the sample reflects the basic shape of that in the population. It treats your data file as a pseudo-population in that cases are randomly selected with replacement to generate other data sets, usually of the same size as the original. Because of sampling with replacement, (a) the same case can be selected in more than one generated data set or at least twice in the same generated sample, and (b) the composition of cases will vary slightly across the generated samples.

When repeated many times (e.g., 1,000) by the computer, bootstrapping simulates the drawing of many random samples. It also constructs an **empirical sampling distribution**, the frequency distribution of the values of a statistic across the generated samples. **Nonparametric percentile bootstrapped confidence intervals** for the parameter estimated by the statistic are calculated in the empirical distribution. The lower and upper bounds of a 95% bootstrapped confidence interval correspond to, respectively, the 2.5th and 97.5th percentiles in the empirical sampling distribution. These limits contain 95% of the bootstrapped values of the statistic.

Presented in Table 2.3 is a small data set where $N = 20$ and $r = .6803$. I used the nonparametric bootstrap procedure of SimStat for Windows (Provalis Research, 1995–2004) to resample from the data in Table 2.3 in order to generate a total of 1,000 bootstrapped samples each with 20 cases.[6] The empirical sampling distribution is presented in Figure 2.5. As expected, this distribution is negatively skewed. SimStat reported that the mean and median of the sampling distribution are, respectively, .6668 and .6837. The standard deviation in the distribution of Figure 2.5 is .1291, which is actually

[5]http://www.statsoft.com/#
[6]http://www.provalisresearch.com/

TABLE 2.3
Example Data Set for Nonparametric Bootstrapping

Case	*X*	*Y*	Case	*X*	*Y*
A	12	16	K	16	37
B	19	46	L	13	30
C	21	66	M	18	32
D	16	70	N	18	53
E	18	27	O	22	52
F	16	27	P	17	34
G	16	44	Q	22	54
H	20	69	R	12	5
I	16	22	S	14	38
J	18	61	T	14	38

the bootstrapped estimate of the standard error. The nonparametric bootstrapped 95% confidence interval for ρ is [.3615, .8626], and the bias-adjusted 95% confidence interval is [.3528, .8602]. The latter controls for lack of independence due to potential selection of the same case multiple times in the same generated sample.

The bias-adjusted bootstrapped 95% confidence interval for ρ, which is [.35, .86] at two-decimal accuracy, is similar to the approximate 95% confidence interval of [.34, .86] calculated earlier using Fisher's approximation for the same data. The bootstrapped estimate of the standard error in correlation units generated by SimStat is .129. Nonparametric bootstrapping is potentially

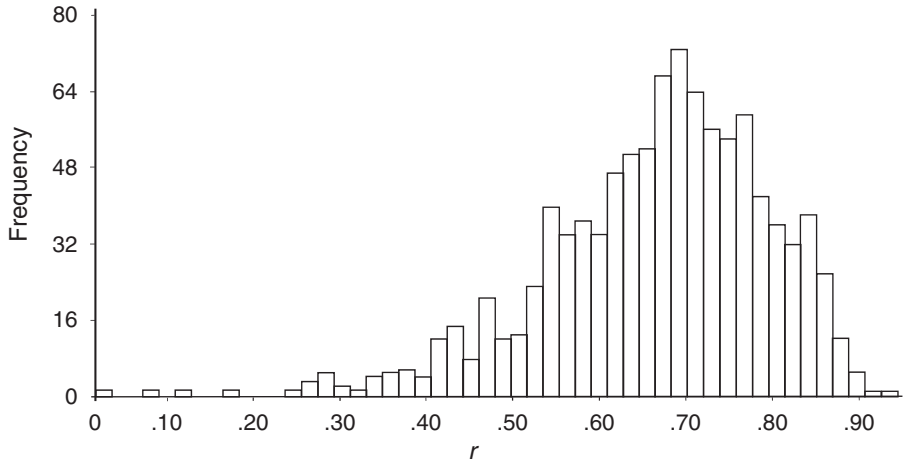

Figure 2.5. Empirical sampling distribution for the Pearson correlation *r* in 1,000 bootstrapped samples for the data in Table 2.3.

more useful when applied to statistics for which there is no approximate method for calculating standard errors and confidence intervals. This is also true when no computer tool for noncentral interval estimation is available for statistics with complex distributions.

The technique of nonparametric bootstrapping seems well suited for interval estimation when the researcher is either unwilling or unable to make a lot of assumptions about population distributions. Wood (2005) demonstrated the calculation of bootstrapped confidence intervals based on means, medians, differences between two means or proportions, correlations, and regression coefficients. His examples are implemented in an Excel spreadsheet[7] and a small stand-alone program.[8] Another computer tool is Resampling Stats (Statistics.com, 2009).[9] Bootstrapping capabilities were recently added to some procedures in SPSS and SAS/STAT.

Outlined next are potential limitations of nonparametric bootstrapping:

1. Nonparametric bootstrapping simulates random sampling, but true random sampling is rarely used in practice. This is another instance of the design–analysis mismatch.
2. It does not entirely free the researcher from having to make assumptions about population distributions. If the shape of the sample distribution is very different compared with that in the population, results of nonparametric bootstrapping may have poor external validity.
3. The "population" from which bootstrapped samples are drawn is merely the original data file. If this data set is small or the observations are not independent, resampling from it will not somehow fix these problems. In fact, resampling can magnify the effects of unusual features in a small data set (Rodgers, 2009).
4. Results of bootstrap analyses are probably quite biased in small samples, but this is true of many traditional methods, too.

The starting point for **parametric bootstrapping** is not a raw data file. Instead, the researcher specifies the numerical and distributional properties of a theoretical probability density function, and then the computer randomly samples from that distribution. When repeated many times by the computer, values of statistics in these synthesized samples vary randomly about the parameters specified by the researcher, which simulates sampling error. Bootstrapped estimation in parametric mode can also approximate standard

[7]http://woodm.myweb.port.ac.uk/nms/resample.xls
[8]http://woodm.myweb.port.ac.uk/nms/resample.exe
[9]Resampling Stats is available for a 10-day trial from http://www.resample.com/

errors for statistics where no textbook equation or approximate method is available, given certain assumptions about the population distribution. These assumptions can be added incrementally in parametric bootstrapping or successively relaxed over the generation of synthetic data sets.

ROBUST ESTIMATION

The least squares estimators M and s^2 are not robust against the effects of extreme scores. This is because their values can be severely distorted by even a single outlier in a smaller sample or by just a handful of outliers in a larger sample. Conventional methods to construct confidence intervals rely on sample standard deviations to estimate standard errors. These methods also rely on critical values in central test distributions, such as t and z, that assume normality or homoscedasticity (e.g., Equation 2.13).

Such distributional assumptions are not always plausible. For example, skew characterizes the distributions of certain variables such as reaction times. Many if not most distributions in actual studies are not even symmetrical, much less normal, and departures from normality are often strikingly large (Micceri, 1989). Geary (1947) suggested that this disclaimer should appear in all introductory statistics textbooks: "Normality is a myth; there never was, and never will be, a normal distribution" (p. 214). Keselman et al. (1998) reported that the ratios across different groups of largest to smallest variances as large as 8:1 were not uncommon in educational and psychological studies, so perhaps homoscedasticity is a myth, too.

One option to deal with outliers is to apply transformations, which convert original scores with a mathematical operation to new ones that may be more normally distributed. The effect of applying a **monotonic transformation** is to compress one part of the distribution more than another, thereby changing its shape but not the rank order of the scores. Examples of transformations that may remedy positive skew include $X^{1/2}$, $\log_{10} X$, and odd-root functions (e.g., $X^{1/3}$). There are many other kinds, and this is one of their potential problems: It can be difficult to find a transformation that works in a particular data set. Some distributions can be so severely nonnormal that basically no transformation will work. The scale of the original scores is lost when scores are transformed. If that scale is meaningful, the loss of the scaling metric creates no advantage but exacts the cost that the results may be difficult (or impossible) to interpret.

An alternative that also deals with departures from distributional assumptions is robust estimation. **Robust (resistant) estimators** are generally less affected than least squares estimators by outliers or nonnormality.

An estimator's quantitative robustness can be described by its **finite-sample breakdown point** (BP), or the smallest proportion of scores that when made arbitrarily very large or small renders the statistic meaningless. The lower the value of BP, the less robust the estimator. For both M and s^2, BP = 0, the lowest possible value. This is because the value of either statistic can be distorted by a single outlier, and the ratio $1/N$ approaches zero as sample size increases. In contrast, BP = .50 for the median because its value is not distorted by arbitrarily extreme scores unless they make up at least half the sample. But the median is not an optimal estimator because its value is determined by a single score, the one at the 50th percentile. In this sense, all the other scores are discarded by the median.

A compromise between the sample mean and the median is the **trimmed mean**. A trimmed mean M_{tr} is calculated by (a) ordering the scores from lowest to highest, (b) deleting the same proportion of the most extreme scores from each tail of the distribution, and then (c) calculating the average of the scores that remain. The proportion of scores removed from each tail is p_{tr}. If $p_{tr} = .20$, for example, the highest 20% of the scores are deleted as are the lowest 20% of the scores. This implies that

1. the total percentage of scores deleted from the distribution is $2p_{tr} = 2(.20)$, or 40%;
2. the number of deleted scores is $2np_{tr} = .40n$, where n is the original group size; and
3. the number of scores that remain is $n_{tr} = n - 2np_{tr} = n - .40n$, where n_{tr} is the trimmed group size.

For an odd number of scores, round the product np_{tr} down to the nearest integer and then delete that number of scores from each tail of the distribution. The statistics M_{tr} and M both estimate μ without bias when the population distribution is symmetrical. But if that distribution is skewed, M_{tr} estimates the trimmed population mean μ_{tr}, which is typically closer to more of the observations than μ.

A common practice is to trim 20% of the scores from each tail of the distribution when calculating trimmed estimators. This proportion tends to maintain the robustness of trimmed means while minimizing their standard errors when sampling from symmetrical distributions; it is also supported by the results of computer simulation studies (Wilcox, 2012). Note that researchers may specify $p_{tr} < .20$ if outliers constitute less than 20% of each tail in the distribution or $p_{tr} > .20$ if the opposite is true. For 20% trimmed means, BP = .20, which says they are robust against arbitrarily extreme scores unless such outliers make up at least 20% of the sample.

A variability estimator more robust than s^2 is the **interquartile range**, or the positive difference between the score that falls at the 75th percen-

tile in a distribution and the score at the 25th percentile. Although BP = .25 for the interquartile range, it uses information from just two scores. An alternative that takes better advantage of the data is the **median absolute deviation** (MAD), the 50th percentile in the distribution of $|X - Mdn|$, the absolute differences between each score and the median. Because it is based on the median, BP = .50 for the MAD. This statistic does not estimate the population standard deviation σ, but the product of MAD and the scale factor 1.483 is an unbiased estimator of σ in a normal population distribution.

The estimator 1.483 (MAD) is part of a **robust method for outlier detection** described by Wilcox and Keselman (2003). The conventional method is to calculate for each score the normal deviate $z = (X - M)/s$, which measures the distance between each score and the mean in standard deviation units. Next, the researcher applies a rule of thumb for spotting potential outliers based on z (e.g., if $|z| > 3.00$, then X is a potential outlier). Masking, or the chance that outliers can so distort values of M or s that they cannot be detected, is a problem with this method. A more robust method is based on this decision rule applied to each score:

$$\frac{|X - Mdn|}{1.483\,(\text{MAD})} > 2.24 \tag{2.28}$$

The value of the ratio in Equation 2.28 is the distance between a score and the median expressed in robust standard deviation units. The constant 2.24 in the equation is the square root of the approximate 97.5th percentile in a central χ^2 distribution with a single degree of freedom. A potential outlier thus has a score on the ratio in Equation 2.28 that exceeds 2.24. Wilcox (2012) described additional robust detection methods.

A robust variance estimator is the **Winsorized variance** s^2_{Win}. (The terms *Winsorized* and *Winsorization* are named after biostatistician Charles P. Winsor.) When scores are Winsorized, they are (a) ranked from lowest to highest. Next, (b) the p_{tr} most extreme scores in the lower tail of the distribution are all replaced by the next highest original score that was not replaced, and (c) the p_{tr} most extreme scores in the upper tail are all replaced by the next lowest original score that was not replaced. Finally, (d) s^2_{Win} is calculated among the Winsorized scores using the standard formula for s^2 (Equation 2.3) except that squared deviations are taken from the **Winsorized mean** M_{Win}, the average of the Winsorized scores, which may not equal M_{tr} in the same sample. The statistic s^2_{Win} estimates the Winsorized population variance σ^2_{Win}, which may not equal σ^2 if the population distribution is nonnormal.

Suppose that $N = 10$ scores ranked from lowest to highest are as follows:

15 16 19 20 22 24 24 29 90 95

The mean and variance of these scores are $M = 35.40$ and $s^2 = 923.60$, both of which are affected by the extreme scores 90 and 95. The 20% trimmed mean is calculated by first deleting the lower and upper .20 (10) = 2 most extreme scores from each end of the distribution, represented next by the strikethrough characters:

~~15~~ ~~16~~ 19 20 22 24 24 29 ~~90~~ ~~95~~

Next, calculate the average based on the remaining 6 scores (i.e., 19–29). The result is $M_{tr} = 23.00$, which as expected is less than the sample mean, $M = 35.40$.

When one Winsorizes the scores for the same trimming proportion (.20), the two lowest scores in the original distribution (15, 16) are each replaced by the next highest score (19), and the two highest scores (90, 95) are each replaced by the next lowest score (29). The 20% Winsorized scores are listed next:

19 19 19 20 22 24 24 29 29 29

The Winsorized mean is $M_{Win} = 23.40$. The total sum of squared deviations of the Winsorized scores from the Winsorized mean is $SS_{Win} = 166.40$, and the degrees of freedom are 10 – 1, or 9. These results imply that the 20% Winsorized variance for this example is $s^2_{Win} = 166.40/9$, or 18.49. The variance of the original scores is greater (923.60), again as expected.

Robust Interval Estimation

The **Tukey–McLaughlin method** (Tukey & McLaughlin, 1963) to calculate robust confidence intervals for μ_{tr} based on trimmed means and Winsorized variances is described next. The standard error of M_{tr} is estimated in this method as

$$s_{TM} = \frac{s_{Win}}{(1 - 2p_{tr})\sqrt{n}} \tag{2.29}$$

For the example where

$$n = 10, p_{tr} = .20, s_{Win} = 18.49^{1/2} = 4.30, \text{ and } M_{tr} = 23.00$$

the standard error of the trimmed mean (23.00) is

$$s_{TM} = \frac{4.30}{[1 - 2(.20)]\sqrt{10}} = 2.266$$

The general form of a robust 100 $(1 - \alpha)$% confidence interval for μ_{tr} in this method is

$$M_{tr} \pm s_{TM}\left[t_{2\text{-tail},\alpha}(n_{tr} - 1)\right] \tag{2.30}$$

where n_{tr} is the number of scores that remain after trimming. For the example where $n = 10$ and $p_{tr} = .20$, the number of deleted scores is 4, so $n_{tr} = 6$. The degrees of freedom are thus $6 - 1 = 5$. The value of $t_{2\text{-tail}, .05}$ (5) is 2.571, so the robust 95% confidence interval for μ_{tr} is

$$23.00 \pm 2.266(2.571)$$

which defines the interval [17.17, 28.83]. It is not surprising that this robust interval is narrower than the conventional 95% confidence interval for μ calculated with the original scores, which is [13.66, 57.14]. (You should verify this result.)

A robust estimator of the standard error for the difference between independent trimmed means when not assuming homoscedasticity is part of the **Yuen–Welch procedure** (e.g., Yuen, 1974). Error variance of each trimmed mean is estimated as

$$w_i = \frac{s^2_{Win_i}(n_i - 1)}{n_{tr_i}(n_{tr_i} - 1)} \tag{2.31}$$

where $s^2_{Win_i}$, n_i, and n_{tr_i} are, respectively, the Winsorized variance, original group size, and effective group size after trimming in the ith group. The Yuen–Welch estimate for the standard error of M_{tr} may be somewhat more accurate than the estimate in the Tukey–McLaughlin method (Equation 2.29), but the two methods usually give similar values (Wilcox, 2012).

The Yuen–Welch standard error of $M_{tr1} - M_{tr2}$ is

$$s_{YW} = \sqrt{w_1 - w_2} \tag{2.32}$$

and the adjusted degrees of freedom in a central t distribution are estimated as

$$df_{YW} = \frac{(w_1 + w_2)^2}{\frac{w_1^2}{n_{tr1} - 1} + \frac{w_2^2}{n_{tr2} - 1}} \tag{2.33}$$

TABLE 2.4
Raw Scores With Outliers and Descriptive Statistics for Two Groups

	Group	
	1	2
	15	3
	16	2
	19	21
	20	18
	22	16
	24	16
	24	13
	28	19
	90	20
	95	82
M	35.40	21.00
M_{tr}	23.00	17.00
M_{Win}	23.40	16.80
s^2	923.600	503.778
s^2_{Win}	18.489	9.067

Note. The trimming proportion is $p_{tr} = .20$.

The general form of a 100 $(1 - \sigma)$% confidence interval for $\mu_{tr1} - \mu_{tr2}$ in this method is

$$M_{tr1} - M_{tr2} \pm s_{YW} \left[t_{2\text{-tail}, \alpha} (df_{YW}) \right] \tag{2.34}$$

Listed in Table 2.4 are raw scores with outliers and descriptive statistics for two groups where $n = 10$. The trimming proportion is $p_{tr} = .20$, so $n_{tr} = 6$ in each group. Outliers in both groups inflate variances relative to their robust counterparts (e.g., $s_2^2 = 503.78$, $s^2_{Win2} = 9.07$). Extreme scores in group 2 (2, 3, 82) fall in both tails of the distribution, so nonrobust versus robust estimates of central tendency are more similar ($M_2 = 21.00$, $M_{tr2} = 17.00$) than in group 1. Exercise 5 asks you to verify the robust estimators for group 2 in Table 2.4.

Summarized next are robust descriptive statistics for the data in Table 2.4:

$$M_{tr1} = 23.00, s^2_{Win1} = 18.489 \quad \text{and} \quad M_{tr2} = 17.00, s^2_{Win2} = 9.067$$

$$M_{tr1} - M_{tr2} = 6.00$$

The standard error of the trimmed mean contrast is estimated in the Yuen–Welch method as

$$w_1 = \frac{18.489\,(9)}{6\,(5)} = 5.547 \quad \text{and} \quad w_2 = \frac{9.067\,(9)}{6\,(5)} = 2.720$$

$$s_{YW} = \sqrt{5.547 + 2.720} = 2.875$$

and the degrees of freedom are calculated as

$$df_{YW} = \frac{(5.547 + 2.720)^2}{\frac{5.547^2}{5} + \frac{2.720^2}{5}} = 8.953$$

The value of $t_{2\text{-tail, }.05}$ (8.953) is 2.264. The robust 95% confidence interval for $\mu_{tr1} - \mu_{tr2}$ is

$$6.00 \pm 2.875\,(2.264)$$

which defines the interval [–.51, 12.51]. Thus, $\mu_{tr1} - \mu_{tr2}$ could be as low as –.51 or it could be as high as 12.51, with 95% confidence and not assuming homoscedasticity. Wilcox (2012) described a robust version of the Welch procedure that is an alternative to the Yuen–Welch method, and Keselman, Algina, Lix, Wilcox, and Deering (2008) outlined robust methods for dependent samples.

A modern alternative in robust estimation to relying on formulas to estimate standard errors and degrees of freedom in central test distributions that assume normality is bootstrapping. There are methods to construct robust nonparametric bootstrapped confidence intervals that protect against repeated selection of outliers in the same generated sample (Salibián-Barrera & Zamar, 2002). Otherwise, bootstrapping is applied in basically the same way as described in the previous section but to generate empirical sampling distributions for robust estimators.

Standard computer programs for general statistical analyses, such as SPSS and SAS/STAT, have limited capabilities for robust estimation. Wilcox (2012) described add-on modules (packages) for conducting robust estimation in R, a free, open source computing environment for statistical analyses, data mining, and graphics.[10] It runs on Unix, Microsoft Windows, and Apple Macintosh families of operating systems. A basic R installation has about the same capabilities as some commercial statistical programs, but there are now over 2,000 packages that further extend its capabilities. Wilcox's (2012) WRS package has routines for robust estimation, outlier detection, comparisons, and confidence interval

[10]http://www.r-project.org/

construction in a variety of univariate or multivariate designs.[11] Additional R packages for robust estimation are available from the Institut universitaire de médecine sociale et préventive (IUMSP).[12] See Erceg-Hurn and Mirosevich (2008) for more information about robust estimation.

CONCLUSION

The basic logic of sampling and estimation was described in this chapter. Confidence intervals based on statistics with simple distributions rely on central test statistics, but statistics with complex distributions may follow noncentral distributions. Special software tools are typically needed for noncentrality interval estimation. The lower and upper bounds of a confidence interval set reasonable limits for the value of the corresponding parameter, but there is no guarantee that a specific confidence interval contains the parameter. Literal interpretation of the percentages associated with a confidence interval assumes random sampling and that all other sources of imprecision besides sampling error are nil. Interval estimates are better than point estimates because they are, as the astronomer Carl Sagan (1996, pp. 27–28) described them, "a quiet but insistent reminder that no knowledge is complete or perfect." Methods for robust interval estimation based on trimmed means and Winsorized variances were introduced. The next chapter deals with the logic and illogic of significance testing.

LEARN MORE

Cumming (2012) gives clear introductions to interval estimation, effect size estimation, and meta-analysis. Chernick (2008) describes bootstrapping methods for estimation, forecasting, and simulation. The accessible book by Wilcox (2003) gives more detail about robust statistics.

Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers* (2nd ed.). Hoboken, NJ: Wiley.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. New York, NY: Academic Press.

[11]http://dornsife.usc.edu/labs/rwilcox/software/
[12]http://www.iumsp.ch/Unites/us/Alfio/msp_programmes.htm

EXERCISES

1. Explain the difference between the standard deviation s and the standard error s_M.
2. Interpret $s = 60.00$ and $s_M = 6.00$ for the same data set. What is the sample size?
3. For $M = 100.00$, $s = 9.00$, and $N = 25$, show that the 99% confidence interval for μ is wider than the corresponding 95% interval.
4. For the data in Table 2.2, calculate the 95% confidence interval for μ_D and the 95% confidence interval for $\mu_1 - \mu_2$.
5. For the data in Table 2.4, verify the values of the robust estimators for group 2.
6. What is the relation between M_{tr} and M_{Win} in the Tukey–McLaughlin method?

This page intentionally left blank

3

LOGIC AND ILLOGIC OF SIGNIFICANCE TESTING

One more asterisk
To rest like eyes of dead fish—
Rigor mortis stars
—Stephen Ziliak and Deirdre McCloskey (2008, p. 87)

This chapter covers the logic of significance testing, including elements that make little sense in most studies. Poor practices are also outlined, such as the specification of arbitrary levels of statistical significance and the failure to estimate a priori power. It is emphasized that all statistical tests rely on assumptions that are generally unrealistic. They are also confounded measures of effect size and sample size. Robust statistical tests that depend on fewer distributional assumptions than do parametric tests are introduced, but robust tests do not solve major shortcomings of significance testing. Some of the topics reviewed next are relatively complicated, but they illustrate limitations of statistical tests with which you may be less familiar. Completing the exercises for this chapter will help you to manage this material.

DOI: 10.1037/14136-003
Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, Second Edition, by R. B. Kline

TWO SCHOOLS

Summarized in Table 3.1 are the basic steps of the Fisher and Neyman–Pearson approaches to statistical inference. In Fisher's method, p from a statistical test measures the strength of the evidence against the null hypothesis H_0. If p is sufficiently small, H_0 can be rejected. Fisher advocated $p < .05$ as a pragmatism for "small enough" (i.e., to reject H_0 due to a sufficiently low value of p) but not as a golden rule. There was no alternative hypothesis H_1 in Fisher's method. Specification of a fixed level of α (e.g., .05 or .01) and an explicit H_1 typifies the Neyman–Pearson model. These steps imply the distinction between Type I error (false rejection of H_0) and Type II error (false retention of H_0). The probability of a Type I error is α, and the likelihood of a Type II error is represented by β. Power is the complement of β, or $1 - \beta$, defined as the probability of correctly rejecting H_0 when H_1 is true. A power analysis estimates the probability of a Type II error as $\beta = 1 -$ power. Recall that power is the probability of getting a statistically significant result over random replications (in the long run) when H_1 is true.

Power analysis concerns a **loss function** for Type II error. A loss function estimates with a single number the cost of a specific decision error. Cost can be measured in monetary terms or in a different metric that represents loss of utility, or relative satisfaction, in some area. A loss function theoretically enables the researcher to weigh the consequences of low power (high β) against the risk of Type I error (α). This mental balancing act could facilitate a better understanding of implications for specifying $\alpha = .05$ versus $\alpha = .01$ (or some other value).

Fisher vehemently opposed loss functions because he believed that statistical inference must respect the pure aim of science, the accumulation and dissemination of knowledge. Entertaining any other consideration would,

TABLE 3.1
Steps in Fisher Significance Testing and Neyman–Pearson Hypothesis Testing

Fisher	Neyman–Pearson
1. State H_0.	1. State H_0 and H_1.
2. Specify test statistic.	2. Specify α (usually .05 or .01)
3. Collect data, calculate test statistic, determine p.	3. Specify test statistic and critical value(s).
4. Reject H_0 if p is small; otherwise, H_0 is retained.	4. Collect data, calculate test statistic, determine p.
	5. Reject H_0 in favor of H_1 if $p < \alpha$; otherwise, H_0 is retained.

Note. H_0 = null hypothesis; H_1 = alternative hypothesis.

in his view, sully the scientific process and censor the researcher in advance (e.g., Fisher, 1956, pp. 102–103). What is probably closer to the truth is that Fisher's rather acerbic personality was unable to tolerate any extension of his original work or share academic credit with Neyman, Pearson, Gosset, and others (see Ziliak & McCloskey, 2008, Chapters 21–22).

It was rarely mentioned in statistics textbooks from the 1960s–2000s that the Intro Stats method is the synthesis of two schools with contradictory elements. Instead, it was typically described in these works without mentioning Fisher, Neyman, or Pearson by name and with no citations at all. This anonymous style of presentation gave the false impression that the Intro Stats method is so universal that citation is unnecessary. Most books also failed to mention the significance testing controversy (Gliner, Leech, & Morgan, 2002), which also discouraged critical thinking on the part of students about what they are reading (Gigerenzer, 2004). Some more recent books do not make these mistakes (e.g., McGrath, 2011), but they are still too rare.

SENSE AND NONSENSE OF THE INTRO STATS MODEL

Emphasized next are aspects of significance testing that are not well understood by many students and researchers.

Null Hypotheses

The standard H_0 is both a point hypothesis and a nil hypothesis. A **point hypothesis** specifies the numerical value of a parameter or the difference between two or more parameters, and a **nil hypothesis** states that this value is zero. The latter is usually a prediction that an effect, difference, or association is zero. Examples of nil hypotheses are presented next:

$$H_0: \mu_1 - \mu_2 = 0 \qquad H_0: \mu_D = 0 \qquad H_0: \rho = 0$$

In contrast, a **non-nil hypothesis** asserts that an effect is not zero. Examples include

$$H_0: \mu_1 - \mu_2 = 5.00 \qquad H_0: \mu_D = 10.00 \qquad H_0: \rho = .30$$

Nil hypotheses as default explanations may be fine in new research areas when it is unknown whether effects exist at all. But they are less suitable in established areas when it is known that some effect is probably not zero. For example, gender differences in certain personality characteristics have remained fairly constant over time, although their magnitudes can

vary with age or context (Hyde, 2005). Specification of a nil hypothesis when measuring gender differences in such characteristics may set the bar too low.

There are also cases where nil hypotheses are indefensible, such as when testing score reliability coefficients for statistical significance. This is because "declaring a reliability coefficient to be nonzero constitutes the ultimate in stupefyingly vacuous information" (Abelson, 1997b, p. 13). Such coefficients should be interpreted in an absolute sense depending on the context (e.g., requiring $r_{XX} > .70$ for test–retest reliability). Nil hypotheses also usually make little sense for validity coefficients. It is more realistic to assume nonzero population correlations because, at some level, everything is related to everything else. Meehl (1990) referred to these expected nonzero associations as a **crud factor**. Although exact values of the crud factor are unknown, correlations may depart even further from zero for variables assessed with the same measurement method. Correlations that result from common method variance may be as high as .20 to .30 in absolute value. Validity coefficients should be interpreted in absolute ways, too, given the context of the study. For example, such coefficients are typically higher in cross-sectional studies than in longitudinal studies. What represents a "significant" (i.e., substantive) correlation depends on the research area, not on the results of a statistical significance test.

Nil hypotheses are tested much more often than non-nil hypotheses even when the former are implausible. Many researchers are unaware of the possibility of specifying non-nil hypotheses, but most statistical computer tools test only nil hypotheses. This means that such tests must be calculated by hand, but doing so is feasible only for simple hypotheses, such as H_0: $\mu_1 - \mu_2 = 10.00$, which can be evaluated without difficulty with the t test. If a nil hypothesis is implausible, estimated probabilities of data will be too low. This means that risk for Type I error is basically zero and a Type II error is the only possible kind when H_0 is known in advance to be false.

The most common context for significance testing is **reject-support testing**, where rejection of H_0 supports the researcher's theory. The opposite is true in **accept-support testing**, where the failure to reject H_0 supports the researcher's expectations (Steiger & Fouladi, 1997). An implication of this distinction is that statistical significance is not always good news for the researcher's hypotheses. Another is that accept-support tests are logically weak because lack of evidence to disprove an assertion does not prove that it is true. Low power can lead to failure to reject H_0, which in accept-support testing favors the researcher's hypothesis. In other words, the researcher is potentially "rewarded" for having a sample size that is too small (i.e., low power) in accept-support testing. In contrast, low power works against the researcher's hypothesis in reject-support testing.

Alternative Hypotheses

The standard H_1 is a **range hypothesis**. A **two-tailed (nondirectional) hypothesis** predicts any result not specified in H_0, but a **one-tailed (directional) hypothesis** predicts values on only one side of H_0. Given H_0: $\rho = 0$, for example, there is only one nondirectional alternative, H_1: $\rho \neq 0$, but there are two possible directional alternatives, H_1: $\rho > 0$ or H_1: $\rho < 0$. The form of H_1 is supposed to be specified before data are collected. The choice also affects the outcome. It is easier to reject H_0 when the data are consistent with a one-tailed H_1, and power is greater, too, if H_0 is actually false. If H_1 is directional but the data indicate an effect in the other direction, H_0 is retained though the results are very inconsistent with it. This rule is not always followed. Sometimes researchers switch from a nondirectional H_1 to a directional H_1 or from one directional H_1 to its opposite in order to reject H_0. Some would consider changing H_1 based on the data a kind of statistical sin to be avoided. Like those against other kinds of sin, such admonitions are not always followed.

Level of Type I Error

Alpha (α) is the probability of making a Type I error over random replications. It is also the conditional prior probability of rejecting H_0 when it is actually true, or

$$\alpha = p\,(\text{Reject } H_0 \,|H_0 \text{ true}) \tag{3.1}$$

Both descriptions are frequentist statements about the likelihood of Type I error.

Too many researchers treat the conventional levels of α, either .05 or .01, as golden rules. If other levels of α are specified, they tend to be even lower, such as .001. Sanctification of .05 as the highest "acceptable" level is problematic. In reject-support testing, where rejecting H_0 favors the researcher's theory, α should be as low as possible from the perspective of journal reviewers and editors, who may wish to guard against bogus claims (Type I error). But in accept-support testing, a greater worry is Type II error because false claims in this context arise from not rejecting H_0, which supports the researcher's theory. Insisting on low values of α in this case may facilitate publication of sham claims, especially when power is low.

Instead of blindly accepting either .05 or .01, one does better to follow Aguinis et al.'s (2010) advice: Specify a level of α that reflects the **desired relative seriousness** (DRS) of Type I error versus Type II error. Suppose a

researcher will study a new treatment for a disorder where no extant treatment is effective. The researcher decides that the risk of a Type II error should be no more than .10. A Type II error in this context means that the treatment really makes a difference, but the null hypothesis of no difference is not rejected. This low tolerance for Type II error reflects the paucity of good treatment options. It is also decided that the risk of a Type I error is half as serious as that of making a Type II error, so DRS = .50.

The desired level of α is computed as

$$\alpha_{des} = \left(\frac{p(H_1)\beta}{1 - p(H_1)} \right) \left(\frac{1}{\text{DRS}} \right) \tag{3.2}$$

where $p\ (H_1)$ is the **prior probability** that the alternative hypothesis is true. This probability could be established rationally based on theory or the researcher's experience, or it could be estimated in a Bayesian analysis given results from prior studies. Suppose the researcher estimates that $p\ (H_1) = .60$ for the example where $\beta = .10$ and DRS = .50. The desired level of α is

$$\alpha_{des} = \left(\frac{.60\,(.10)}{1 - .60} \right) \left(\frac{1}{.50} \right) = .30$$

which says that $\alpha = .30$ reflects the desired balance of Type I versus Type II error. The main point is that researchers should not rely on a mechanical ritual (i.e., automatically specify .05 or .01) to control risk for Type I error that ignores the consequences of Type II error. Note that the estimate of $p\ (H_1)$ could come from a Bayesian analysis based on results of prior studies. In this case, the form of the probability that H_1 is true would be that of the conditional probability $p\ (H_1 | \text{Data})$, where "Data" reflects extant results and the whole conditional probability is estimated with Bayesian methods (Chapter 10).

The level of α sets the risk of Type I error for a single test. There is also **experimentwise (familywise) error rate**, or the likelihood of making at least one Type I error across a set of tests. If each individual test is conducted at the same level of α, then

$$\alpha_{ew} = 1 - (1 - \alpha)^c \tag{3.3}$$

where c is the number of tests. Suppose that 20 statistical tests are conducted, each at $\alpha = .05$. The experimentwise error rate is

$$\alpha_{ew} = 1 - (1 - .05)^{20} = .64$$

which says that the risk of making a Type I error across all tests is .64. Equation 3.3 assumes independent hypotheses or outcomes; otherwise, the estimate of .64 is too low. This result is the probability of one or more Type I errors, but it does not indicate how many errors may have been committed (it could be 1, or 2, or 3 . . .) or on which tests they occurred.

Experimentwise Type I error is controlled by reducing the number of tests or specifying a lower α for each one. The former is realized by prioritizing hypotheses (i.e., testing just the most important ones), which means that "fishing expeditions" in data analysis are to be avoided. Another way is to use multivariate techniques that can test hypotheses across several variables at once. The **Bonferroni correction** is a simple method to set α for individual tests: Just divide a target value of α_{ew} by the total number of tests; the result is α_{Bon}. Suppose a researcher wishes to limit the experimentwise error rate to .30 for a set of 20 tests. Thus, $\alpha_{Bon} = .30/20 = .015$, which is the level of α for each individual test. Not all methodologists believe that controlling experimentwise Type I error is generally a desirable goal, especially when power is already low in reject-support testing at the conventional levels of statistical significance, such as .05.

A danger of conducting too many significance tests is **HARKing** (hypothesizing after the results are known). This happens when the researcher keeps testing until H_0 is rejected—which is practically guaranteed by the phenomenon of experimentwise error—and then positions the paper as if those findings were the object of the study (Ellis, 2010). Austin, Mamdani, Juurlink, and Hux (2006) demonstrated how testing multiple hypotheses not specified in advance increases the likelihood of discovering implausible associations. Using a database of about 10 million cases, they conducted statistical tests until finding 2 out of 200 disorders for which people born under particular astrological signs had significantly higher base rates. These associations "disappeared" (were no longer statistically significant) after controlling for multiple comparisons.

Another argument against using statistical tests to "snoop" for results by checking for significance is **Feynman's conjecture**, named after the physicist and Nobel laureate Richard Feynman (Gigerenzer, 2004). It is the assertion that the $p < \alpha$ is meaningful only when hypotheses are specified *before* the data are collected. Otherwise, significance testing capitalizes on chance, which is not taken into account by p values. That is, the appropriate way to test a pattern found by accident in one sample is to repeat the analysis in a replication sample with new cases. Only in the latter case would p for the test of the original pattern be meaningful. A related misuse is **testing to a foregone conclusion**, or the practice of collecting data until $p < .05$ is observed. The problem with this tactic is that the actual rate of Type I error for the last statistical test is > .05, and the reason is Feynman's conjecture. A better

stopping rule for data collection is a minimum sample size needed to attain a target level of power (e.g., ≥ .80) in a particular study. Power analysis is described later.

p Values

All statistical tests do basically the same thing: The difference between a sample result and the value of the corresponding parameter(s) specified in H_0 is divided by the estimated sampling error, and this ratio is then summarized as a test statistic (e.g., t, F, χ^2). That ratio is converted by the computer to a probability based on a theoretical sampling distribution (i.e., random sampling is assumed). Test probabilities are often printed in computer output under the column heading *p*, which is the same abbreviation used in journal articles. You should not forget that *p* actually stands for the conditional probability

$$p(\text{Data} + \mid H_0 \text{ and all other assumptions})$$

which represents the likelihood of a result or outcomes even more extreme (Data +) assuming

1. the null hypothesis is exactly true;
2. the sampling method is random sampling;
3. all distributional requirements, such as normality and homoscedasticity, are met;
4. the scores are independent;
5. the scores are also perfectly reliable; and
6. there is no source of error besides sampling or measurement error.

In addition to the specific observed result, *p* values reflect outcomes never observed and require many assumptions about those unobserved data. If any of these assumptions are untenable, *p* values may be inaccurate. If *p* is too low, there is positive bias, and (a) H_0 is rejected more often than it should be and (b) the nominal rate of Type I error is higher than the stated level of α. Negative bias means just the opposite—*p* is too high—and it also reduces statistical power because it is now more difficult to reject the null hypothesis.

Although *p* and α are derived in the same theoretical sampling distribution, *p* does *not* estimate the conditional probability of a Type I error (Equation 3.1). This is because *p* is based on a range of results under H_0, but α has nothing to do with actual results and is supposed to be specified before any

data are collected. Confusion between p and α is widespread (e.g., Hubbard, Bayarri, Berk, & Carlton, 2003). To differentiate the two, Gigerenzer (1993) referred to p as the **exact level of significance**. If $p = .032$ and $\alpha = .05$, H_0 is rejected at the .05 level, but .032 is not the long-run probability of Type I error, which is .05 for this example.

The exact level of significance is the conditional probability of the data (or any result even more extreme) assuming H_0 is true, given all other assumptions about sampling, distributions, and scores. Some other correct interpretations for $p < .05$ are listed next:

1. Suppose the study were repeated many times by drawing many random samples from very large population(s) where H_0 is true. Less than 5% of these hypothetical results would be even more inconsistent with H_0 than the actual result.
2. Less than 5% of test statistics from many random samples are further away from the mean of the sampling distribution under H_0 than the one for the observed result.

That is about it; other correct definitions may just be variations of those listed. The range of correct interpretations of p is thus actually narrow. It also depends on many assumptions about idealized sampling methods or measurement that do not apply in most studies.

Because p values are estimated assuming that H_0 is true, they do not somehow measure the likelihood that H_0 is correct. At best they provide only indirect evidence against H_0, but some statisticians object to even this mild characterization (e.g., Schervish, 1996). The false belief that p is the probability that H_0 is true, or the inverse probability error (see Chapter 1), is widespread. Many other myths about p are described in the next chapter.

Cumming (2008) studied **prediction intervals for *p***, which are intervals with an 80% chance of including p values from random replications. Summarized next for designs with two independent samples are computer simulation results for one-tailed prediction intervals for one-tailed replications, or results that are in the same direction as the initial study. The lower bound is 0, and the upper bound is the 80th percentile in the sampling distribution of p values from replications. This interval contains the lower 80% of p values in the sampling distribution and the rest, or 20%, exceed the upper bound. The one-tailed prediction interval for $p = .05$ regardless of sample size is (0, .22), so 80% of replication p values are between 0 and .22, but 20% exceed .22. The corresponding interval for $p = .01$ is (0, .083), which is narrower than that for $p = .05$ but still relatively wide. Cumming (2008) concluded that p values generally provide unreliable information about what is likely to happen in replications unless p is very low, such as $< .001$.

Probabilities from significance tests say little about effect size. This is because essentially any test statistic (TS) can be expressed as the product

$$TS = ES \times f(N) \tag{3.4}$$

where ES is an effect size and $f(N)$ is a function of sample size. This equation explains how it is possible that (a) trivial effects can be statistically significant in large samples or (b) large effects may not be statistically significant in small samples. So p is a confounded measure of effect size and sample size. Statistics that directly measure effect size are introduced in Chapter 5.

POWER

Power is the probability of getting statistical significance over many random replications when H_1 is true. It varies directly with sample size and the magnitude of the population effect size. Other factors that influence power include

1. the level of statistical significance (e.g., $\alpha = .05$ vs. $\alpha = .01$);
2. the directionality of H_1 (i.e., directional vs. nondirectional);
3. whether the design is between-subjects or within-subjects;
4. the particular test statistic used; and
5. the reliability of the scores.

This combination leads to the greatest power: a large population effect size, a large sample, a higher level of α (e.g., .05 instead of .01), a within-subjects design, a parametric test rather than a nonparametric test (e.g., t instead of Mann–Whitney), and very reliable scores.

A computer tool for power analysis estimates the probability of rejecting H_0, given specifications about population effect size and study characteristics. Power can be calculated for a range of estimates about population effect size or study characteristics, such as sample size. The resulting **power curves** can then be compared. A variation is to specify a desired level of power and then estimate the minimum sample size needed to obtain it.

There are two kinds of power analysis, proper and improper. The former is a **prospective (a priori) power analysis** conducted before the data are collected. Some granting agencies require a priori power analyses. This is because if power is low, it is unlikely that the expected effect will be detected, so why waste money? Power $\geq .80$ is generally desirable, but an even higher standard may be need if consequences of Type II error are severe. There are some free power analysis computer tools, including G*Power 3 (Faul, Erdfelder, Lang,

& Buchner, 2007)[1] and Lenth's (2006–2009) online Java applets for power analysis.[2]

The improper kind is a **retrospective (post hoc, observed) power analysis** conducted after the data are collected. The effect size observed in the sample is treated as the population effect size, and the computer estimates the probability of rejecting H_0, given the sample size and other characteristics of the analysis. Post hoc power is inadequate for a few reasons (Ellis, 2010; O'Keefe, 2007). Observed effect sizes are unlikely to equal corresponding population effect sizes. The *p* values from statistical tests vary inversely with power, so if results are not statistically significant, observed power must be low. Low post hoc power suggests that the sample is too small, but the researcher should have known better in the first place. A retrospective power analysis is more like an autopsy conducted after things go wrong than a diagnostic procedure (i.e., a priori power analysis). Do not bother to report observed power.

Reviews from the 1970s and 1980s indicated that the typical power of behavioral science research is only about .50 (e.g., Sedlmeier & Gigerenzer, 1989), and there is little evidence that power is any higher in more recent studies (e.g., Brock, 2003). Ellis (2010) estimated that < 10% of studies have samples sufficiently large to detect smaller population effect sizes. Increasing sample size would address low power, but the number of additional cases necessary to reach even nominal power when studying smaller effects may be so great as to be practically impossible (F. L. Schmidt, 1996). Too few researchers, generally < 20% (Osborne, 2008), bother to report prospective power despite admonitions to do so (e.g., Wilkinson & the TFSI, 1999).

The concept of power does not stand without significance testing. As statistical tests play a smaller role in the analysis, the relevance of power also declines. If significance tests are not used, power is irrelevant. Cumming (2012) described an alternative called **precision for research planning**, where the researcher specifies a target margin of error for estimating the parameter of interest. Next, a computer tool, such as ESCI (see footnote 4, Chapter 2), is used to specify study characteristics before estimating the minimum sample size needed to meet the target. The advantage over power analysis is that researchers must consider both effect size and precision in study planning.

Reviewed next are the *t* and *F* tests for means and the χ^2 test for two-way contingency tables. Familiarity with these basic test statistics will help you to better appreciate limitations of significance testing. It is also possible in many cases to calculate effect sizes from test statistics, so learning about *t*, *F*, and χ^2 gives you a head start toward understanding effect size estimation.

[1]http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/
[2]http://www.stat.uiowa.edu/~rlenth/Power/

t TESTS FOR MEANS

The *t* tests reviewed next compare means from either two independent or two dependent samples. Both are special cases of the *F* test for means such that $t^2 = F$ for the same contrast and a nil hypothesis. The general form of *t* for independent samples is

$$t(N-2) = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{M_1 - M_2}} \tag{3.5}$$

where $N - 2$ are the pooled within-groups degrees of freedom df_W, $M_1 - M_2$ and $s_{M_1 - M_2}$ are, respectively, the observed mean contrast and its standard error (Equation 2.12), and $\mu_1 - \mu_2$ is the population contrast specified in H_0. For a nil hypothesis, $\mu_1 - \mu_2 = 0$.

Suppose that patients given an established treatment score on average 10 points more than control cases on an outcome where higher scores are better. A new treatment is devised that is hoped to be even more effective; otherwise, there is no need to abandon the old treatment. If population 1 corresponds to the new treatment and population 2 is control, the non-nil hypothesis

$$H_0: \mu_1 - \mu_2 = 10.00$$

is more appropriate than the nil hypothesis

$$H_0: \mu_1 - \mu_2 = 0$$

because the effect of the old treatment is not zero. Results from a study of the new treatment are

$$M_1 - M_2 = 15.00, n_1 = n_2 = 25, s_{M_1 - M_2} = 4.50$$

For a two-tailed H_1, results of the *t* test for the two null hypotheses are

$$t_{\text{non-nil}}(48) = \frac{15.00 - 10.00}{4.50} = 1.11, p = .272$$

$$t_{\text{nil}}(48) = \frac{15.00}{4.50} = 3.33, p = .002$$

This example illustrates the principle that the relative rareness of data under implausible null hypotheses compared with more null plausible hypotheses is exaggerated (respectively, .002 vs. .272). This is why Rouder, Speckman, Sun, and Morey (2009) wrote, "As a rule of thumb, hypothesis testing should be reserved for those cases in which the researcher will entertain the null as theoretically interesting and plausible, at least approximately" (p. 235).

Computer tools for statistical analyses generally assume nil hypotheses and offer no option to specify non-nil hypotheses. It is no problem to calculate the *t* test by hand for non-nil hypotheses, but it is practically impossible to do so for many other tests (e.g., *F*). It is ironic that modern computer tools for statistics are so inflexible in their nil-hypothesis-centric focus.

Power of the independent samples *t* test is greatest in balanced designs with the same number of cases in each group. There is loss of power in unbalanced designs even if the total number of cases is equal for a balanced versus unbalanced design. Rosenthal, Rosnow, and Rubin (2000) showed that the power loss for an unbalanced design where $n_1 = 70$ and $n_2 = 30$ is equivalent to losing 16 cases (16% of the sample) from the balanced design where $n_1 = n_2 = 50$. Relative power generally decreases as the group size disparity increases in unbalanced designs.

The form of the *t* test for a dependent mean contrast is

$$t(n-1) = \frac{M_D - \mu_D}{s_{M_D}} \tag{3.6}$$

where the degrees of freedom are the group size (n) minus 1, M_D and s_{M_D} are, respectively, the observed mean difference score and its standard error (Equation 2.20), and μ_D is the population dependent mean contrast specified in H_0. The latter is zero for a nil hypothesis.

Assuming a nil hypothesis, both forms of the *t* test defined express a contrast as the proportion of its standard error. If $t = 1.50$, for example, the first mean is 1½ standard errors higher than the second, but the sign of *t* is arbitrary because it depends on the direction of subtraction. You should know that the standard error metric of *t* is affected by sample size. Suppose descriptive statistics for two groups in a balanced design are

$$M_1 = 13.00,\ s_1^2 = 7.50 \quad \text{and} \quad M_2 = 11.00,\ s_2^2 = 5.00$$

which imply $M_1 - M_2 = 2.00$. Reported in Table 3.2 are results of the independent samples *t* test for these data at three different group sizes, $n = 5$, 15, and 30. Note that the pooled within-groups variance, $s_{\text{pool}}^2 = 6.25$ (Equation 2.13), is unaffected by group size. This is not true for the denominator of *t*, $s_{M_1 - M_2}$, which gets smaller as *n* increases. This causes the value of *t* to go up and its *p* value to go down for the larger group sizes. Consequently, the test for $n = 5$ is not statistically significant at $p < .05$, but it is for the larger group sizes. Results for the latter indicate less sampling error but not a larger effect size. Exercise 1 asks you to verify the results in Table 3.2 for $n = 15$.

TABLE 3.2
Results of the Independent Samples *t* Test at Three Different Group Sizes

	Group size (*n*)		
Statistic	5	15	30
$s_{M_1-M_2}$	1.581	.913	.645
t	1.26	2.19	3.10
df_W	8	28	58
p	.242	.037	.003
$t_{2\text{-tail, }.05}$	2.306	2.048	2.002
95% CI for $\mu_1 - \mu_2$	–1.65, 5.65	.13, 3.87	.71, 3.29

Note. For all analyses, $M_1 = 13.00$, $s_1^2 = 7.50$, $M_2 = 11.00$, $s_2^2 = 5.00$, $s_{pool}^2 = 6.25$, and *p* values are two-tailed and for a nil hypothesis. CI = confidence interval.

The standard error metric of *t* is also affected by whether the means are independent or dependent. Look back at Table 2.2, which lists raw scores and descriptive statistics for two samples where $M_1 - M_2 = 2.00$, $s_1^2 = 7.50$, and $s_2^2 = 5.00$. Summarized next for these data are results of the *t* test and confidence intervals assuming $n = 5$ in each of two independent samples:

$$M_1 - M_2 = 2.00,\ s_{M_1-M_2} = 1.581,\ t_{ind}\,(8) = 1.26,\ p = .242$$
$$95\%\ \text{CI for}\ \mu_1 - \mu_2\ \ [-1.65, 5.65]$$

Results for the same data but now assuming $n = 5$ pairs of scores across dependent samples are

$$M_D = 2.00,\ r_{12} = .735,\ s_{M_D} = .837,\ t_{dep}\,(4) = 2.39,\ p = .075$$
$$95\%\ \text{CI for}\ \mu_D\ [-.32, 4.32]$$

Note the smaller standard error, higher value of *t* and its lower *p* value, and the narrower 95% confidence interval in the dependent samples analysis relative to the independent samples analysis of the same raw scores. The assumptions of the *t* tests are the same as those of the independent samples *F* test, which are considered in the next section.

The **Welch *t* test**, also called the **Welch–James *t* test** (e.g., James, 1951), for independent samples assumes normality but not homoscedasticity. Its equation is

$$t_{Wel}\,(df_{Wel}) = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{Wel}} \qquad (3.7)$$

where s_{Wel} and the estimated degrees of freedom df_{Wel} are defined by, respectively, Equations 2.15 and 2.16. The Welch *t* test is generally more accurate than the standard *t* test when the population distributions are heteroscedastic but normal (Keselman et al., 2008). But neither test may be accurate when the population distributions are not normal, the group sizes are both small and unequal, or there are outliers.

F TESTS FOR MEANS

The *t* test analyzes **focused comparisons** (contrasts) between two means. A contrast is a single-*df* effect that addresses a specific question, such as whether treatment and control differ. The *F* test can analyze focused comparisons, too ($t^2 = F$ for a contrast). But only *F* can also be used in **omnibus comparisons** that simultaneously compare at least three means for equality.

Suppose factor A has $a = 3$ levels. Its omnibus effect has two degrees of freedom ($df_A = 2$), and the hypotheses tested by *F* for this effect are listed next:

$$H_0: \mu_1 = \mu_2 = \mu_3 \quad \text{and} \quad H_1: \mu_1 \neq \mu_2 \neq \mu_3 \text{ (i.e., not } H_0)$$

Rejecting H_0 says only that differences among M_1, M_2, and M_3 are unlikely. This result alone is not very informative. A researcher may be more interested in focused comparisons, such as whether each of two treatment conditions differs from control, which break down the omnibus effect into specific directional effects. Thus, it is common practice either to follow an omnibus comparison with contrasts or to forgo the omnibus test and analyze only contrasts. The logic of the *F* test in single-factor designs with $a \geq 3$ levels is considered next. Chapter 7 addresses contrast analysis in such designs, and Chapter 8 covers designs with multiple factors.

Independent Samples

The general form of the *F* test for independent samples is

$$F(df_A, df_W) = \frac{MS_A}{MS_W} \tag{3.8}$$

where $df_A = a - 1$ and df_W are the pooled within-groups degrees of freedom, or

$$df_W = \sum_{i=1}^{a} df_i = \sum_{i=1}^{a} (n_i - 1) = N - a \tag{3.9}$$

The numerator is the between-groups mean square. Its equation is

$$MS_A = \frac{SS_A}{df_A} = \frac{\sum_{i=1}^{a} n_i(M_i - M_T)^2}{a-1} \tag{3.10}$$

where SS_A is the between-groups sum of squares, n_i and M_i are, respectively, the size and mean of the ith group, and M_T is the grand mean. This term reflects group size and sources of variation that lead to unequal group means, including sampling error or a real effect of factor A.

The denominator of F is the pooled within-groups variance MS_W, and it measures only error variance. This is because cases within each group are all treated the same, so variation of scores around group means has nothing to do with any effect of the factor. This error term is not affected by group size because functions of n appear in both its numerator and its denominator,

$$MS_W = \frac{SS_W}{df_W} = \frac{\sum_{i=1}^{a} df_i(s_i^2)}{\sum_{i=1}^{a} df_i} \tag{3.11}$$

where s_i^2 is the variance of the ith group. If there are only two groups, $MS_W = s_{\text{pool}}^2$, and only in a balanced design can MS_W also be computed as the average of the within-groups variances. The total sum of squares SS_T is the sum of SS_A and SS_W; it can also be computed as the sum of squared deviations of individual scores from the grand mean.

Presented next are descriptive statistics for three groups:

$M_1 = 13.00, s_1^2 = 7.50$ $\quad$ $M_2 = 11.00, s_2^2 = 5.00$ $\quad$ $M_3 = 12.00, s_3^2 = 4.00$

Reported in Table 3.3 are the results of F tests for these data at group sizes $n = 5$, 15, and 30. Note in the table that $MS_W = 5.50$ regardless of group size. But both MS_A and F increase along with the group size, which also progressively lowers p values from .429 for $n = 5$ to .006 for $n = 30$. Exercise 2 asks you to verify results in Table 3.3 for $n = 30$.

Equation 3.10 for MS_A defines a **weighted means analysis** where squared deviations of group means from the grand mean are weighted by group size. If the design is unbalanced, means from bigger groups get more weight. This may not be a problem if unequal group sizes reflect unequal population base rates. Otherwise, an **unweighted means analysis** may be preferred where all means are given the same weight by (a) computing the grand mean as the

TABLE 3.3
Results of the Independent Samples *F* Test at Three Different Group Sizes

Source	*SS*	*df*	*MS*	*F*
		$n = 5$		
Between (*A*)	10.00	2	5.00	.91[a]
Within (error)	66.00	12	5.50	
Total	76.00	14		
		$n = 15$		
Between (*A*)	30.00	2	15.00	2.73[b]
Within (error)	231.00	42	5.50	
Total	261.00	44		
		$n = 30$		
Between (*A*)	60.00	2	30.00	5.45[c]
Within (error)	478.50	87	5.50	
Total	538.50	89		

Note. For all analyses, $M_1 = 13.00$, $s_1^2 = 7.50$, $M_2 = 11.00$, $s_2^2 = 5.00$, $M_3 = 12.00$, and $s_3^2 = 4.00$.
[a]$p = .429$. [b]$p = .077$. [c]$p = .006$.

simple arithmetic average of the group means and (b) substituting the harmonic mean n_h for the actual group sizes in Equation 3.10:

$$n_h = \frac{a}{\sum_{i=1}^{a} \frac{1}{n_i}} \tag{3.12}$$

Results of weighted versus unweighted analysis for the same data tend to diverge as group sizes are increasingly unbalanced.

The assumptions of the *t* tests are the same as for the independent samples *F* test. They are stated in many introductory books as independence, normality, and homoscedasticity, but there are actually more. Two are that (a) the factor is fixed and (b) all its levels are represented in the study. Levels of **fixed effects factors** are intentionally selected for investigation, such as the equally spaced drug dosages 0 (control), 3, 6, 9, and 12 mg · kg^{-1}. Because these levels are not randomly selected, the results may not generalize to other dosages not studied, such as 15 mg · kg^{-1}. Levels of **random effects factors** are randomly selected, which yields over replications a representative sample from all possible levels. A **control factor** is a special kind of random factor that is not itself of interest but is included for the sake of generality (Keppel & Wickens, 2004). An example is when participants are randomly assigned to receive different versions of a vocabulary test. Using different word lists

may enhance generalizability compared with using a single list. Designs with random factors are dealt with in Chapters 7 and 8.

A third additional requirement is that the factor affects only means; that is, it does not also change the shapes or variances of distributions. Some actual treatments affect both means and variances, including certain medications for high blood pressure (Webb, Fischer, & Rothwell, 2011). Such a pattern could arise due to a nonadditive effect where treatment does not have the same efficacy for all cases. For example, a drug may be more effective for men than women. A conditional treatment effect in human studies is called a **person × treatment interaction**. Interactions can be estimated in factorial designs but not in single-factor designs, because there is no systematic effect other than that of the sole factor. Altogether, these additional requirements are more restrictive than many researchers realize.

It is beyond the scope of this section to review the relatively large literature about consequences of violating the assumptions of the *F* test (e.g., Glass, Peckham, & Sanders, 1972; Keppel & Wickens, 2004; Winer, Brown, & Michels, 1991), so this summary is selective. The independence assumption is critical because nonindependence can seriously bias *p* values. Too many researchers believe that the normality and homoscedasticity assumptions can be violated with relative impunity. But results by Wilcox (1998) and Wilcox and Keselman (2003), among others, indicated that (a) even small departures from normality can seriously distort the results and (b) the combination of small and unequal group sizes and homoscedasticity can have similar consequences. Outliers can also distort outcomes of the *F* test.

Dependent Samples

The variances MS_A and MS_W are calculated the same way regardless of whether the samples are independent or dependent (Equations 3.10–3.11), but the latter no longer reflects only error variance in correlated designs. This is due to the subjects effect. It is estimated for factors with ≥ 3 levels as M_{cov}, the average covariance over all pairs of conditions. The subtraction $MS_W - M_{cov}$ literally removes the subjects effect from the pooled within-conditions variance and also defines the error term for the dependent samples *F* test. A similar subtraction removes the subjects effect from the error term of the dependent samples *t* test (Equation 2.21).

An **additive model** assumes that the quantity $MS_W - M_{cov}$ reflects only sampling error. In some sources, this error term is designated as MS_{res}, where the subscript refers to residual variance after removal of the subjects effect. A **nonadditive model** assumes that the error term reflects both random error and a true person × treatment interaction where some unmeasured

characteristic of cases either amplifies or diminishes the effect of factor A for some, but not all, cases. The error term for a nonadditive model may be called $MS_{A\times S}$, where the subscript reflects this assumption. Unfortunately, it is not possible to separately estimate variability due to random error versus true person × treatment interaction when each case is measured just once in each condition in a single-factor design. This implies that $MS_{res} = MS_{A\times S}$ in the same data set, so the distinction between them for now is more conceptual than practical. In Chapter 7, I consider cases where assumptions of additive versus nonadditive models in correlated designs can make a difference in effect size estimation.

For a nonadditive model, the general form of the dependent samples *F* test is

$$F(df_A, df_{A\times S}) = \frac{MS_A}{MS_{A\times S}} \tag{3.13}$$

where $df_{A\times S} = (a - 1)\,(n - 1)$ and $MS_{A\times S} = MS_W - M_{cov}$. The latter can also be expressed as

$$MS_{A\times S} = \frac{SS_{A\times S}}{df_{A\times S}} = \frac{SS_W - SS_S}{df_W - df_S} \tag{3.14}$$

where SS_S is the sum of squares for the subjects effect with $df_S = n - 1$ degrees of freedom. Equation 3.14 shows the decomposition of the total within-conditions sum of squares into two parts, one due to the subjects effect and the other related to error, or $SS_W = SS_S + SS_{A\times S}$.

The potential power advantage of the dependent samples *F* test over the independent samples *F* test is demonstrated next. Data for three samples are presented in Table 3.4. Results of two different *F* tests with these data are reported in Table 3.5. The first analysis assumes $n = 5$ cases in each of three independent samples, and the second analysis assumes $n = 5$ triads of scores across three dependent samples. Only the second analysis takes account of the positive correlations between each pair of conditions (see Table 3.4). Observe the higher *F* and the lower *p* values for the dependent sample analysis (Table 3.5). Exercise 3 asks you to verify the results of the dependent samples *F* test in Table 3.5.

The dependent samples *F* test assumes normality. Expected dependency among scores due to the subjects effect is removed from the error term (Equation 3.14), so the assumptions of homoscedasticity and independence concern error variances across the levels of the factor. The latter implies that error variance in the first condition has nothing to do with error variance in

TABLE 3.4
Raw Scores and Descriptive Statistics for Three Samples

	Sample		
	1	2	3
	9	8	13
	12	12	14
	13	11	16
	15	10	14
	16	14	18
M	13.00	11.00	15.00
s^2	7.50	5.00	4.00

Note. In a dependent samples analysis, $r_{12} = .7348$, $r_{13} = .7303$, and $r_{23} = .8385$.

the second condition, and so on. This is a strong assumption and probably often implausible, too. This is because error variance from measurements taken close in time, such as adjacent trials in a learning task, may well overlap. This **autocorrelation of the errors** may be less with longer measurement intervals, but autocorrelated error occurs in many within-subjects designs.

Another assumption for factors with ≥ 3 levels is **sphericity (circularity)**, or the requirement for equal population variances of difference scores between every pair of conditions, such as

$$\sigma^2_{D_{12}} = \sigma^2_{D_{13}} = \sigma^2_{D_{23}}$$

TABLE 3.5
Results of the Independent Samples *F* Test and the Dependent Samples *F* Test for the Data in Table 3.4

Source	*SS*	*df*	*MS*	*F*
	Independent samples analysis			
Between (*A*)	40.00	2	20.00	3.64[a]
Within (error)	66.00	12	5.50	
Total	106.00	14		
	Dependent samples analysis			
Between (*A*)	40.00	2	20.00	14.12[b]
Within	66.00	12	5.50	
Subjects	54.67	4	13.67	
$A \times S$ (error)	11.33	8	1.42	
Total	106.00	14		

[a]$p = .058$. [b]$p = .002$.

in designs with three dependent samples. Even relatively small violations of this requirement lead to positive bias (H_0 is rejected too often). There are statistical tests intended to detect violation of sphericity, such as **Mauchly's test**, but they lack power in smaller samples and rely on other assumptions, such as normality, that may be untenable. Such tests are not generally useful (see Baguley, 2004). Keppel and Wickens (2004) suggested that (a) sphericity is doubtful in most studies and (b) researchers should direct their efforts to controlling bias. Exercise 4 asks you to explain why the sphericity requirement does not apply to the dependent samples *t* test.

Summarized next are basic options for dealing with the sphericity assumption in correlated designs; see Keselman, Algina, and Kowalchuk (2001) for more information:

1. Assume maximal violation of sphericity, compute *F* in the usual way, but compare it against a higher critical value with 1, $n - 1$ degrees of freedom, which makes the test more conservative. This method is the **Geisser–Greenhouse conservative test**.
2. Measure the degree of departure from sphericity with estimated epsilon, $\hat{\varepsilon}$. It ranges from $1/(a - 1)$ for maximal departure to 1.00 for no departure. The two degrees of freedom for the critical value for *F* (between-conditions, within-conditions) are then computed as, respectively, $\hat{\varepsilon}\,(a - 1)$ and $\hat{\varepsilon}\,(a - 1)\,(n - 1)$, which makes the test more conservative for $\hat{\varepsilon} < 1.00$. Names for $\hat{\varepsilon}$ include the **Box correction, Geisser–Greenhouse epsilon**, and **Huynh–Feldt epsilon**.
3. Conduct focused comparisons instead of the omnibus comparison, where each contrast has its own error term, so the sphericity requirement does not apply. This contrast test is actually a form of the dependent samples *t* test.
4. Analyze the data with multivariate analysis of variance (MANOVA), which treats difference scores as multiple, correlated outcomes in univariate within-subjects designs (Huberty & Olejnik, 2006). Equal error variances are assumed in MANOVA, but autocorrelation is allowed.
5. Use a statistical modeling technique, such as structural equation modeling or hierarchical linear modeling, that allows for both unequal and correlated error variances in within-subjects designs (e.g., Kline, 2010).
6. Use nonparametric bootstrapping to construct an empirical sampling distribution for dependent samples *F*. The critical value for $\alpha = .05$ falls at the 95th percentile in this distribution. Sphericity is not assumed, but this tactic is not ideal in small samples.

7. Use a robust test for correlated designs based on trimmed means and nonparametric bootstrapping (e.g., Keselman, Kowalchuk, Algina, Lix, & Wilcox, 2000).

Analysis of Variance as Multiple Regression

All forms of ANOVA are nothing more than special cases of multiple regression. In the latter, predictors can be either continuous or categorical (Cohen, 1968). It is also possible in multiple regression to estimate interaction or curvilinear effects. In theory, one needs just a regression computer procedure to conduct any kind of ANOVA. The advantage of doing so is that regression output routinely contains effect sizes in the form of regression coefficients and the overall multiple correlation (or R^2). Unfortunately, some researchers do not recognize these statistics as effect sizes and emphasize only patterns of statistical significance. Some ANOVA computer procedures print source tables with no effect sizes, but it is easy to calculate some of the same effect sizes seen in regression output from values in source tables (see Chapter 5).

χ^2 TEST OF ASSOCIATION

Whether there is a statistical association between two categorical variables is the question addressed by the χ^2 test. A two-way contingency table summarizes the data analyzed by this test. Presented in the top half of Table 3.6 is a 2 × 2 cross-tabulation with frequencies of treatment and control cases ($n = 40$ each) that either recovered or did not recover. A total of 24 cases in the treatment group recovered, or .60. Among control cases, 16 cases recovered, or .40. The recovery rate among treated cases is thus .20 higher than among untreated cases.

The χ^2 statistic for two-way contingency tables is

$$\chi^2[(r-1)(c-1)] = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{o_{ij}} - f_{e_{ij}})^2}{f_{e_{ij}}} \tag{3.15}$$

where the degrees of freedom are the product of the number of rows (r) minus one and the number of columns (c) minus one; $f_{o_{ij}}$ is the observed frequency for the cell in the ith row and jth column; and $f_{e_{ij}}$ is the expected frequency for the same cell under the nil hypothesis that the two variables are unrelated. There is a quick way to compute by hand the expected frequency for any cell: Divide the product of the row and column (marginal) totals for that cell by the total number of cases, N. It is that simple. Assumptions of

TABLE 3.6
Results of the Chi-Square Test of Association for the Same Proportions at Different Group Sizes

		Outcome			
		Observed frequencies			
Group	*n*	Recovered	Not recovered	Recovery rate	χ^2 (1)
		n = 40			
Treatment	40	24	16	.60	3.20[a]
Control	40	16	24	.40	
Total	80	40	40		
		n = 80			
Treatment	80	48	32	.60	6.40[b]
Control	80	32	48	.40	
Total	160	80	80		

[a]$p = .074$. [b]$p = .011$.

the χ^2 test include independence, mutually exclusive cells in the contingency table, and a sample size large enough so that the minimum expected frequency is at least 5 in 2×2 tables.

For the contingency table in the top part of Table 3.6, the expected value for every cell is

$$f_e = (40 \times 40)/80 = 20$$

which shows the pattern under H_0 where the recovery rate is identical in the two groups (20/40, or .50). The test statistic for $n = 40$ is

$$\chi^2(1) = 3.20, p = .074$$

so H_0 is not rejected at the .05 level. (You should verify this result.) The effect of increasing the group size on χ^2 while keeping all else constant is demonstrated in the lower part of Table 3.6. For example, H_0 is rejected at the .05 level for $n = 80$ because

$$\chi^2(1) = 6.40, p = .011$$

even though the difference in the recovery rate is still .20. Exercise 5 asks you to verify the results of the χ^2 test for the group size $n = 80$ in Table 3.6.

ROBUST TESTS

Classical nonparametric tests are alternatives to the parametric *t* and *F* tests for means (e.g., the Mann–Whitney test is the nonparametric analogue to the *t* test). Nonparametric tests generally work by converting the original scores to ranks. They also make fewer assumptions about the distributions of those ranks than do parametric tests applied to the original scores. Nonparametric tests date to the 1950s–1960s, and they share some limitations. One is that they are not generally robust against heteroscedasticity, and another is that their application is typically limited to single-factor designs (Erceg-Hurn & Mirosevich, 2008).

Modern robust tests are an alternative. They are generally more flexible than nonparametric tests and can be applied in designs with multiple factors. The robust tests described next assume a trimming proportion of .20. Selecting a different trimming proportion based on inspecting the data may yield incorrect results. This is because these robust tests do not control for post hoc trimming, so their *p* values may be wrong if any trimming proportion other than .20 is specified (Keselman et al., 2008).

Presented next is the equation for **Yuen–Welch *t* test** based on trimmed means:

$$t_{YW}\left(df_{YW}\right) = \frac{\left(M_{tr1} - M_{tr2}\right) - \left(\mu_{tr1} - \mu_{tr2}\right)}{s_{YW}} \tag{3.16}$$

where the robust standard error s_{YW} and degrees of freedom df_{YW} adjusted for heteroscedasticity are defined by, respectively, Equations 2.32 and 2.33. There is also a **robust Welch–James *t* test**, but the robust Yuen–Welch *t* test may yield effective levels of Type I error that are slightly closer to stated levels of α over random samples (Wilcox, 2012).

Listed next are values of robust estimators for the data from two groups in Table 2.4:

$$M_{tr1} = 23.00,\ s^2_{Win\,1} = 18.489,\ w_1 = 5.547$$

$$M_{tr2} = 17.00,\ s^2_{Win\,2} = 9.067,\ w_2 = 2.720$$

$$M_{tr1} - M_{tr2} = 6.00,\ s_{YW} = 2.875,\ df_{YW} = 8.953$$

The value of the Yuen–Welch *t* statistic is

$$t_{YW}\left(8.953\right) = \frac{6.00}{2.875} = 2.09$$

and $t_{2\text{-tail},\ .05}$ (8.953) = 2.264. Thus, the nil hypothesis H_0: $\mu_{tr1} - \mu_{tr2} = 0$ is not rejected at the .05 level. This outcome is consistent with the robust 95% confidence interval for $\mu_{tr1} - \mu_{tr2}$ of [–.51, 12.51] computed for the same data in Chapter 2 with the Yuen–Welch method.

Robust nonparametric bootstrapping is another way to estimate critical values for a robust test. For example, the critical values for the test of H_1: $\mu_{tr1} - \mu_{tr2} \neq 0$ fall at the 2.5th and 97.5th percentiles in the empirical sampling distribution of the robust Yuen–Welch *t* statistic generated in nonparametric bootstrapping. Unlike critical values based on central *t* distributions, these bootstrapped critical values do not assume normality. The bootstrapping option for robust *t* tests generally requires group sizes of $n > 20$.

Keselman et al. (2000, 2008) described extensions of the robust Welch *t* test for contrasts in between-subjects factorial designs, and Keselman et al. (2000) found that a version of the robust Welch test controlled Type I error reasonably well in correlated designs where the sphericity assumption is violated. Source code by Keselman et al. (2008) in the SAS/IML programming language that conducts the robust Welch *t* test with nonparametric bootstrapping can be downloaded.[3] See Wilcox (2012) for descriptions of additional robust tests based on trimmed means and Winsorized variances.

Erceg-Hurn and Mirosevich (2008) described a class of robust tests based on ranks that are generally better than classical nonparametric techniques. They can also be applied in designs with multiple factors. One is the **Brunner–Dette–Munk test** based on the **ANOVA-type statistic (ATS)**, which tests the null hypothesis that both the population distributions and relative treatment effects are identical over conditions. Relative treatment effects are estimated by converting the original scores to ranks and then computing the proportion of scores in each condition that are higher (or lower) on the outcome variable than all cases in the whole design. Relative treatment effects range from 0 to 1.00, and they all equal .50 under a nil hypothesis. Macros in the SAS/STAT programming language for calculating the ATS in factorial designs can be downloaded from the website of the Abteilung Medizinische Statistik at Universitätsmedizin Göttingen.[4] Wilcox (2012) described packages for R that calculate the ATS.

At the end of the day, robust statistical tests are subject to many of the same limitations as other statistical tests. For example, they assume random sampling albeit from population distributions that may be nonnormal or heteroscedastic; they also assume that sampling error is the only source of error

[3]http://supp.apa.org/psycarticles/supplemental/met_13_2_110/met_13_2_110_supp.html
[4]http://www.ams.med.uni-goettingen.de/amsneu/ordinal-de.shtml

variance. Alternative tests, such as the Welch–James and Yuen–Welch versions of a robust *t* test, do not always yield the same *p* value for the same data, and it is not always clear which alternative is best (Wilcox, 2003). They are also subject to most of the cognitive distortions described in the next chapter. Researchers should not imagine that they hear in robust tests a siren's song to suspend critical judgment about significance testing.

CONCLUSION

Outcomes of statistical tests rely on many assumptions that are farfetched in most studies. Classical parametric tests depend on distributional assumptions, such as normality and homoscedasticity, that are probably untenable in many analyses. Robust tests ease some distributional assumptions, but their *p* values may still be generally incorrect in actual data sets, especially in samples that are not random. The default null hypothesis in significance testing is a nil hypothesis, which is often known to be false before the data are collected. That most researchers disregard power also complicates the interpretation of outcomes of statistical tests. For all these reasons, *p* values in computer output should never be literally interpreted. Additional problems of statistical tests and related myths are considered in the next chapter.

LEARN MORE

F. L. Schmidt (1996) reviews problems with overreliance on significance testing. Appropriate types of power analysis are considered by O'Keefe (2007), and Keselman et al. (2008) introduces robust hypothesis testing with trimmed means in various designs.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13,* 110–129. doi: 10.1037/1082-989X.13.2.110

O'Keefe, D. J. (2007). Post hoc power, observed power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of power analyses. *Communication Methods and Measures, 1,* 291–299. doi:10.1080/19312450701641375

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1,* 115–129. doi:10.1037/1082-989X.1.2.115

EXERCISES

1. For the results reported in Table 3.2, conduct the *t* test for independent samples for $n = 15$ and construct the 95% confidence interval for $\mu_1 - \mu_2$.
2. For the results listed in Table 3.3, conduct the *F* test for independent samples for $n = 30$.
3. For the data in Table 3.4, verify the results of the dependent samples *F* test in Table 3.5. Calculate the source table by hand using four-decimal accuracy for the error term.
4. Explain why the dependent samples *t* test does not assume sphericity.
5. For the data in Table 3.6, verify the results of the χ^2 test for $n = 80$.

This page intentionally left blank

4

COGNITIVE DISTORTIONS IN SIGNIFICANCE TESTING

> If psychologists are so smart, why are they so confused? Why is statistics carried out like compulsive hand washing?
>
> —Gerd Gigerenzer (2004, p. 590)

Many false beliefs are associated with significance testing. Most involve exaggerating what can be inferred from either rejecting or failing to reject a null hypothesis. Described next are the "Big Five" misinterpretations with estimates of their base rates among psychology professors and students. Also considered in this chapter are variations on the Intro Stats method that may be helpful in some situations. Reject-support testing is assumed instead of accept-support testing, but many of the arguments can be reframed for the latter. I assume also that $\alpha = .05$, but the issues dealt with next apply to any other criterion level of statistical significance.

BIG FIVE MISINTERPRETATIONS

Please take a moment to review the correct interpretation of statistical significance (see Chapter 3). Briefly, $p < .05$ means that the likelihood of the data or results even more extreme given random sampling under the

DOI: 10.1037/14136-004
Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, Second Edition, by R. B. Kline

TABLE 4.1
The Big Five Misinterpretations of $p < .05$ and Base Rates Among Psychology Professors and Students

Fallacy	Description	Base rate (%)	
		Professors[a]	Students[b]
	Misinterpretations of p		
Odds against chance	Likelihood that result is due to chance is < 5%	—	—
Local Type I error	Likelihood that Type I error was just committed is < 5%	67–73	68
Inverse probability	Likelihood that H_0 is true is < 5%	17–36	32
	Misinterpretations of $1 - p$		
Validity	Likelihood that H_1 is true is > 95%	33–66	59
Replicability	Likelihood that result will be replicated is > 95%	37–60	41

Note. Table adapted from R. B. Kline, 2009, *Becoming a Behavioral Science Researcher: A Guide to Producing Research That Matters,* p. 125, New York, Guilford Press. Copyright 2009 by Guilford Press. Adapted with permission. Dashes (—) indicate the absence of estimated base rates.
[a]Haller and Krauss (2002), Oakes (1986). [b]Haller and Krauss (2002).

null hypothesis is < .05, assuming that all distributional requirements of the test statistic are satisfied and there are no other sources of error variance. Let us refer to any correct definition as p (D+$|H_0$), which emphasizes p as the conditional probability of the data under H_0 given all the other assumptions just mentioned.

Listed in Table 4.1 are the Big Five false beliefs about statistical significance. Three concern p values, but two others involve their complements, or $1 - p$. Also reported in the table are base rates in samples of psychology professors or students (Haller & Krauss, 2002; Oakes, 1986). Overall, psychology students are no worse than their professors regarding erroneous beliefs. These poor results are not specific to psychology (e.g., forecasting; Armstrong, 2007). It is also easy to find similar misunderstandings in journal articles and statistics textbooks (e.g., Cohen, 1994; Gigerenzer, 2004). These results indicate that myths about significance testing are passed on from teachers and published works to students.

Odds-Against-Chance Fallacy

This myth is so pervasive that I believe the **odds-against-chance fallacy** (Carver, 1978) is the biggest of the Big Five. It concerns the false belief that p indicates the probability that a result happened by sampling error; thus,

$p < .05$ says that there is less than a 5% likelihood that a particular finding is due to chance. There is a related misconception I call the **filter myth**, which says that p values sort results into two categories, those that are a result of "chance" (H_0 not rejected) and others that are due to "real" effects (H_0 rejected). These beliefs are wrong for the reasons elaborated next.

When p is calculated, it is already assumed that H_0 is true, so the probability that sampling error is the only explanation is already taken to be 1.00. It is thus illogical to view p as measuring the likelihood of sampling error. Thus, p does not apply to a particular result as the probability that sampling error was the sole causal agent. There is no such thing as a statistical technique that determines the probability that various causal factors, including sampling error, acted on a particular result. Instead, inference about causation is a rational exercise that considers results within the context of design, measurement, and analysis. Besides, virtually all sample results are affected by error of some type, including measurement error.

I am not aware of an estimate of the base rate of this fallacy, but I believe that it is nearly universal. This is because one can find this misinterpretation just about everywhere. Try this exercise: Enter the term *define: statistical significance* in the search box of Google. What you will then find displayed on your computer monitor are hundreds of incorrect definitions, most of which invoke the odds-against-chance fallacy. Granted, these web pages are not academic sources, but similar errors are readily found in educational works, too.

Local Type I Error Fallacy

Most psychology students and professors may endorse the **local Type I error fallacy** (Table 4.1). It is the mistaken belief that $p < .05$ given $\alpha = .05$ means that the likelihood that the decision just taken to reject H_0 is a Type I error is less than 5%. Pollard (1993) described this fallacy as confusing the conditional probability of a Type I error, or

$$\alpha = p(\text{Reject } H_0 | H_0 \text{ true})$$

with the conditional posterior probability of a Type I error given that H_0 has been rejected, or

$$p(H_0 \text{ true} | \text{Reject } H_0)$$

But p values from statistical tests are conditional probabilities of data, so they do not apply to any specific decision to reject H_0. This is because any particular decision to do so is either right or wrong, so no probability is associated with it (other than 0 or 1.0). Only with sufficient replication could one determine whether a decision to reject H_0 in a particular study was correct.

Inverse Probability Fallacy

About one third of psychology students and professors endorse the inverse probability error (Table 4.1), also known as the **fallacy of the transposed conditional** (Ziliak & McCloskey, 2008), the **Bayesian Id's wishful thinking error** (Gigerenzer, 1993), and the **permanent illusion** (Gigerenzer & Murray, 1987) due to its persistence over time and disciplines. It was defined earlier as the false belief that p measures the likelihood that H_0 is true, given the data. A researcher who interprets the result $p < .05$ as saying that H_0 is true with a probability < .05 commits this error. It stems from forgetting that p values are conditional probabilities of the data, or p (D+|H_0), and not of the null hypothesis, or p (H_0|D+). There are ways in Bayesian statistics to estimate conditional probabilities of hypotheses (see Chapter 10) but not in traditional significance testing.

Validity Fallacy

Two of the Big Five misunderstandings concern $1 - p$. One is the **valid research hypothesis fallacy** (Carver, 1978), which refers to the false belief that the probability that H_1 is true is > .95, given $p < .05$. The complement of p is a probability, but $1 - p$ is just the probability of getting a result even less extreme under H_0 than the one actually found. This fallacy is endorsed by most psychology students and professors (see Table 4.1).

Replicability Fallacy

About half of psychology students and professors endorse the **replicability fallacy** (see Table 4.1), or the erroneous belief that the complement of p indicates the probability of finding a statistically significant result in a replication study (Carver, 1978). Under this falsehood, given $p < .05$, a researcher would infer that the probability of replication is > .95. If this fallacy were true, knowing the probability of replication would be very useful. Alas, p is just the probability of the data in a particular study under H_0 under many stringent assumptions.

You should know that there *is* a sense in which p values indirectly concern replication, but the probability of the latter is not generally $1 - p$. Greenwald, Gonzalez, Harris, and Guthrie (1996) showed there is a curvilinear relation between p values and the average statistical power in hypothetical random replications based on the same number of cases. In general, if the population effect size is the same as that in a specific sample needed to obtain $p < .05$, the probability that the same H_0 will be rejected in a replication is about .50, not .95.

Killeen (2005, 2006) described a point estimate of the probability of getting statistical significance over random replications in the same direction as in an original sample known as p_{rep}. The method for estimating p_{rep} assumes that the observed effect size is the same as the population effect size, which is unlikely. It is based on a **random effects model** for population effect sizes, which assumes a distribution of population effect sizes, or that there is a different true effect size for each study. Estimation of p_{rep} relies on accurate estimation of variation in population effect sizes, or the **realization variance**.

Killeen (2006) described an inference model that replaces significance testing with a utility theory approach based on p_{rep} and interval estimation. It takes account of the seriousness of different types of decisions errors about whether to replicate a particular study. An advantage of this framework is that it makes apparent the falsehood that $1 - p$ is the probability of replication. Killeen suggested that p_{rep} may be less subject to misinterpretation, but this remains to be seen. The estimate of p_{rep} also assumes random sampling, which is difficult to justify in most studies; see J. Miller (2009) and Iverson and Lee (2009) for additional criticisms. I believe it is better to actually conduct replications than to rely on statistical prediction.

Ubiquitary Nature of the Big Five

Results by Oakes (1986) and Haller and Krauss (2002) indicated that virtually all psychology students and about 80 to 90% of psychology professors endorsed at least one of the Big Five false beliefs. So it seems that most researchers believe for the case $\alpha = .01$ and $p < .01$ that the result is very unlikely to be due to sampling error and that the probability a Type I error was just committed is just as unlikely (< .01 for both). Most researchers might also conclude that H_1 is very likely to be true, and many would also believe that the result is very likely to replicate (> .99 for both). These (misperceived) odds in favor of the researcher's hypothesis are so good that it must be true, right? The next (il)logical step would be to conclude that the result must also be important. Why? Because it is *significant*! Of course, none of these things are true, but the Big Five are hardly the end of cognitive distortions in significance testing.

MISTAKEN CONCLUSIONS AFTER MAKING A DECISION ABOUT THE NULL HYPOTHESIS

Several different false conclusions may be reached after deciding to reject or fail to reject H_0. Most require little explanation about why they are wrong.

Magnitude Fallacy

The **magnitude fallacy** is the false belief that low p values indicate large effects. Cumming (2012) described a related error called the **slippery slope of significance** that happens when a researcher ambiguously describes a result for which $p < \alpha$ as "significant" without the qualifier "statistically" and then later discusses the effect as if it were automatically "important" or "large." These conclusions are unwarranted because p values are confounded measures of effect size and sample size (see Equation 3.4). Thus, effects of trivial magnitude need only a large enough sample to be statistically significant. If the sample size is actually large, low p values just confirm a large sample, which is circular logic (B. Thompson, 1992).

Meaningfulness Fallacy and Causality Fallacy

Under the **meaningfulness fallacy**, the researcher believes that rejection of H_0 confirms H_1. This myth actually reflects two cognitive errors. First, the decision to reject H_0 in a single study does not imply that H_1 is "proven." Second, even if the *statistical* hypothesis H_1 is correct, it does not mean that the *substantive* hypothesis behind H_1 is also correct. Statistical significance does not "prove" any particular hypothesis, and there are times when the same numerical result is equally consistent with more than one substantive hypothesis.

Statistical versus substantive hypotheses not only differ in their levels of abstraction (statistical: lowest; scientific: highest) but also have different implications following rejection of H_0. If H_0 and H_1 reflect merely statistical hypotheses, there is little to do after rejecting H_0 except replication. But if H_1 stands for a scientific hypothesis, the work just begins after rejecting H_0. Part of the task involves evaluating competing substantive hypotheses that are also compatible with the statistical hypothesis H_1. If alternative explanations cannot be ruled out, confidence in the original hypothesis must be tempered. There is also the strategy of **strong inference** (Platt, 1964), in which experiments are devised that would yield different results depending on which competing explanation is correct. For the same reasons, the **causality fallacy** that statistical significance means that the underlying causal mechanism is identified is just that.

Zero Fallacy and Equivalence Fallacy

The **zero fallacy** or the **slippery slope of nonsignificance** (Cumming, 2012) is the mistaken belief that the failure to reject a nil hypothesis means that the population effect size is zero. Maybe it is, but you cannot tell based

on a result in one sample, especially if power is low. In this case, the decision to not reject a nil hypothesis would be a Type II error. Improper design, procedures, or measures can also lead to Type II errors. The **equivalence fallacy** occurs when the failure to reject H_0: $\mu_1 = \mu_2$ is interpreted as saying that the populations are equivalent. This is wrong because even if $\mu_1 = \mu_2$, distributions can differ in other ways, such as variability or distribution shape. The inference of equivalence would be just as wrong if this example concerned reliability coefficients or validity coefficients that were not statistically different (B. Thompson, 2003). Proper methods for equivalence testing are described later.

Quality Fallacy and Success Fallacy

The beliefs that getting statistical significance confirms the quality of the experimental design and also indicates a successful study are, respectively, the **quality fallacy** and the **success fallacy**. Poor study design or just plain old sampling error can lead to incorrect rejection of H_0, or Type I error. Failure to reject H_0 can also be the product of good science, especially when a false claim is not substantiated by other researchers. You may have heard about the case in the 1990s about a group of physics researchers who claimed to have produced cold fusion (a low energy nuclear reaction) with a simple laboratory apparatus. Other scientists were unable to replicate the phenomenon, and the eventual conclusion was that the original claim was an error. In an article about the warning signs of bogus science, Park (2003) noted that

> a PhD in science is not an inoculation against foolishness or mendacity, and even some Nobel laureates seem to be a bit strange. The sad truth is that there is no claim so preposterous that a PhD scientist cannot be found to vouch for it. (p. 33)

In this case, lack of positive results from replication studies is informative.

Failure Fallacy

The **failure fallacy** is the mistaken belief that lack of statistical significance brands the study as a failure. Gigerenzer (2004) recited this older incantation about doctoral dissertations and the critical ratio, the predecessor of *p* values: "A critical ratio of three [i.e., $p < .01$], or no PhD" (p. 589). Although improper methods or low power can cause Type II errors, the failure to reject H_0 can be an informative result. Researchers tend to attribute failure to reject H_0 to poor design rather than to the validity of the substantive hypothesis behind H_1 (Cartwright, 1973).

Reification Fallacy

The **reification fallacy** is the faulty belief that failure to replicate a result is the failure to make the same decision about H_0 across studies (Dixon & O'Reilly, 1999). In this view, a result is not considered replicated if H_0 is rejected in the first study but not in the second study. This sophism ignores sample size, effect size, and power across different studies. Suppose a mean difference is found in an initial study and a nil hypothesis is rejected. The same contrast is found in a replication, but H_0 is not rejected due to a smaller sample size. There is evidence for replication even though different decisions about H_0 were made across the two studies (e.g., see Table 2.1).

Objectivity Fallacy

The myth that significance testing is an objective method of hypothesis testing but all other inference models are subjective is the **objectivity fallacy** (Gorard, 2006). To the contrary, there are many decisions to be made in significance testing, some of which have little to do with substantive hypotheses (see Chapter 3). Significance testing is objective in appearance only. It is also not the only framework for testing hypotheses. Bayesian estimation as an alternative to significance testing is considered in Chapter 10.

Sanctification Fallacy

The **sanctification fallacy** refers to dichotomous thinking about continuous p values. If $\alpha = .05$, for example, $p = .049$ versus $p = .051$ are practically identical in terms of test outcomes. Yet a researcher may make a big deal about the first but ignore the second. There is also evidence for the **cliff effect**, which refers to an abrupt decline in the degree of confidence that an effect exists for p just higher than .05 (Nelson, Rosenthal, & Rosnow, 1986). Nelson et al. (1986) also found a second decline when p values were just above .10. These changes in rated confidence are out of proportion to changes in continuous p values. More recently, Poitevineau and Lecoutre (2001) found that a minority of researchers exhibited the cliff effect, which is consistent with the prior specification of α in the Neyman–Pearson approach. Other researchers showed more gradual declines in confidence as p values increased, which is more consistent with the Fisher approach. Just as the Intro Stats method is a mishmash of Fisher's and Neyman–Pearson's models, so it seems are researchers' patterns of rated confidence as a function of p.

Differences between results that are "significant" versus "not significant" by close margins, such as $p = .03$ versus $p = .07$ when $\alpha = .05$, are themselves often not statistically significant. That is, relatively large changes in p

can correspond to small, nonsignificant changes in the underlying variable (Gelman & Stern, 2006). This is another reason not to make big distinctions among results with similar *p* values. There is also the bizarre practice of describing results where *p* is just higher than α as "trends" or as "approaching significance." These findings are also typically interpreted along with statistically significant ones. The problem is that results with *p* values just lower than α (e.g., $p = .049$, $\alpha = .05$) are almost never described as "approaching nonsignificance" and subsequently discounted.

Robustness Fallacy

Classical parametric statistical tests are not robust against outliers or violations of distributional assumptions, especially in small, unrepresentative samples. But many researchers believe just the opposite, which is the **robustness fallacy**. Indirect support for this claim comes from observations that most researchers do not provide evidence about whether distributional or other assumptions are met (e.g., Keselman et al., 1998; Onwuegbuzie, 2002). These surveys reflect a large gap between significance testing as described in textbooks and its use in practice. The fact that most articles fail to reassure readers that the results are trustworthy is part of the reporting crisis (see Chapter 1). It is too bad that most researchers ignore this sound advice of Wilkinson and the TFSI (1999): Address data integrity before presenting the results. This includes any complications, such as missing data or distributional anomalies and steps taken to remedy them (e.g., robust estimation, transformations). This quote attributed to the Scottish author George MacDonald is apropos for researchers: "To be trusted is a greater compliment than being loved."

WHY SO MANY MYTHS?

Many fallacies involve wishful thinking about things that researchers really want to know. These include the probability that H_0 or H_1 is true, the likelihood of replication, and the chance that a particular decision to reject H_0 is wrong. Alas, statistical tests tell us only the conditional probability of the data. Cohen (1994) noted that no statistical technique applied in individual studies can fulfill this wish list. Bayesian methods are an exception because they estimate conditional probabilities of hypotheses (see Chapter 10). But there is a method that can tell us what we want to know. It is not a statistical technique; rather, it is good, old-fashioned replication, which is also the best way to deal with the problem of sampling error.

Most people who use statistical tests have science backgrounds, so why is there so much misunderstanding about significance testing? Two possibilities

were offered in Chapter 1: It is hard for people (scientists included) to change bad habits. There is conflict of interest in playing the significance game, or the well-trodden routine of promising asterisks (statistical significance) for money from granting agencies, and then delivering those asterisks in written summaries with little understanding of what the presence versus absence of those asterisks really means (if anything) before promising more asterisks, and so on. It is also true that careers in behavioral research are generally based on amassing large piles of asterisks over the years, and some researchers may resist, as Leo Tolstoy put it, any pressure to "admit the falsity of conclusions which they have delighted in explaining to colleagues, which they have proudly taught to others, and which they have woven, thread by thread, into the fabric of their lives" (as cited in Gleick, 1987, p. 38).

An additional factor is that it is hard to explain the convoluted logic of significance testing and dispel confusion about it (Pollard, 1993). Authors who defend significance testing concede widespread false beliefs but note that users are responsible for misinterpretations (e.g., Hurlbert & Lombardi, 2009, p. 337, "Researchers, heal thyselves! No fault lies with the significance test!") Critics counter that any method with so much potential to be misconstrued must bear some blame. It is also clear that the cliché of "better teaching" about significance testing has not in more than 60 years improved the situation (Fidler et al., 2004).

Another problem is that people readily make judgment errors based on perceived probabilities of events, sometimes at great cost to their personal well-being (e.g., gambler's fallacy). It is especially difficult for people to think correctly and consistently about conditional probabilities, which are ratios of event probabilities (e.g., Dawes, 2001). A complication is the phenomenon of **illusory correlation**, or the expectation that two things should be correlated when in fact they are not. If such expectations are based on apparently logical associations, the false belief that two things go together can be very resistant to disconfirmation. Semantic associations between the concepts of "chance" and "data" combined with poor inherent probabilistic reasoning may engender illusory correlation in significance testing (e.g., the odds-against-chance fallacy). Once engrained, such false beliefs are hard to change. But stripping away the many illusions linked with significance testing should precede the realization that

> there is no automatic method for statistical induction which consists of a simple recipe. Analyzing research results and drawing inferences therefrom require rethinking the situation in each case and devising an appropriate method. Such a challenge is not always welcome. (Falk & Greenbaum, 1995, p. 76)

It is regrettable that the word *significant* was ever associated with rejecting H_0. Connotations of this word in everyday language, which are illus-

Figure 4.1. Visual map of words with meanings similar to that of "significant." Image and text from the Visual Thesaurus (http://www.visualthesaurus.com). Copyright 1998–2011 by Thinkmap, Inc. All rights reserved. Reprinted with permission.

trated in Figure 4.1 created with the Thinkmap Visual Thesaurus,[1] include "important," "noteworthy," and "monumental," but *none* of them automatically apply to H_0 rejections. One way to guard against overinterpretation is to drop the word *significant* from our data analysis vocabulary altogether. Hurlbert and Lombardi (2009) reminded us that there is no obligation is use the word *significant* at all in research. Another is to always use the phrase *statistically significant* (see B. Thompson, 1996), which signals that we are not talking about significance in the everyday sense (e.g., Figure 4.1). Using just the word *statistical* should also suffice. For example, rejection of H_0: $\mu_1 = \mu_2$ could be described as evidence for a statistical mean difference (Tryon, 2001). Calling an effect statistical implies that it was observed but not also necessarily important or real.

Another suggested reform is to report exact *p* values, such as

$$t(20) = 2.40,\ p = .026 \text{ instead of } t(20) = 2.40,\ p < .05$$

[1]http://www.visualthesaurus.com/

If *p* values are incorrect in most studies, though, reporting them at even greater precision could give a false impression. In large samples, *p* values are often very low, such as .00012, and reporting such small probabilities could invite misunderstanding. I think that 3-decimal accuracy is fine, but do *not* take *p* values literally and help your readers to do the same.

Whatever cognitive mechanisms engender or maintain false beliefs about significance testing, it is clear that data analysis practices in psychology and other areas are based more on myth than on fact. This is why Lambdin (2012) described significance testing as a kind of statistical sorcery or shamanism where researchers hold up *p* values that they do not understand as the assurance of confirmatory evidence in a farcical imitation of science passed on from one generation to the next. For the same reasons, Bakan (1966) basically referred to significance testing as psychology's "dirty little secret" (Lambdin, 2012) and Lykken (1991) described its continued practice as a kind of cargo-cult ritual where we mimic the natural sciences but are not really doing science at all.

ADDITIONAL PROBLEMS

Considered next are negative consequences of letting *p* values do our thinking for us:

1. *The ease with which a researcher can report statistically significant evidence for untrue hypotheses fills the literature with false positive results*. Many decisions influence outcomes of statistical tests, including ones about sample size (e.g., when to stop collecting data) and the specification of directional versus nondirectional alternative hypotheses, among other factors not always disclosed. Simmons et al. (2011) described these kinds of decisions as **researcher degrees of freedom**. They also demonstrated through computer simulations that practically any result can be presented as statistically significant due to flexibility in choices. Simmons et al. (2011) suggested that researchers disclose all decisions that affect *p* values, but this rarely happens in practice.
2. *Significance testing diverts attention from the data and the measurement process*. If researchers are too preoccupied with H_0 rejections, they might lose sight of more important aspects of their data, such as whether constructs are properly defined and measured. By focusing exclusively on sampling error, researchers may neglect dealing with the greater problem of real error (see Chapter 2).
3. *The large amount of time required to learn significance testing limits exposure to other approaches*. Students in the behavioral sciences typically know very little about other inference models, such as

Bayesian estimation, even in graduate school. That extensive training in traditional significance testing seems to mainly fill many students' heads with fairy tales about this procedure is another problem.

4. *Strongly embedded significance testing mentality hinders continued learning.* Next I describe a phenomenon I call the **great *p* value blank-out**. This happens when students put little effort into learning what is for them a new statistical technique until they get to the part about significance testing. Once they figure out whether rejecting H_0 is "good" or "bad" for their hypotheses, the technique is then applied with little comprehension of its theory or requirements. In computer output, the student skips right to the *p* values but shows little comprehension of other results. This cognitive style is a kind of self-imposed trained incapacity that obstructs new learning. Rodgers (2010) described similar experiences with students trained mainly in traditional significance testing.
5. *Overemphasis on statistical significance actually dampens enthusiasm for research.* Low power (e.g., 50) may result in a research literature where only about half the results are positive, and this ambiguity cannot be resolved by additional studies if power remains low. This may explain why even undergraduates know that "the three most commonly seen terms in the [research] literature are 'tentative,' 'preliminary,' and 'suggest.' As a default, 'more research is needed'" (Kmetz, 2002, p. 62). It is not just a few students who are skeptical of the value of research but also practitioners, for whom statistical significance often does not translate into relevance or substantive significance (Aguinis et al., 2010).

ILLEGITIMATE USES OF SIGNIFICANCE TESTING

Some applications of statistical tests are inappropriate in just about any context. Two were mentioned in the previous chapter: One is testing score reliabilities or validity coefficients for statistical significance against nil hypotheses. A second concerns statistical tests that evaluate the distributional assumptions of other statistical tests, such as Mauchley's test for sphericity in correlated designs. The problem with such "canary in a coal mine" tests (i.e., that evaluate assumptions of other statistical tests) is that they often depend on other assumptions, such as normality, that may be indefensible. This is why Erceg-Hurn and Mirosevich (2008) advised that "researchers should not rely on statistical tests to check assumptions because of the frequency with which they produce inaccurate results" (p. 594).

A third example is the **stepwise method** in multiple regression or discriminant function analysis where predictors are entered into the equation based solely on statistical significance (e.g., which predictor, if selected, would have the lowest *p* value for its regression weight?). After they are selected, predictors at a later step can be removed from the equation, again according to statistical test outcomes (e.g., if $p > .05$ for a predictor's regression weight). The stepwise process stops when there could be no statistically significant increase in R^2 by adding more predictors. There are variations on stepwise methods, but all such methods are directed by the computer, not the researcher.

Problems of stepwise methods are so severe that they are actually banned in some journals (e.g., B. Thompson, 1995) and for good reason, too. One is extreme capitalization on chance. Because every result in these methods is determined by *p* values, the results are unlikely to replicate in a new sample. Another is that *p* values in computer output for stepwise methods are typically wrong because they are not adjusted for nonchance selection (i.e., Feynman's conjecture; see Chapter 3). Worst of all, stepwise methods give the false impression that the researcher does not have to think about the problem of predictor selection and entry order; see Whittingham, Stephens, Bradbury, and Freckleton (2006) for additional criticisms.

DEFENSES OF SIGNIFICANCE TESTING

The litany of complaints reviewed so far raises the question of whether anything is right with significance testing. Some authors defend its use while acknowledging that significance testing should be supplemented with effect sizes or confidence intervals (e.g., Aguinis et al., 2010; Hurlbert & Lombardi, 2009). Some supportive arguments are summarized next:

1. *If nothing else, significance testing addresses sampling error.* Significance testing provides a method for managing the risk of sampling error and controlling the long-run risk of Type I error. Thus, some researchers see significance testing as addressing important needs and therefore may be less like passive followers of tradition than supposed by critics. Those critics rightly point out that confidence intervals convey more direct information about sampling error. They also suggest that excessive fixation on *p* values is one reason why confidence intervals are not reported more often.
2. *Significance testing is a gateway to decision theory.* In situations where researchers can estimate costs of Type I versus Type II

error and benefits of correct decisions, decision theory offers a framework for estimating overall utility in the face of uncertainty. Decision theory may also be able to detect long-term negative consequences of an intervention even while statistical tests fail to reject the nil hypothesis of no short-term effect (D. H. Johnson, 1999). But it is rare in the behavioral sciences that researchers can estimate costs versus benefits regarding their decisions. This is why Hurlbert and Lombardi (2009) and others have suggested that mass manufacturing is the ideal application for the Intro Stats method. Manufacturing processes are susceptible to random error. If this error becomes too great, products fail. In this context, H_0 represents a product specification that is reasonable to assume is true, samples can be randomly selected, and exact deviations of sample statistics from the specification can be accurately measured. Costs for corrective actions, such as a product recall, may also be known. Perhaps the behavioral sciences are just the wrong context for significance testing.

3. *Some research questions require a dichotomous answer.* There are times when the question that motivates research is dichotomous (e.g., Should this intervention be implemented? Is this drug more effective than placebo?). The outcome of significance testing is also dichotomous. It deals with whether observed effects stand out against sampling error. But significance testing does not estimate the magnitudes of those effects, nor does it help with real-world decisions of the kind just stated. The latter involve judging whether observed effect sizes are sufficiently large to support the investment of resources. Evidence for replication would also be crucial, but the final yes-or-no decision is ultimately a matter of informed judgment based on all available evidence.
4. *Nil hypotheses are sometimes appropriate.* Robinson and Wainer (2002) noted that nil hypotheses are sometimes justified in multiple factor designs when there is no reason to expect an effect when just one independent variable is manipulated. In most studies, though, nil hypotheses are feeble arguments.

VARIATIONS ON THE INTRO STATS METHOD

This section identifies variations on significance testing that may be useful in some situations.

Neo-Fisherian Significance Assessments

Hurlbert and Lombardi (2009) recommended replacing the Intro Stats method, or the **paleo-Fisherian and Neyman–Pearsonian approaches**, with a modified version, **Neo-Fisherian significance assessments**. This model has the three characteristics listed next:

1. *No dichotomization of* p *values*. This means that (a) exact *p* values are reported, but they are not compared with an arbitrary standard such as .05. (b) The terms *significant* versus *not significant* when describing *p* values are dropped. The former modification is closer to the original Fisher approach than the Neyman–Pearson approach (see Table 3.1), and the latter is intended to minimize the cliff effect.
2. *High* p *values lead to the decision to suspend judgment*. The label for this outcome reminds researchers not to "accept" H_0 in this case or somehow believe $p > \alpha$ is evidence that H_0 is true. This variation may protect researchers against the inverse probability, zero, or equivalence fallacies when *p* values are high.
3. *Adjunct analyses may include effect size estimation and interval estimation*. When metrics of outcome variables are meaningful, effect sizes should be estimated in that original metric. That is, researchers should report unstandardized effect sizes. Standardized effect sizes that are metric free, such as squared correlations (i.e., proportions of explained variance), should be reported only when scales of outcomes of variables are arbitrary. The distinction between unstandardized and standardized effect sizes is elaborated in the next chapter.

The decision to suspend judgment derives from an approach to significance testing based on **three-valued logic** (e.g., Harris, 1997), which allows split-tailed alternative hypotheses that permit statistically significant evidence against a substantive hypothesis if the direction of the observed effect is not as predicted. Hurlbert and Lombardi (2009) also emphasized the distinction between statistical hypotheses and substantive hypotheses. In particular, researchers should not mistake support for a statistical hypothesis as automatically meaning support for any corresponding substantive hypothesis. The Hulbert–Lombardi model features some welcome reforms, but it relies on the assumption that *p* values are generally accurate in most studies.

Customer-Centric Science

Aguinis et al. (2010) described an approach to significance testing known as **customer-centric science**. It is intended to make results more rel-

evant for stakeholders other than researchers, especially practitioners. The basic ideas are listed next:

1. *Set the level of* α *rationally, not arbitrarily, and report the exact value of* p. This means that the researcher estimates the probability of a Type II error β, calculates the desired relative serious of a Type I versus Type II error, and estimates the prior probability that H_1 is true. Next, the researcher applies Equation 3.2 to compute the optimal level of α, given the information just mentioned. After the test is conducted, the exact *p* value is reported.
2. *Report effect sizes that indicate the degree to which an outcome is explained or predicted.* Aguinis et al. (2010) emphasized that researchers should also avoid applying arbitrary standards for describing observed effect sizes as "small," "medium," or "large," or any other qualitative description that may not apply in a particular research area. This point is elaborated in the next chapter.
3. *Interpret statistically significant results and their magnitudes in ways that are meaningful for stakeholders other than researchers.* This part of Aguinis et al.'s (2010) approach explicitly recognizes that neither statistical significance nor effect size directly addresses the practical, clinical, or, more generally, substantive significance of the findings. These authors do not suggest that this kind of communication should be mandatory in research reports, but doing so presents potential opportunities to narrow the communication gap between researchers and practitioners.

Suggestions for how to convey the substantive meaning of research results are offered in the next chapter (and are discussed by Aguinis et al., 2010), but this process requires that researchers become aware of the concerns of stakeholders and speak in language relevant to this audience. It also means that the overly technical, jargon-filled discourse that avoids any consideration of substantive significance and is seen in far too many journal articles is to be avoided. A key difference between customer-centric science and my recommendations presented in the next section is a central role for significance testing in the former.

Equivalence Testing

The method of **equivalence testing** is better known in pharmacology and environmental sciences. It concerns the problem of how to establish equivalence between groups or treatments. Suppose that a researcher wishes to determine whether a generic drug can be substituted for a more expensive drug. In traditional significance testing, the failure to reject H_0: $\mu_1 = \mu_2$ is not

evidence that the drugs are equivalent. In equivalence testing, a single point null hypothesis is replaced by two range subhypotheses. Each subhypothesis expresses a range of $\mu_1 - \mu_2$ values that corresponds to substantive mean differences. For example, the pair of subhypotheses

$$H_0: \begin{cases} H_{0_1}: (\mu_1 - \mu_2) < -10.00 \\ H_{0_2}: (\mu_1 - \mu_2) > 10.00 \end{cases}$$

says that the population means cannot be considered equivalent if the absolute value of their difference is greater than 10.00. The complementary interval for this example is

$$-10.00 \leq (\mu_1 - \mu_2) \leq 10.00$$

which is a **good-enough belt** for a hypothesis of equivalence, also called a **range of practical equivalence**. It is a range hypothesis that indicates the value(s) of the parameter(s) considered equivalent and uninteresting. Standard statistical tests are used to contrast the observed mean difference against each of these one-sided null hypotheses for a directional H_1. Only if both range subhypotheses are rejected at the same level of α can the compound null hypothesis of nonequivalence be rejected.

In the approach just outlined, Type I error is the probability of declaring two populations or conditions to be equivalent when in truth they are not. In a drug study, this risk is the patient's (consumer's) risk. McBride (1999) showed that if Type I error risk is to be the producer's instead of the patient's, the null hypothesis appropriate for this example would be the range hypothesis

$$H_0 : -10.00 \leq (\mu_1 - \mu_2) \leq 10.00$$

and it would be rejected either if the lower end of a one-sided confidence interval about the observed mean difference is greater than 10.00 or if the upper end of a one-sided confidence interval is less than –10.00. See Wellek (2010) for more information.

Inferential Confidence Intervals

Tryon (2001) proposed an integrated approach to testing means for statistical difference, equivalence, or indeterminancy (neither statistically different or equivalent). It is based on **inferential confidence intervals**, which are modified confidence intervals constructed around individual means. The width of an inferential confidence interval is the product of the standard error of the mean (Equation 2.6) and a two-tailed critical *t* value reduced

by a correction factor that equals the ratio of the standard error of the mean difference (Equation 2.12) over the sum of the individual standard errors. Because values of this correction factor range from about .70 to 1.00, widths of inferential confidence intervals are generally narrower than those of standard confidence intervals about the same means.

A statistical difference between two means occurs in this approach when their inferential confidence intervals do not overlap. The probability associated with this statistical difference is the same as that from the standard *t* test for a nil hypothesis and a nondirectional H_1. Statistical equivalence is concluded when the **maximum probable difference** between two means is less than an amount considered inconsequential as per an equivalence hypothesis. The maximum probable difference is the difference between the highest upper bound and the lowest lower bound of two inferential confidence intervals. For example, if [10.00, 14.00] and [12.00, 18.00] are the inferential confidence intervals based on two different means, the maximum probable difference is 18.00 – 10.00, or 8.00. If this difference lies within the range set by the equivalence hypothesis, statistical equivalence is inferred. A contrast neither statistically different nor equivalent is indeterminant, and it is not evidence for or against any hypothesis.

Tryon and Lewis (2008) extended this approach when testing for statistical equivalence over two or more populations. Tryon (2001) claimed that the method of inferential confidence intervals is less susceptible to misinterpretation because (a) the null hypothesis is implicit instead of explicit, (b) the model covers tests for both differences and equivalence, and (c) the availability of a third outcome—statistical indeterminancy—may help to prevent the interpretation of marginally nonsignificant differences as "trends."

BUILDING A BETTER FUTURE

Outlined next are recommendations that call for varying degrees of use of statistical tests—from none at all to somewhat more pivotal depending on the context—but with strict requirements for their use. These suggestions are intended as a constructive framework for reform and renewal. I assume that reasonable people will disagree with some of the specifics put forward. Indeed, a lack of consensus has characterized the whole debate about significance testing. Even if you do not endorse all the points elaborated next, you may at least learn new ways of looking at the controversy over statistical tests or, even better, data, which is the ultimate goal of this discussion.

A theme underlying these recommendations can be summarized like this: Significance testing may have helped us in psychology and other behavioral sciences through a difficult adolescence during which we struggled to

differentiate ourselves from the humanities while at the same time strived to become more like the natural sciences. But just as few adults wear the same style of clothes, listen to the same types of music, or have the same values they did as teenagers, behavioral science needs to leave its adolescence behind. Growing up is a series of conscious choices followed by more mature actions. Continued arrested development and stagnation of our research literature are possible consequences of failing the challenge of statistics reform.

A second theme is the realization that statistical significance provides even in the best case nothing more than low-level support for the existence of an effect, relation, or difference. That best case occurs when researchers estimate a priori power, specify the correct construct definitions and operationalizations, work with random or at least representative samples, analyze highly reliable scores in distributions that respect test assumptions, control other major sources of imprecision besides sampling error, and test plausible null hypotheses. In this idyllic scenario, *p* values from statistical tests may be reasonably accurate and potentially meaningful, if they are not misinterpreted. But science should deal with more than just the existence question, a point that researchers overly fixated on *p* values have trouble understanding. As Ziliak and McCloskey (2008) put it, two other vital questions are "how much?" (i.e., effect size) and "so what?" (i.e., substantive significance).

A third theme is that behavioral science of the highest caliber is possible without significance testing *at all.* Here it is worth noting that some of the most influential empirical work in psychology, including that of Piaget, Pavlov, and Skinner, was conducted without rejecting null hypotheses (Gigerenzer, 1993). The natural sciences have thrived without relying on significance testing. Ziliak and McCloskey (2008) argued that this is one of the reasons why the natural sciences have fared better than the behavioral sciences over recent decades. This provocative argument ignores some differences between the subject matter in the natural sciences and that in the behavioral sciences (e.g., Lykken, 1991). But there seems little doubt that collective overconfidence in significance testing has handicapped the behavioral sciences.

Times of change present opportunities for both progress and peril. Guthery et al. (2001) counted the following potential advantages of the decline of significance testing: Researchers might pay less attention to statistical hypotheses and more attention to the good, creative ideas that drive scientific progress. Researchers may better resist the temptation to become preoccupied with statistical tools at the expense of seeking true cumulative knowledge. A risk is that another mechanically applied statistical ritual will simply replace significance testing. There does not appear at this point to be any contender, including Bayesian estimation, that could take the place of

significance testing across whole disciplines, so the likelihood that we will simply swap one set of bad habits for another seems remote for now.

Recommendations

Specific suggestions are listed next and then discussed:

1. Routine use of significance testing without justification is no longer acceptable.
2. If statistical tests are used, (a) information about a priori power must be reported, (b) the representativeness of the sample must be addressed (i.e., is the population inference model, which assumes random sampling, tenable?), and (c) distributional or other assumptions of the test must be verified. If nil hypotheses are tested, the researcher should explain why such hypotheses are appropriate.
3. If an a priori level of α is specified, do so based on rational grounds, not arbitrary ones. Otherwise, do not dichotomize *p* values; just report them.
4. Drop the word *significant* from our data analysis vocabulary. Use it only in its everyday sense to describe something actually noteworthy or important.
5. If scores from psychological tests are analyzed, report reliability coefficients and describe other relevant psychometric information.
6. It is the researcher's responsibility to report and interpret effect sizes and confidence intervals whenever possible. This does not mean the researcher should report effect sizes only for results with low *p* values.
7. It is also the researcher's responsibility to consider the substantive significance of the results. Statistical tests are inadequate for this purpose. This means no more knee-jerk claims of importance based solely on low *p* values.
8. Replication is the decisive way to deal with sampling error. The best journals should require evidence for replication.
9. Statistics education needs more reform than is apparent to date. The role of significance testing should be greatly reduced so that more time can be spent showing students how to determine whether a result has substantive significance and how to replicate it.
10. Researchers need more help from their statistical software to compute effect sizes and confidence intervals and also to test non-nil hypotheses.

No Unjustified Use of Statistical Tests

Today nearly all researchers use significance testing, but most fail to explain why its use makes sense in a particular study. For example, there is little point to significance testing when power is low, but few researchers estimate power or even mention this issue. This failure is critical when expected results are not statistically significant. If readers knew that power was low in such analyses, it would be clear that the absence of asterisks may be due more to the design (e.g., N is too small) than to the validity of the research hypotheses. We probably see so few examples of reporting power when results are mainly negative because of bias for publishing studies with H_0 rejections. In a less biased literature, p values that exaggerate the relative infrequency of the results are expected under implausible null hypotheses. If it is feasible to test only a nil hypothesis but such a null hypothesis is dubious, interpretation of statistical test outcomes should be modified accordingly.

If the use of significance testing is justifiable, other suggested reforms are relevant. One is to specify the level of α based on rational grounds that also take account of the serious of Type II error. If a researcher cannot think of a reason to set the level of statistical significance to a value other than the "defaults" of .05 or .01, that researcher has not thought sufficiently about the problem. This also means that significance testing should be applied in an informed way. An alternative is to just report p values without dichotomizing them. If so, the word *significant* would not apply to any result, but the researcher should be careful not to base interpretations on undisclosed dichotomization of p values (e.g., results are not interpreted unless $p < .05$).

The capability of significance tests to address the dichotomous question of whether effects, relations, or differences are greater than expected levels of sampling error may be useful in some new research areas. Due to the many limitations of statistical tests, this period of usefulness should be brief. Given evidence that an effect exists, the next steps should involve estimation of its magnitude and evaluation of its substantive significance, both of which are beyond what significance testing can tell us. More advanced study of the effect may require statistical modeling techniques (Rodgers, 2010). It should be a hallmark of a maturing research area that significance testing is not the primary inference method.

Report Psychometrics for Test Scores

There is a misconception that reliability is an attribute of tests rather than of the scores for a particular population of examinees (B. Thompson, 2003). This misconception may discourage researchers from reporting the reliabilities of their own data. Interpretation of effect size estimates also requires assessments of score reliability (Wilkinson & the TFSI, 1999). The best type

of estimate is calculated in the researcher's own sample. If these coefficients are satisfactory, readers are reassured that the scores were reasonably precise. Reliability induction is a second-best practice where only coefficients from previous studies are reported. Too many authors who depend on reliability induction fail to explicitly compare characteristics of their sample with those from cited studies of score reliability (e.g., Vacha-Haase & Thompson, 2011).

Report and Interpret Effect Sizes and Confidence Intervals

That some journals require effect sizes supports this recommendation. Reporting confidence intervals for effect sizes is even better: Not only does the width of the confidence interval directly indicate the amount of sampling error associated with a particular effect size, it also estimates a range of effect sizes in the population that may have given rise to the observed result. Although it not always possible to compute effect sizes in certain kinds of complex designs or construct confidence intervals based on some types of statistics, this concerns a minority of studies. In contrast to authors who recommend calculating effect sizes only for statistically significant results (Onwuegbuzie & Levin, 2003), I urge reporting of effect sizes for all substantive analyses, especially if power is low or *p* values are deemed untrustworthy.

Demonstrate Substantive Significance

Null hypothesis rejections do not imply substantive significance, so researchers need other frames of reference to explain to their audiences why the results are interesting or important. A start is to learn how to describe your results without mention of statistical significance at all. In its place, refer to descriptive statistics and effect sizes and explain why those effect sizes matter in a particular context. Doing so may seem odd at first, but you should understand that statistical tests are not generally necessary to detect meaningful or noteworthy effects, which should be obvious to visual inspection of relatively simply kinds of graphical displays (Cohen, 1994). The description of results at a level closer to the data may also help researchers to develop better communication skills.

Replicate, Replicate

The rationale for this recommendation is obvious. A replication requirement would help to filter out some of the fad research topics that bloom for a short time but then disappear. Such a requirement could be relaxed for original results with the potential for a large impact in their field, but the need to replicate studies with unexpected or surprising results is even greater (Robinson & Levin, 1997). Chapter 9 deals with replication in more detail.

Statistics Education Should Be Less Significance-Centric

Significance testing is presented as the pinnacle in many introductory statistics courses. Graduate courses often do little more than inform students about additional kinds of statistical tests and strategies for their use. The situation is little better in undergraduate psychology programs, which emphasize traditional approaches to analysis (i.e., statistical tests) and have not generally kept pace with changes in the field. And too many statistics textbooks still fail to tell students about effect size estimation (Capraro & Capraro, 2002).

Some topics already taught in introductory courses should be given more prominence. Many effect sizes are nothing more than correlations, proportions of standard deviations, or percentages of scores that fall at certain points. These are all basic kinds of statistics covered in many introductory courses. But potential application outside classical descriptive or inferential statistics is often unexplained. For example, students usually learn about the *t* test for comparing independent means. The same students often do not know about the point-biserial correlation, r_{pb}. In a two-sample design, r_{pb} is the correlation between a dichotomous independent variable (group membership) and a quantitative dependent variable. It is easily derived from the *t* test and is just a special case of the Pearson correlation, *r*. These ideas are elaborated in the next chapter.

Better integration between courses in research methods and statistics is also needed. In many undergraduate programs, these subjects are taught in separate courses, and there is often little connection between the two. The consequence is that students learn about data analysis methods without getting a good sense of their potential applications. This may be an apt time to rethink the partition of the teaching of research skills into statistics versus methods courses.

Statistical Software Should Be Modernized

Most general statistical software programs are still overly significance-test-centric. That more of them now optionally print at least some kinds of effect sizes is encouraging. It should also be the case that, for a given analytical choice, different kinds of effect size are options. Given these discussions, perhaps it is results of statistical tests that should be the optional output. Some programs also optionally print confidence intervals based on means or regression coefficients, but they should give confidence intervals based on effect sizes, too.

CONCLUSION

Significance testing has been like a collective Rorschach inkblot test for the behavioral sciences: What we see in it has more to do with wish fulfillment than reality. This magical thinking has impeded the development of psychology and other disciplines as cumulative sciences. There would be no

problem with significance testing if researchers all routinely specified plausible null hypotheses, set the level of α based on rational grounds, estimated power before collecting the data in randomly selected samples, verified distributional and other assumptions, analyzed scores with little measurement error, and understood the correct meaning of *p* values. But the gap between what is required for significance tests to be accurate and characteristics of real world studies is just too great. I offered suggestions in this chapter, all of which involve a smaller role—including none at all—for significance testing. Replication is the most important reform of all. The next chapter introduces effect size estimation in comparative studies with continuous outcomes.

LEARN MORE

Aguinis et al. (2010) and Hurlbert and Lombardi (2009) give spirited defenses of modified forms of significance testing. Ziliak and McCloskey (2008) deliver an eloquent but hard-hitting critique of significance testing, and Lambdin (2012) takes psychology to task for its failure to abandon statistical witchcraft.

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods, 13,* 515–539. doi:10.1177/1094428109333339

Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman–Pearson decision theory framework and rise of the neoFisherian. *Annales Zoologici Fennici, 46,* 311–349. Retrieved from http://www.sekj.org/AnnZool.html

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory & Psychology, 22,* 67–90. doi:10.1177/0959354311429854

Ziliak, S., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives.* Ann Arbor: University of Michigan Press.

EXERCISES

Explain why each statement about statistical significance listed next is incorrect.

1. Statistically significant: "Said of a sample size which is large enough to be considered representative of the overall population being studied."[2]

[2]http://www.investorwords.com/4704/statistically_significant.html

2. "Many researchers get very excited when they have discovered a 'statistically significant' finding, without really understanding what it means. When a statistic is significant, it simply means that you are very sure that the statistic is reliable."[3]
3. "Statistical tests are used because we want to do the experiment once and avoid the enormous cost of repeating it many times. The test will tell us how likely a particular mean difference would be to occur by chance; those unlikely to occur by chance are termed significant differences and form the basis for scientific conclusions" (M. K. Johnson & Liebert, 1977, p. 60).
4. "A long-standing convention in psychology is to label results as *statistically significant* if the probability is less than 5% that the research hypothesis is wrong" (Gray, 2002, p. 41).
5. "The message here is that in judging a study's results, there are two questions. First, is the result statistically significant? If it is, you can consider there to be a real effect. The next question is then, is the effect size large enough for the result to be useful or interesting" (Aron & Aron, 2002, p. 147).

[3]http://www.statpac.com/surveys/statistical-significance.htm

This page intentionally left blank

This page intentionally left blank

5

CONTINUOUS OUTCOMES

> Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude—not just, does a treatment affect people, but how much does it affect them.
>
> —Gene Glass (quoted in M. Hunt, 1997, pp. 29–30)

Effect size estimation with continuous outcomes is introduced in this chapter. Also considered are conceptual issues and limitations of effect size estimation including the challenge of establishing substantive significance. Two major effect size types, standardized mean differences and measures of association, are described. These effect sizes are among the most widely reported in the literature, both in primary studies and in meta-analytic studies. Interval estimation for effect sizes is also covered. Research designs considered next compare only two independent or dependent samples, but later chapters extend effect size estimation to more complex designs. Exercises for this chapter involve the computation and interpretation of effect size measures, which are introduced next.

DOI: 10.1037/14136-005
Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, Second Edition, by R. B. Kline

DEFINITIONS OF EFFECT SIZE

Kelley and Preacher (2012) defined **effect size** as a quantitative reflection of the magnitude of some phenomenon used for the sake of addressing a specific research question. In this sense, an effect size is a statistic (in samples) or parameter (in populations) with a purpose, that of quantifying a phenomenon of interest. More specific definitions may depend on study design. For example, effect size in experimental studies usually refers to the magnitude of the impact of the independent variable on the dependent (outcome) variable. Thus, effect size is measured on the latter. In contrast, **cause size** refers to the independent variable and specifically to the amount of change in it that produces a given effect on the dependent variable. A related idea is that of **causal efficacy**, or the ratio of effect size to the size of its cause. The greater the causal efficacy, the more that a given change on an independent variable results in proportionally bigger changes on the dependent variable. The idea of cause size is most relevant when the factor is experimental and its levels are quantitative.

In nonexperimental studies, effect size can be described as the degree of covariation between variables of interest. If there is a distinction between predictor and criterion variables, effect size is generally measured on the criterion, when doing so addresses a question of interest. But there are times when there is no clear distinction between predictors and criteria. In this case, an effect size would correspond to something more akin to a correlation (standardized) or covariance (unstandardized) than to a regression coefficient. Kelley and Preacher (2012) noted that even simpler kinds of outcomes, such as proportions, can be considered as effect sizes if they describe what exists in a sample or population in a way that addresses a particular question.

An **effect size measure**, as defined by Kelley and Preacher (2012), is a named expression that maps data, statistics, or parameters onto a quantity that represents the magnitude of the phenomenon of interest. This expression connects dimensions or generalized units that are abstractions of variables of interest with a specific operationalization of those units. For example, the abstract quality of "variability" can be operationalized in terms of variances, standard deviations, or ranges, among other units, and the quality of "relatedness" can be quantified in units that correspond to correlations, regression coefficients, or covariances, among other possibilities. An effect size measure is thus a particular implementation of the dimension(s) of interest. Likewise, an **effect size value** is the real number (e.g., .35, 1.65) that results from applying the effect size measure to data, statistics, or parameters, and it is this outcome that is interpreted in terms of a particular research question.

A good effect size measure has the characteristics listed next (Kelley & Preacher, 2012):

1. Its scale (metric) should be appropriate for the research question. Specific metrics for effect sizes are considered below.
2. It should be independent of sample size. Recall that test statistics reflect both sample size and effect size (e.g., Equation 3.4).
3. As a point estimate, an effect size should have good statistical properties; that is, it should be unbiased, consistent (its values converge to that of the corresponding parameter as sample size increases), and efficient (it has minimum error variance).
4. The effect size is reported with a confidence interval.

Not all effect size measures considered in this book have all the properties just listed. But it is possible to report multiple effect sizes that address the same question in order to improve the communication of the results. An example is when an effect size is commonly reported in a particular literature, but its properties may not be optimal. If so, there is no problem with reporting both the "expected" effect size and one with better properties.

CONTEXTS FOR ESTIMATING EFFECT SIZE

Major contexts for effect size estimation and the difference between unstandardized and standardized effect sizes are outlined next.

Meaningful Versus Arbitrary Metrics

Examples of outcomes with meaningful metrics include salaries in dollars and post-treatment survival time in years. Means or contrasts for variables with meaningful units are **unstandardized effect sizes** that can be directly interpreted. For example, Bruce et al. (2000) evaluated the effect of moderate doses of caffeine on the 2,000-meter rowing performance times in seconds (s) of competitive male rowers. Average times in the placebo and caffeine conditions were, respectively, 415.39 s and 411.01s. The unstandardized contrast of 4.38 s is both directly interpretable and practically significant: Four seconds can make a big difference in finish order among competitive rowers over this distance.

In medical research, physical measurements with meaningful metrics are often available. Examples include milligrams of cholesterol per deciliter of blood, inches of mercury displacement in blood pressure, and patterns of cardiac cycle tracings generated by an electrocardiogram device, among others that could be considered gold standard measures in particular areas of health research. But in psychological research there are typically no "natural" units for abstract, nonphysical constructs such as intelligence, scholastic

achievement, or self-concept. Unlike in medicine, there are also typically no universally accepted measures of such constructs.

Therefore, metrics in psychological research are often arbitrary instead of meaningful. An example is the total score for a set of true-false items. Because responses can be coded with any two different numbers, the total is arbitrary. Standard scores such as percentiles and normal deviates are arbitrary, too, because one standardized metric can be substituted for another. **Standardized effect sizes** can be computed for results expressed in arbitrary metrics. Such effect sizes can also be directly compared across studies where outcomes have different scales. This is because standardized effect sizes are based on units that have a common meaning regardless of the original metric.

Summarized next are relative advantages of unstandardized versus standardized effect sizes (Baguley, 2009):

1. It is better to report unstandardized effect sizes for outcomes with meaningful metrics. This is because the original scale is lost when results are standardized.
2. Unstandardized effect sizes are best for comparing results across different samples measured on the same outcomes. Because standardized effect sizes reflect the variances in a particular sample, the basis for that standardization is not comparable when cases in one sample are more or less variable than in another.
3. Standardized effect sizes are better for comparing conceptually similar results based on different units of measure. Suppose that different rating scales are used as outcome variables in two studies of the same treatment. They measure the same construct but have different score metrics. Calculating the same standardized effect size in each study converts the results to a common scale.
4. Standardized effect sizes are affected by the corresponding unstandardized effect sizes plus characteristics of the study, including its design (e.g., between-subjects vs. within-subjects), whether factors are fixed or random, the extent of error variance, and sample base rates. This means that standardized effect sizes are less directly comparable over studies that differ in their designs or samples.
5. There is no such thing as **T-shirt effect sizes** (Lenth, 2006–2009) that classify standardized effect sizes as "small," "medium," or "large" and apply over all research areas. This is because what is considered a large effect in one area may be seen as small or trivial in another. B. Thompson (2001) advised that we should avoid "merely being stupid in another metric" (pp. 82–83) by interpreting effect sizes in the same rigid way that characterizes significance testing.

6. There is usually no way to directly translate standardized effect sizes into implications for substantive significance. This means that neither statistical significance nor observation of effect sizes considered large in some T-shirt metric warrants concluding that the results are meaningful.

Meta-Analysis

It is standardized effect sizes from sets of related studies that are analyzed in most meta-analyses. Consulting a meta-analytic study provides a way for researchers to gauge whether their own effects are smaller or larger than those from other studies. If no meta-analytic study yet exists, researchers can calculate, using equations presented later, effect sizes based on descriptive or test statistics reported by others. Doing so permits direct comparison of results across different studies of the same phenomenon, which is part of meta-analytic thinking.

Power Analysis and Significance Testing

A priori power analysis requires specification of population effect sizes, or the parameters of statistics introduced next. Thus, one needs to know about effect size in order to use a computer tool for power analysis. Estimating sample effect sizes can help to resolve two interpretational quandaries that can arise in significance testing: Trivial effects can lead to rejection of H_0 in large samples, and it may be difficult to reject H_0 in small samples even for larger effects. Measurement of effect magnitudes apart from the influence of sample size distinguishes effect size estimation from significance testing (see Equation 3.4).

LEVELS OF ANALYSIS

Effect sizes for analysis at the group or variable level are based on aggregated scores. Consequently, they do not directly reflect the status of individual cases, and there are times when group- or variable-level effects do not tell the whole story. Knowledge of descriptive statistics including correlation coefficients is required in order to understand group- or variable-level effect sizes. Not so for case-level effect sizes, which are usually proportions of scores that fall above or below certain reference points. These proportions may be observed or predicted, and the reference points may be relative, such as the median of one group, or more absolute, such as a minimum score on an admissions test. Huberty (2002) referred to such effect sizes as **group overlap indexes**, and they are suitable for communication with general audiences. There is an old saying that goes, "The more you know, the more simply you

should speak." Case-level analysis can help a researcher do just that, especially for audiences who are not formally trained in research.

FAMILIES OF EFFECT SIZES

There are two broad classes of standardized effect sizes for analysis at the group or variable level, the ***d* family**, also known as **group difference indexes**, and the ***r* family**, or **relationship indexes** (Huberty, 2002; Rosenthal et al., 2000). Both families are **metric- (unit-) free effect sizes** that can compare results across studies or variables measured in different original metrics. Effect sizes in the *d* family are **standardized mean differences** that describe mean contrasts in standard deviation units, which can exceed 1.0 in absolute value. Standardized mean differences are **signed effect sizes**, where the sign of the statistic indicates the direction of the corresponding contrast.

Effect sizes in the *r* family are scaled in correlation units that generally range from −1.0 to +1.0, where the sign indicates the direction of the relation between two variables. For example, the point-biserial correlation r_{pb} is an effect size for designs with two unrelated samples, such as treatment versus control, and a continuous outcome. It is a form of the Pearson correlation *r* in which one of the two variables is dichotomous. If $r_{pb} = .30$, the correlation between group membership and outcome is .30, and the former explains $.30^2 = .09$, or 9%, of the total variance in the latter. A squared correlation is a **measure of association**, which is generally a **proportion of variance explained effect size**. Measures of association are **unsigned effect sizes** and thus do not indicate directionality.

Because squared correlations can make some effects look smaller than they really are in terms of their substantive significance, some researchers prefer unsquared correlations. If $r = .30$, for example, it may not seem very impressive to explain $.30^2 = .09$, or <10%, of the total variance. McCloskey and Ziliak (2009) described examples in medicine, education, and other areas where potentially valuable findings may have been overlooked due to misinterpretation of squared correlations. Rutledge and Loh (2004) calculated correlation effect sizes for 15 widely cited studies in behavioral health (e.g., heart disease, smoking, depression). They found that proportions of explained variance were typically <.10, yet these studies are considered to be landmark investigations that demonstrated clinically meaningful results. For example, the Steering Committee of the Physicians' Health Study Research Group (1988) found that the clinical value of small doses of aspirin in preventing heart attack was so apparent that it terminated a randomized clinical trial early so that the results could be reported. The correlation effect size was .034. This means that taking aspirin versus placebo explained about

$.034^2 = .0012$, or .12%, of the variability in cardiovascular health outcomes. See Ferguson (2009) for cautions about comparing effect sizes in medicine with those in psychology or other behavioral sciences.

Due to capitalization on chance, squared sample correlations estimate population proportions of explained variance with positive bias. This is a greater problem when the sample size is small. There are methods to correct squared correlations for bias, and they do not all yield the same results for the same effect. Methods described in the regression literature generate corrected squared multiple correlations such that $\hat{R}^2 < R^2$, where $\hat{R}^2$ is a bias-adjusted result that controls for sample size and the number of predictors (e.g., Snyder & Lawson, 1993). In very large samples, $\hat{R}^2$ and R^2 are asymptotically equal (i.e., there is virtually no bias). In smaller samples, some researchers prefer $\hat{R}^2$ over R^2 because the former is a more conservative estimator of ρ^2, the population proportion of explained variance.

The technique of ANOVA is just a special case of multiple regression, but in the ANOVA literature R^2 is often called **estimated eta-squared**, $\hat{\eta}^2$. Two bias-adjusted estimators for designs with fixed factors are **estimated omega-squared**, $\hat{\omega}^2$, and **estimated epsilon-squared**, $\hat{\varepsilon}^2$. Note that some authors use the symbols η^2, ω^2, and ε^2 to refer to sample statistics, but this is potentially confusing because Greek letters without the hat symbol (^) usually refer to parameters (e.g., μ). It is generally true for the same data that

$$\hat{\eta}^2 > \hat{\varepsilon}^2 > \hat{\omega}^2$$

but their values converge in large samples. The effect size $\hat{\omega}^2$ is reported more often than $\hat{\varepsilon}^2$, so the latter is not covered further; see Olejnik and Algina (2000) for more information about $\hat{\varepsilon}^2$. Kirk (1996) described a category of miscellaneous effect size indexes that includes some statistics not described in this book, including the binomial effect size display and the counternull value of an effect size; see also Rosenthal et al. (2000), Ellis (2010), and Grissom and Kim (2011).

STANDARDIZED MEAN DIFFERENCES

The parameter estimated by a sample standardized mean difference is

$$\delta = \frac{\mu_1 - \mu_2}{\sigma^*} \tag{5.1}$$

where the numerator is the population mean contrast and the denominator is a population standard deviation on the outcome variable. The

parameter δ (Greek lowercase delta) expresses the contrast as the proportion of a standard deviation. For example, $\delta = .75$ says that the mean of population 1 is three quarters of a standard deviation higher than the mean of population 2. Likewise, $\delta = -1.25$ says that the mean of the first population is 1¼ standard deviations lower than the mean of the second. The sign of δ is arbitrary because the direction of the subtraction between the two means is arbitrary. Always indicate the meaning of the sign for an estimate of δ.

The denominator of δ is σ^*, a population standard deviation (Equation 5.1). This denominator is the **standardizer** for the contrast. There is more than one population standard deviation in a comparative study. For example, σ^* could be the standard deviation in just one of the populations (e.g., $\sigma^* = \sigma_1$), or, assuming homoscedasticity, it could be the common population standard deviation (i.e., $\sigma^* = \sigma_1 = \sigma_2$). Because there is more than one potential standardizer, you should always describe the denominator in any estimate of δ.

The general form of a sample standardized mean difference is

$$d = \frac{M_1 - M_2}{\hat{\sigma}^*} \tag{5.2}$$

where the numerator is the observed contrast and the denominator is an estimator of σ^* that is not the same in all kinds of d statistics. This means that d statistics with different standardizers can—and usually do—have different values for the same contrast. Thus, to understand what a particular d statistic measures, you need to know which population standard deviation its standardizer estimates. This is critical because there are no standard names for d statistics. For example, some authors use the term Cohen's d to refer any sample standardized mean difference, but Cohen (1988) used the symbol d to refer to δ, the parameter. Others authors use the same term to refer to d statistics with a particular standardizer. This ambiguity in names is why specific d statistics are designated in this book by their standardizers.

Presented in the top part of Table 5.1 are the results of two hypothetical studies in which the same group contrast is measured on variables that reflect the same construct but with different scales. The unstandardized contrast is larger in the first study (75.00) than in the second (11.25), but the estimated population standard deviation is greater in the first study (100.00) than in the second (15.00). Because $d_1 = d_2 = .75$, we conclude equal effect size magnitudes in standard deviation units across studies 1 and 2. Reported in the bottom part of the table are results of two other hypothetical studies with the same unstandardized contrast, 75.00. Because the standard devia-

TABLE 5.1
Standardized Mean Differences for Two Hypothetical Contrasts

Study	$M_1 - M_2$	$\hat{\sigma}^*$	d
	Different mean contrast, same effect size		
1	75.00	100.00	.75
2	11.25	15.00	.75
	Same mean contrast, different effect size		
3	75.00	500.00	.15
4	75.00	50.00	1.50

tion in the third study (500.00) is greater than that in the fourth (50.00), we conclude unequal effect sizes across studies 3 and 4 because $d_3 = .15$ and $d_4 = 1.50$.

Specific types of d statistics seen most often in the literature and their corresponding parameters are listed in Table 5.2 and are discussed next (see also Keselman et al., 2008). From this point, the subscript for d designates its standardizer.

d_{pool}

The parameter estimated by d_{pool} is $\delta = (\mu_1 - \mu_2)/\sigma$, where the denominator is the common population standard deviation assuming homoscedasticity. The estimator of σ is s_{pool}, the square root of the pooled within-groups variance (Equation 2.13). When the samples are independent, d_{pool} can also be calculated given just the group sizes and the value of t_{ind}, the independent samples t statistic with $df_W = N - 2$ degrees of freedom for a nil hypothesis:

$$d_{pool} = t_{ind}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{5.3}$$

This equation is handy when working with secondary sources that do not report sufficient group descriptive statistics to calculate d_{pool} as $(M_1 - M_2)/s_{pool}$. It is also possible to transform the correlation r_{pb} to d_{pool} for the same data:

$$d_{pool} = r_{pb}\sqrt{\left(\frac{df_W}{1 - r_{pb}^2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{5.4}$$

TABLE 5.2
Types of Standardized Mean Differences for Two-Sample Designs

Statistic	Equation	Parameter
	Nonrobust	
d_{pool}	$\frac{M_1 - M_2}{s_{pool}}$	$\frac{\mu_1 - \mu_2}{\sigma}$
d_{s1}	$\frac{M_1 - M_2}{s_1}$	$\frac{\mu_1 - \mu_2}{\sigma_1}$
d_{s2}	$\frac{M_1 - M_2}{s_2}$	$\frac{\mu_1 - \mu_2}{\sigma_2}$
d_{total}	$\frac{M_1 - M_2}{s_T}$	$\frac{\mu_1 - \mu_2}{\sigma_{total}}$
d_{diff}	$\frac{M_D}{s_D}$	$\frac{\mu_D}{\sigma\sqrt{2\,(1 - \rho_{12})}}$
	Robust	
$d_{Win\ p}$	$\frac{M_{tr1} - M_{tr2}}{s_{Win\ p}}$	$\frac{\mu_{tr1} - \mu_{tr2}}{\sigma_{Win}}$
d_{Win1}	$\frac{M_{tr1} - M_{tr2}}{s_{Win1}}$	$\frac{\mu_{tr1} - \mu_{tr2}}{\sigma_{Win1}}$
d_{Win2}	$\frac{M_{tr1} - M_{tr2}}{s_{Win2}}$	$\frac{\mu_{tr1} - \mu_{tr2}}{\sigma_{Win2}}$

Equation 5.4 shows that d_{pool} and r_{pb} describe the same contrast but in different standardized units. An equation that converts d_{pool} to r_{pb} is presented later.

In correlated designs, d_{pool} is calculated as M_D/s_{pool}, where M_D is the dependent mean contrast. The standardizer s_{pool} in this case assumes that the cross-conditions correlation r_{12} is zero (i.e., any subjects effect is ignored). The parameter estimated when the samples are dependent is μ_D/σ. The value of d_{pool} can also be computed from t_{dep}, the dependent samples t with $n - 1$ degrees of freedom for a nil hypothesis, and group size, the within-condition variances, and the variance of the difference scores (Equation 2.21) as

$$d_{pool} = t_{dep}\sqrt{\frac{2s_D^2}{n\,(s_1^2 + s_2^2)}} \tag{5.5}$$

d_{s1} or d_{s2}

The standardizer s_{pool} assumes homoscedasticity. An alternative is to specify the standard deviation in either group, s_1 or s_2, as the standardizer. If one group is treatment, the other is control, and treatment is expected to affect both central tendency and variability, it makes sense to specify the standardizer as s_{con}, the standard deviation in the control group. The resulting standardized mean difference is $(M_1 - M_2)/s_{con}$, which some authors call **Glass's delta**. It estimates the parameter $(\mu_1 - \mu_2)/\sigma_{con}$, and its value describes the treatment effect only on means.

Suppose that the two groups do not correspond to treatment versus control, and their variances are heterogeneous, such as $s_1^2 = 400.00$ and $s_2^2 = 25.00$. Rather than pool these dissimilar variances, the researcher will specify one of the group standard deviations as the standardizer. Now, which one? The choice determines the value of the resulting d statistic. Given $M_1 - M_2 = 5.00$, for instance, the two possible results for this example are

$$d_{s1} = \frac{5.00}{\sqrt{400.00}} = .25 \quad \text{or} \quad d_{s2} = \frac{5.00}{\sqrt{25.00}} = 1.00$$

The statistic d_{s2} indicates a contrast four times larger in standard deviation units than d_{s1}. The two results are equally correct if there are no conceptual grounds to select one group standard deviation or the other as the standardizer. It would be best in this case to report values of both d_{s1} and d_{s2}, not just the one that gives the most favorable result. When the group variances are similar, d_{pool} is preferred because s_{pool} is based on larger sample sizes (yielding presumably more precise statistical estimates) than s_1 or s_2. But if the ratio of the largest over the smallest variance exceeds, say, 4.0, then d_{s1} or d_{s2} would be better.

d_{total}

Olejnik and Algina (2000) noted that s_{pool}, s_1, and s_2 estimate the full range of variation for experimental factors but perhaps not for individual difference (nonexperimental) factors. Suppose there is a substantial gender difference on a continuous variable. In this case, s_{pool}, s_1, and s_2 all reflect a partial range of individual differences. The unbiased variance estimator for the whole data set is $s_T^2 = SS_T/df_T$, where the numerator and denominator are, respectively, the total sum of squares and total degrees of freedom, or $N - 1$. Gender contrasts standardized against s_T would be smaller in absolute value than when the standardizer is s_{pool}, s_1, or s_2, assuming a group difference. Whether standardizers reflect partial or full ranges of variability is a crucial problem in factorial designs and is considered in Chapter 8.

Correction for Positive Bias

Absolute values of d_{pool}, d_{s1}, d_{s2}, and d_{total} are positively biased, but the degree of bias is slight unless the group sizes are small, such as $n < 20$. Multiplication of any of these statistics by the correction factor

$$c(df) = 1 - \frac{3}{4df - 1} \tag{5.6}$$

where df refers to the standardizer's degrees of freedom, yields a numerical approximation to the unbiased estimator of δ. Suppose that d_{pool} is calculated in a balanced two-sample design where $n = 10$. The degrees of freedom are $df_W = 18$, so the approximate unbiased estimator is .9578 d_{pool}. But for $n = 20$ and $df_W = 38$, the approximate unbiased estimator is .9801 d_{pool}. For even larger group sizes, the correction factor is close to 1.0, which implies little adjustment for bias. Some authors refer to $c\,(df)\,d_{pool}$ as **Hedges's g**, but others apply this term to d_{pool}. Given this ambiguity, I do not recommend using the term to describe either statistic.

Suppose that the means and variances of two samples in a balanced design (i.e., the groups have the same number of cases, n) are

$$M_1 = 13.00,\ s_1^2 = 7.50 \quad \text{and} \quad M_2 = 11.00,\ s_2^2 = 5.00$$

which implies $M_1 - M_2 = 2.00$ and $s_{pool}^2 = 6.25$. Reported in Table 5.3 are results of the independent samples t test and values of d statistics for $n = 5$, 15, and 30. The t test shows the influence of group size. In contrast, $d_{pool} = .80$ for all three analyses and in general is invariant to group size, keeping all else constant. The approximate unbiased estimator $c\,(df_W)\,d_{pool}$ is generally less than d_{pool}, but their values converge as n increases. The two possible values of d for these data when the standardizer is a group standard deviation are $d_{s1} = .73$ and $d_{s2} = .89$. Values of d_{total} are generally similar to those of other d statistics for these data (see Table 5.3), but, in general, d_{total} is increasingly dissimilar to d_{pool}, d_{s1}, and d_{s2} for progressively larger contrasts on nonexperimental factors. Exercise 1 asks you to verify some of the results in Table 5.3.

d_{diff} for Dependent Samples

Standardized mean differences in correlated designs are called **standardized mean changes** or **standardized mean gains**. There are two different ways to standardize contrasts in these designs. The first method does so just as one would in designs with independent samples (i.e., calculate any of the d statistics described so far). For example, one possibility is $d_{pool} = M_D/s_{pool}$,

TABLE 5.3
Results of the *t* Test and Effect Sizes at Three Different Group Sizes

Statistic	Group size (*n*)		
	5	15	30
	t test		
t	1.26	2.19	3.10
df_W	8	28	58
p	.242	.037	.003
	Standardized mean differences		
d_{pool}	.80	.80	.80
$c\,(df_W)\ d_{pool}$	.72	.78	.79
d_{s1}	.73	.73	.73
d_{s2}	.89	.89	.89
d_{total}	.77	.75	.75
	Point-biserial correlation		
r_{pb}	.41	.38	.38

Note. For all analyses, $M_1 = 13.00$, $s_1^2 = 7.50$, $M_2 = 11.00$, $s_2^2 = 5.00$, $s_{pool}^2 = 6.25$, and p values are two-tailed and for a nil hypothesis.

where the standardizer is the pooled within-conditions standard deviation that assumes homoscedasticity. This assumption may be untenable in a repeated measures design, however, when treatment is expected to change variability among cases from pretest to posttest. A better alternative in this case is d_{s1}, where the denominator is the standard deviation from the pretest condition. Some authors refer to M_D/s_1 as **Becker's g**.

The second method is to calculate a standardized mean change as $d_{diff} = M_D/s_D$, where the denominator is the standard deviation of the difference scores, which takes account of the cross-conditions correlation r_{12}. In contrast, s_{pool}, s_1, and s_2 all assume that this correlation is zero. If r_{12} is reasonably high and positive, it can happen that s_D is smaller than the other standardizers just mentioned. This implies that d_{diff} can be bigger in absolute value than d_{pool}, d_{s1}, and d_{s2} for the same contrast. The effect size d_{diff} estimates the parameter

$$\delta = \frac{\mu_D}{\sigma\sqrt{2\,(1-\rho_{12})}} \tag{5.7}$$

where σ and ρ_{12} are, respectively, the common population standard deviation and cross-conditions correlation. Note in this equation that the denominator is less than σ only if $\rho_{12} > .50$.

A drawback is that d_{diff} is scaled in the metric of difference scores, not original scores. An example from Cumming and Finch (2001) illustrates this problem: A verbal test is given before and after an intervention. Test scores are based on the metric $\mu = 100.00$, $\sigma = 15.00$. The observed standard deviations at both occasions are also 15.00, and the standard deviation of the difference scores is 7.70. The result is $M_D = 4.10$ in favor of the intervention. If our natural reference for thinking about scores on the verbal test is their original metric, it makes sense to report a standardized mean change as 4.10/15.00, or .27, instead of 4.10/7.70, or .53. This is true even though the latter standardized effect size estimate is about twice as large as the former.

Cortina and Nouri (2000) argued that *d* should have a common meaning regardless of the design, which implies a standardizer in the metric of the original scores. This advice seems sound for effects that could theoretically be studied in either between-subjects or within-subjects designs. But standardized mean differences based on s_D may be preferred when the emphasis is on measurement of change. The researcher should explain the choice in any event. Exercise 2 asks you to verify that $d_{pool} = .80$ but $d_{diff} = 1.07$ for the data in Table 2.2.

Robust Standardized Mean Differences

The *d* statistics described so far are based on least squares estimators (i.e., *M*, *s*) that are not robust against outliers, nonnormality, or heteroscedasticity. Algina, Keselman, and Penfield (2005a, 2005b) described the robust *d* statistics listed in Table 5.2. One is

$$d_{\text{Win p}} = \frac{M_{tr1} - M_{tr2}}{s_{\text{Win p}}} \tag{5.8}$$

where $M_{tr1} - M_{tr2}$ is the contrast between 20% trimmed means and $s_{\text{Win p}}$ is the 20% pooled Winsorized standard deviation that assumes homoscedasticity. The latter in a squared metric is

$$s^2_{\text{Win p}} = \frac{df_1(s^2_{\text{Win1}}) + df_2(s^2_{\text{Win2}})}{df_W} \tag{5.9}$$

where s^2_{Win1} and s^2_{Win2} are the 20% Winsorized group variances and $df_1 = n_1 - 1$, $df_2 = n_2 - 1$, and $df_W = N - 2$. The parameter estimated by $d_{\text{Win p}}$ is

$$\delta_{rob} = \frac{\mu_{tr1} - \mu_{tr2}}{\sigma_{\text{Win}}} \tag{5.10}$$

which is not $\delta = (\mu_1 - \mu_2)/\sigma$ when population distributions are skewed or heteroscedastic. Otherwise, multiplication of $d_{\text{Win p}}$ by the scale factor .642 estimates δ when the scores are selected from normal distributions with equal variances for 20% trimming and Winsorization.

If it is unreasonable to assume equal Winsorized population variances, two alternative robust effect sizes are

$$d_{\text{Win1}} = \frac{M_{\text{tr1}} - M_{\text{tr2}}}{s_{\text{Win1}}} \quad \text{or} \quad d_{\text{Win2}} = \frac{M_{\text{tr1}} - M_{\text{tr2}}}{s_{\text{Win2}}} \tag{5.11}$$

where the standardizer is the Winsorized standard deviation in either group 1 or group 2. The product .642 d_{Win1} estimates $(\mu_1 - \mu_2)/\sigma_1$, and the product .642 d_{Win2} estimates $(\mu_1 - \mu_2)/\sigma_2$, assuming normality for 20% trimming and Winsorization (see Table 5.2).

Look back at Table 2.4, in which raw scores with outliers for two groups ($n = 10$ each) are presented. Listed next are the 20% trimmed means and Winsorized variances for each group:

$$M_{\text{tr1}} = 23.00,\ s^2_{\text{Win1}} = 18.489 \quad \text{and} \quad M_{\text{tr2}} = 17.00,\ s^2_{\text{Win2}} = 9.067$$

which implies $s^2_{\text{Win p}} = 13.778$. (You should verify this result using Equation 5.9.) The robust d statistic based on the pooled standardizer is

$$d_{\text{Win p}} = \frac{23.00 - 17.00}{\sqrt{13.778}} = 1.62$$

so the size of the trimmed mean contrast is 1.62 Winsorized standardized deviations, assuming homoscedasticity. If it is reasonable to also assume normality, the robust estimate of δ would be .642 (1.62), or 1.04. Exercise 3 asks you to calculate d_{Win1} and d_{Win2} for the same data.

Limitations of Standardized Mean Differences

Limitations considered next apply to all designs considered in this book, not just two-sample designs. Heteroscedasticity across studies limits the usefulness of d as a standardized effect size. Suppose that $M_1 - M_2 = 5.00$ in two different studies on the same outcome measure. The pooled within-groups variance is 625.00 in the first study but is only 6.25 in the second. As a consequence, d_{pool} for these two studies reflects the differences in their variances:

$$d_{\text{pool1}} = \frac{5.00}{\sqrt{625.00}} = .20 \quad \text{and} \quad d_{\text{pool2}} = \frac{5.00}{\sqrt{6.25}} = 2.00$$

These results indicate a contrast tenfold greater in the second study than in the first. In this case, it is better to compare the unstandardized contrast $M_1 - M_2 = 5.00$ across studies.

CORRELATION EFFECT SIZES

The point-biserial correlation r_{pb} is the correlation between membership in one of two different groups and a continuous variable. A conceptual equation is

$$r_{pb} = \left(\frac{M_1 - M_2}{S_T} \right) \sqrt{pq} \tag{5.12}$$

where S_T is the standard deviation in the total data set computed as $(SS_T/N)^{1/2}$ and p and q are the proportions of cases in each group ($p + q = 1.0$). The expression in parentheses in Equation 5.12 is a d statistic with the standardizer S_T. It is the multiplication of this quantity by the standard deviation of the dichotomous factor, $(pq)^{1/2}$, that transforms the whole expression into correlation units. It is also possible to convert d_{pool} to r_{pb} for the same data:

$$r_{pb} = \frac{d_{pool}}{\sqrt{d_{pool}^2 + df_W \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \tag{5.13}$$

It may be easier to compute r_{pb} from t_{ind} with $df_W = N - 2$ for a nil hypothesis:

$$r_{pb} = \frac{t_{ind}}{\sqrt{t_{ind}^2 + df_W}} \tag{5.14}$$

The absolute value of r_{pb} can also be derived from the independent samples F statistic with 1, df_W degrees of freedom for the contrast:

$$|r_{pb}| = \sqrt{\frac{F_{ind}}{F_{ind} + df_W}} = \sqrt{\frac{SS_A}{SS_T}} = \hat{\eta} \tag{5.15}$$

This equation also shows that $|r_{pb}|$ is a special case of $\hat{\eta}^2 = SS_A/SS_T$, where SS_A is the between-groups sum of squares for the dichotomous factor A. In particular, $r_{pb}^2 = \hat{\eta}^2$ in a two-group design. Note that $\hat{\eta}$ is an unsigned correlation, but r_{pb} is signed and thus indicates directionality.

Reported in the lower part of Table 5.3 are values of r_{pb} at three different group sizes, $n = 5$, 15, and 30, for the same contrast. For the smallest group size, $r_{pb} = .41$, but for the larger group sizes, $r_{pb} = .38$. This pattern illustrates a characteristic of r_{pb} and other sample correlations that approach their maximum absolute values in very small samples. In the extreme case where the size of each group is $n = 1$ and the two scores are not equal, $r_{pb} = \pm 1.00$. This happens out of mathematical necessity and is not real evidence for a perfect association. Taking $r_{pb} = .38$ as the most reasonable value, we can say that the correlation between group membership and outcome is .38 and that the former explains about $.38^2 = .144$, or 14.4%, of the total observed variance. Exercise 4 involves reproducing some of the results in Table 5.3 for r_{pb}.

Two Dependent Samples

The correlation r_{pb} is for designs with two unrelated samples. For dependent samples, we can instead calculate the correlation of which r_{pb} is a special case, $\hat{\eta}$. It is derived as $(SS_A/SS_T)^{1/2}$ whether the design is between-subjects or within-subjects. A complication is that $\hat{\eta}$ may not be directly comparable when the same factor is studied with independent versus dependent samples. This is because SS_T is the sum of SS_A and SS_W when the samples are unrelated, but it comprises SS_A, SS_S, and $SS_{A \times S}$ for dependent samples. Thus, SS_T reflects only one systematic effect (A) when the means are independent but two systematic effects (A, S) when the means are dependent.

A partial correlation that controls for the subjects effect in correlated designs assuming a nonadditive model is

$$\text{partial } \hat{\eta} = \sqrt{\frac{SS_A}{SS_A + SS_{A \times S}}} \tag{5.16}$$

where the denominator under the radical represents just one systematic effect (A). The square of Equation 5.16 is partial $\hat{\eta}^2$, a measure of association that refers to a residualized total variance, not total observed variance. Given partial $\hat{\eta}^2 = .25$, for example, we can say that factor A explains 25% of the variance controlling for the subjects effect.

If the subjects effect is relatively large, partial $\hat{\eta}^2$ can be substantially higher than $\hat{\eta}^2$ for the same contrast. This is not contradictory because only $\hat{\eta}^2$ is in the metric of the original scores. This fact suggests that partial $\hat{\eta}^2$ from a correlated design and $\hat{\eta}^2$ from a between-subjects design with the same factor and outcome may not be directly comparable. But $\hat{\eta}^2 =$ partial $\hat{\eta}^2$ when the means are unrelated because there is no subjects effect. Exercise 5 asks you to verify that $\hat{\eta}^2 = .167$ and partial $\hat{\eta}^2 = .588$ in the dependent samples analysis of the data in Table 2.2.

Robust Measures of Association

The technique of robust regression is an option to calculate *r*-type effect sizes that are resistant against outliers, nonnormality, or heteroscedasticity. A problem is that there are different robust regression methods (e.g., Wilcox, 2003), and it is not always clear which is best in a particular sample. For instance, the ROBUSTREG procedure in SAS/STAT offers a total of four methods that vary in their efficiencies (minimization of error variance) or breakdown points. It is more straightforward to work with robust *d* statistics than robust measures of association in two-sample designs.

Limitations of Measures of Association

The correlation r_{pb} (and $\hat{\eta}^2$, too) is affected by base rate, or the proportion of cases in one group versus the other, p and q. It tends to be highest in balanced designs. As the design becomes more unbalanced holding all else constant, r_{pb} approaches zero. Suppose that $M_1 - M_2 = 5.00$ and $S_T = 10.00$ in each of two different studies. The first study has equal group sizes, or $p_1 = q_1 = .50$. The second study has 90% of its cases in the first group and 10% of them in the second group, or $p_2 = .90$ and $q_2 = .10$. Using Equation 5.12, we get

$$r_{pb1} = \left(\frac{5.00}{10.00}\right)\sqrt{.50(.50)} = .25 \quad \text{and} \quad r_{pb2} = \left(\frac{5.00}{10.00}\right)\sqrt{.90(.10)} = .15$$

The values of these correlations are different even though the mean contrast and standard deviation are the same. Thus, r_{pb} is not directly comparable across studies with dissimilar relative group sizes (d_{pool} is affected by base rates, too, but d_{s1} or d_{s2} is not). The correlation r_{pb} is also affected by the total variability (i.e., S_T). If this variation is not constant over samples, values of r_{pb} may not be directly comparable. Assuming normality and homoscedasticity, *d*- and *r*-type effect sizes are related in predictable ways; otherwise, it can happen that *d* and *r* appear to say different things about the same contrast (McGrath & Meyer, 2006).

CORRECTING FOR MEASUREMENT ERROR

Too many researchers neglect to report reliability coefficients for scores analyzed. This is regrettable because effect sizes cannot be properly interpreted without knowing whether the scores are precise. The general effect of measurement error in comparative studies is to attenuate absolute stan-

dardized effect sizes and reduce the power of statistical tests. Measurement error also contributes to variation in observed results over studies. Of special concern is when both score reliabilities and sample sizes vary from study to study. If so, effects of sampling error are confounded with those due to measurement error.

There are ways to correct some effect sizes for measurement error (e.g., Baguley, 2009), but corrected effect sizes are rarely reported. It is more surprising that measurement error is ignored in most meta-analyses, too. F. L. Schmidt (2010) found that corrected effect sizes were analyzed in only about 10% of the 199 meta-analytic articles published in *Psychological Bulletin* from 1978 to 2006. This implies that (a) estimates of mean effect sizes may be too low and (b) the wrong statistical model may be selected when attempting to explain between-studies variation in results. If a fixed effects model is mistakenly chosen over a random effects model, confidence intervals based on average effect sizes tend to be too narrow, which can make those results look more precise than they really are. Underestimating mean effect sizes while simultaneously overstating their precision is a potentially serious error.

The **correction for attenuation** formula for the Pearson correlation shows the relation between r_{XY} and $\hat{r}_{XY}$, which estimates the population correlation between *X* and *Y* if scores on both variables were perfectly reliable:

$$\hat{r}_{XY} = \frac{r_{XY}}{\sqrt{r_{XX}\, r_{YY}}} \tag{5.17}$$

where r_{XX} and r_{YY} are the score reliabilities for the two variables. For example, given $r_{XY} = .30$, $r_{XX} = .80$, and $r_{YY} = .70$,

$$\hat{r}_{XY} = \frac{.30}{\sqrt{.80(.70)}} = .40$$

which says that the estimated correlation between *X* and *Y* is .40 controlling for measurement error. Because disattenuated correlations are only estimates, it can happen that $|\hat{r}_{XY}| > 1.0$.

In comparative studies where the factor is presumably measured with nearly perfect reliability, effect sizes are usually corrected for measurement error in only the outcome variable, designated next as *Y*. Forms of this correction for *d*- and *r*-type effect sizes are, respectively,

$$\hat{d} = \frac{d}{\sqrt{r_{YY}}} \quad \text{and} \quad \hat{r}_{pb} = \frac{r_{pb}}{\sqrt{r_{YY}}} \tag{5.18}$$

For example, given $d = .75$ and $r_{YY} = .90$,

$$\hat{d} = \frac{.75}{\sqrt{.90}} = .79$$

which says that the contrast is predicted to be .79 standard deviations large controlling for measurement error. The analogous correction for the correlation ratio is $\hat{\eta}^2/r_{YY}$.

Appropriate reliability coefficients are needed to apply the correction for attenuation, and best practice is to estimate these coefficients in your own samples. The correction works best when reliabilities are good, such as $r_{XX} > .80$, but otherwise it is less accurate. The capability to correct effect sizes for unreliability is no substitute for good measures. Suppose that $d = .15$ and $r_{YY} = .10$. A reliability coefficient so low says that the scores are basically random numbers and random numbers measure nothing. The disattenuated effect size is $\hat{d} = .15/.10^{1/2}$, or .47, an adjusted result over three times larger than the observed effect size. But this estimate is not credible because the scores should not have been analyzed in the first place. Correction for measurement error increases sampling error compared with the original effect sizes, but this increase is less when reliabilities are higher. Hunter and Schmidt (2004) described other kinds of corrections, such as for range restriction.

INTERVAL ESTIMATION WITH EFFECT SIZES

Links to computer tools, described next, are also available on this book's web page.

Approximate Confidence Intervals

Distributions of *d*- and *r*-type effect sizes are complex and generally follow, respectively, noncentral *t* distributions and noncentral *F* distributions. Noncentral interval estimation requires specialized computer tools. An alternative for *d* is to construct approximate confidence intervals based on hand-calculable estimates of standard error in large samples. An approach outlined by Viechtbauer (2007) is described next.

The general form of an approximate 100 $(1 - \alpha)$% confidence interval for δ is

$$d \pm s_d \left(z_{2\text{-tail}, \sigma}\right) \tag{5.19}$$

where s_d is an asymptotic standard error and $z_{2\text{-tail}, \alpha}$ is the positive two-tailed critical value of the normal deviate at the α level of statistical significance.

If the effect size is d_{pool}, d_{s1}, d_{s2}, or d_{total} (or any of these statistics multiplied by c (df); Equation 5.6) and the means are treated as independent, an approximate standard error is

$$s_{d\ ind} = \sqrt{\frac{d^2}{2\,df} + \frac{N}{n_1 n_2}} \tag{5.20}$$

where df are the degrees of freedom for the standardizer of the corresponding d statistic. An estimated standard error when treating the means based on n pairs of scores as dependent is

$$s_{d\ dep} = \sqrt{\frac{d^2}{2(n-1)} + \frac{2(1-r_{12})}{n}} \tag{5.21}$$

where r_{12} is the cross-conditions correlation. Finally, if the effect size in a dependent samples analysis is d_{diff} or c (df) d_{diff}, the asymptotic standard error is

$$s_{d\,diff} \sqrt{\frac{d^2}{2(n-1)} + \frac{1}{n}} \tag{5.22}$$

There are versions of these standard error equations for very large samples where the sample size replaces the degrees of freedom, such as N instead of $df_W = N - 2$ in Equation 5.20 for the effect size d_{pool} (e.g., Borenstein, 2009). In very large samples, these two sets of equations (N, df) give similar results, but I recommend the versions presented here if the sample size is not very large.

Suppose that $n = 30$ in a balanced design and $d_{pool} = .80$. The estimated standard error is

$$s_{d\ pool} = \sqrt{\frac{.80^2}{2(58)} + \frac{60}{30(30)}} = .2687$$

Because $z_{2\text{-tail, }.05} = 1.96$, the approximate 95% confidence interval for δ is

$$.80 \pm .2687\ (1.96)$$

which defines the interval [.27, 1.33]. This wide range of imprecision is due to the small group size ($n = 30$). Exercise 6 asks you to construct the approximate 95% confidence interval based on the same data but for a dependent contrast where $r_{12} = .75$.

A method to construct an approximate confidence interval for ρ using Fisher's transformation of the Pearson *r* was described in Chapter 2. Another method by Hunter and Schmidt (2004) builds approximate confidence intervals directly in correlation units. These approximate methods may not be very accurate when the effect size is r_{pb}. An alternative is to use a computer tool that constructs noncentral confidence intervals for η^2 based on $\hat{\eta}^2$, of which r_{pb} is a special case.

Noncentral Confidence Intervals for δ

When the means are independent, *d* statistics follow noncentral *t* distributions, which have two parameters, *df* and Δ, the noncentrality parameter (e.g., Figure 2.4). Assuming normality and homoscedasticity, Δ is related to the population effect size δ and the group sizes,

$$\Delta = \delta\sqrt{\frac{n_1 n_2}{N}} \tag{5.23}$$

When the nil hypothesis is true, $\delta = 0$ and $\Delta = 0$; otherwise, Δ has the same sign as δ. Equation 5.23 can be rearranged to express δ as a function of Δ and group sizes:

$$\delta = \Delta\sqrt{\frac{N}{n_1 n_2}} \tag{5.24}$$

Steiger and Fouladi (1997) showed that if we can obtain a confidence interval for Δ, we can also obtain a confidence interval for δ using the confidence interval transformation principle. In theory, we first construct a 100 $(1 - \alpha)$% confidence interval for Δ. The lower bound Δ_L is the value of the noncentrality parameter for the noncentral *t* distribution in which the observed *t* statistic falls at the 100 $(1 - \alpha/2)$th percentile. The upper bound Δ_U is the value of the noncentrality parameter for the noncentral *t* distribution in which the observed *t* falls at the 100 $(\alpha/2)$th percentile. If $\alpha = .05$, for example, the observed *t* falls at the 97.5th percentile in the noncentral *t* distribution where the noncentrality parameter equals Δ_L. The same observed *t* also falls at the 2.5th percentile in the noncentral *t* distribution where the noncentrality parameter is Δ_U. But we need to find which particular noncentral *t* distributions are most consistent with the data, and it is this problem that can be solved with the right computer tool. The same tool may also automatically convert the lower and upper bounds of the interval for Δ to δ units.

The resulting interval is the noncentral 100 $(1 - \alpha)$% confidence interval for δ, which can be asymmetrical around the sample value of d.

Suppose that $n = 30$ in a balanced design and $d_{pool} = .80$, which implies $t\ (58) = 3.10$ (Equation 5.3). I used ESCI (Cumming, 2012; see footnote 4, Chapter 2) to construct the 95% noncentral confidence interval for δ for these data. It returned these results

95% CI for Δ [1.0469, 5.1251]

95% CI for δ [.270, 1.323]

We can say that the observed t of 3.10 falls at the

97.5th percentile in the noncentral $t\ (58,1.0469)$ distribution,

and the same observed t falls at the

2.5th percentile in the noncentral $t\ (58,\ 5.1251)$ distribution

You can verify these results with an online noncentral t percentile calculator[1] or J. H. Steiger's Noncentral Distribution Calculator (NDC), a freely available Windows application for noncentrality interval estimation.[2] Thus, the observed effect size of $d_{pool} = .80$ is just as consistent with a population effect size as low as $\delta = .27$ as it is with a population effect size as high as $\delta = 1.32$, with 95% confidence. The approximate 95% confidence interval for δ for the same data is [.27, 1.33], which is similar to the noncentral interval just described.

Smithson (2003) described a set of freely available SPSS scripts that calculate noncentral confidence intervals based on d_{pool} in two-sample designs when the means are treated as independent.[3] Corresponding scripts for SAS/STAT and R are also available.[4] Kelley (2007) described the Methods for the Behavioral, Educational, and Social Sciences (MBESS) package for R, which calculates noncentral confidence intervals for many standardized effect sizes.[5] The Power Analysis module in STATISTICA Advanced also calculates noncentral confidence intervals based on standardized effect sizes (see footnote 5, Chapter 2). In correlated designs, distributions of d_{pool} follow neither central nor noncentral t distributions (Cumming & Finch, 2001). For dependent mean contrasts, ESCI uses Algina and Keselman's (2003) method for finding approximate noncentral confidence intervals for μ_D/σ.

[1] http://keisan.casio.com/
[2] http://www.statpower.net/Software.html
[3] http://core.ecu.edu/psyc/wuenschk/SPSS/SPSS-Programs.htm
[4] http://dl.dropbox.com/u/1857674/CIstuff/CI.html
[5] http://cran.r-project.org/web/packages/MBESS/index.html

Noncentral Confidence Intervals for η^2

Methods for constructing confidence intervals for η^2 based on noncentral *F* distributions in designs with independent samples and fixed factors were described by Fidler and Thompson (2001), Smithson (2003), and Kelley (2007). Briefly, noncentral *F* distributions for single-factor designs have three parameters, df_A, df_W, and the noncentrality parameter λ (Greek lowercase lambda). The general method for obtaining a noncentral confidence interval for η^2 is similar to that for obtaining a noncentral confidence interval for δ. Assuming the 95% confidence level, a computer tool first finds the lower bound λ_L, the noncentrality parameter of the noncentral *F* distribution in which the observed *F* for the contrast falls at the 97.5th percentile. The upper bound λ_U is the noncentrality parameter of the noncentral *F* distribution in which the observed *F* falls at the 2.5th percentile. The endpoints of the interval in λ units are then converted to η^2 units with the equation

$$\eta^2 = \frac{\lambda}{\lambda + N} \tag{5.25}$$

I use the same data as in the previous example. In a balanced design, $n = 30$, $d_{pool} = .80$, and $t\,(58) = 3.10$. These results imply $r_{pb} = .377$ (see Equation 5.13), $\hat{\eta}^2 = .142$, and $F\,(1, 58) = 3.10^2$, or 9.60. I used Smithson's (2003) SPSS scripts to compute these results:

95% CI for λ [1.0930, 26.2688]

95% CI for η^2 [.0179, .3045]

Thus, the observed effect size of $\hat{\eta}^2 = .142$ is just as consistent with a population effect size as low as $\eta^2 = .018$ as it is with a population size as high as $\eta^2 = .305$, with 95% confidence. The range of imprecision is wide due to the small sample size. You should verify with Equation 5.25 that the bounds of the confidence interval in λ units convert to the corresponding bounds of the confidence interval in η^2 units for these data.

Other computer tools that generate noncentral confidence intervals for η^2 include the aforementioned MBESS package for R and the Power Analysis module in STATISTICA Advanced. There is a paucity of programs that calculate noncentral confidence intervals for η^2 in correlated designs. This is because the distributions of η^2 in this case may follow neither central nor noncentral test distributions. An alternative is bootstrapped confidence intervals.

Bootstrapped Confidence Intervals

In theory, bootstrapping could be used to generate confidence intervals based on any effect size described in this book. It is also the method generally used for interval estimation based on robust *d* statistics. Results of computer simulation studies by Algina et al. (2005a, 2005b); Algina, Keselman, and Penfield (2006); and Keselman et al. (2008) indicated that nonparametric percentile bootstrap confidence intervals on robust *d* statistics for independent or dependent means are reasonably accurate compared with the method of noncentrality interval estimation when distributions are nonnormal or heteroscedastic. Described next are some freely available computer tools for constructing bootstrapped confidence intervals based on various robust or nonrobust estimators of δ in two-sample designs:[6]

1. The program ES Bootstrap: Independent Groups (Penfield, Algina, & Keselman, 2004b) generates bootstrapped confidence intervals for δ based on the estimator .642 $d_{\text{Win p}}$ in designs with two unrelated groups.
2. The ES Bootstrap 2 program (Penfield, Algina, & Keselman, 2006) extends this functionality for the same design to additional robust or nonrobust estimators of δ.
3. In correlated designs, ES Bootstrap: Correlated Groups (Penfield, Algina, & Keselman, 2004a) generates bootstrapped confidence intervals for δ based on various robust or nonrobust estimators.

Keselman et al. (2008) described scripts in the SAS/IML programming language that generate robust tests and bootstrapped confidence intervals in single- or multiple-factor designs with independent or dependent samples.[7] Wilcox's (2012) WRS package has similar capabilities for robust *d* statistics (see footnote 11, Chapter 2).

I used Penfield et al.'s (2004b) ES Bootstrap: Independent Groups program to construct a bootstrapped 95% confidence interval for δ based on the raw data for two groups with outliers in Table 2.4. For these data, $d_{\text{Win p}} = 1.62$, and the estimate of δ is the product of the scale factor .642 and 1.62, or 1.04. Based on the empirical sampling distribution of 1,000 generated samples, the bootstrapped 95% confidence interval returned by the computer tool is [–.19, 2.14]. Thus, the observed robust effect size of 1.04 standard deviations is just as consistent with a population effect size as low as δ = –.19 as it is with a population effect size as high as δ = 2.14, with 95% confidence. The wide range of imprecision is due to the small sample size ($N = 20$).

[6]http://plaza.ufl.edu/algina/index.programs.html
[7]http://supp.apa.org/psycarticles/supplemental/met_13_2_110/met_13_2_110_supp.html

ANALYSIS OF GROUP DIFFERENCES AT THE CASE LEVEL

The methods outlined next describe effect size at the case level.

Measures of Overlap

Presented in Figure 5.1 are two pairs of frequency distributions, each of which illustrates one of Cohen's (1988) overlap measures, U_1 and U_3. Both pairs depict $M_1 > M_2$, normal distributions and equal group sizes and variances. The shaded regions in Figure 5.1(a) represent areas where the two distributions do not overlap, and U_1 is the proportion of scores across both groups within these areas. The difference $1 - U_1$ is thus the proportion of scores within the area of overlap. If the mean difference is nil, $U_1 = 0$, but if the contrast is so great that no scores overlap, $U_1 = 1.00$. The range of U_1 is

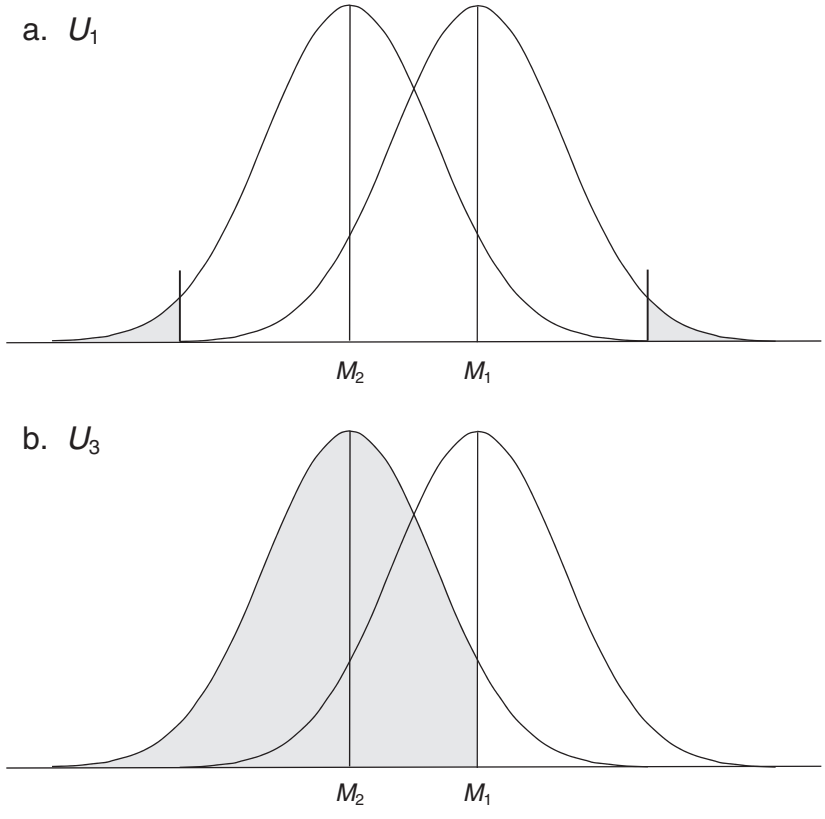

Figure 5.1. Measures of distribution overlap, U_1 (a) and U_3 (b).

thus 0–1.00. Illustrated in Figure 5.1(b) is U_3, the proportion of scores in the lower group exceeded by a typical score in the upper group. A typical score could be the mean, median, or some other measure of central tendency, but the mean and median are equal in symmetrical distributions. If two distributions are identical, $U_3 = .50$, but if $U_3 = 1.00$, the distributions are so distinct that the typical score in the upper group exceeds all scores in the lower group. The range of U_3 is thus .50–1.00.

In real data sets, U_1 and U_3 are derived by inspecting group frequency distributions. For U_1, count the total number of scores from each group outside the range of the other group and divide this number by N. For U_3, locate the typical score from the upper group in the frequency distribution of the lower group and then calculate the percentile equivalent of that score expressed as a proportion. A potential problem with U_1 is that if the range of scores is limited, the proportion of nonoverlapping scores may be zero even if the contrast is relatively large.

Suppose that treated cases have a higher mean than control cases on an outcome where a higher score is a better result. If $U_1 = .15$ and $U_3 = .55$, we can say that only 15% of the scores across the two groups are distinct. The rest, or 85%, fall within the area of overlap. Thus, most treated cases look like most untreated cases and vice versa. The typical treated case scores higher than 55% of untreated cases. Whether these results are clinically significant is another matter, but U_1 and U_3 describe case-level effects in ways that general audiences can understand.

It is also possible in graphical displays to give information about distribution overlap at the case level. The display in Figure 5.2(a) shows the means for two groups based on the scores with outliers in Table 2.4. The format of this line graphic is often used to show the results of a *t* test, but it conveys no case-level information. Plotting error bars around the dots that represent group means in Figure 5.2(a) might help, but M and s_M are not robust estimators. The display in Figure 5.2(b) with group **box plots (box-and-whisker plots)** is more informative. The bottom and top borders of the rectangle in a box plot correspond to, respectively, the 25th percentile and 75th percentile. The total region in the rectangle thus represents 50% of the scores, and comparing these regions across groups says something about overlap. The line inside the rectangle of a box plot represents the median (50th percentile). The "whiskers" are the vertical lines that connect the rectangle to the lowest and highest scores that are not extreme. The box plot for group 1 in Figure 5.2(b) has no whisker because the score at the 75th percentile (29) is also the last nonextreme score at the upper end of the distribution. Tukey (1977) described other visual methods of exploratory data analysis that show case-level information.

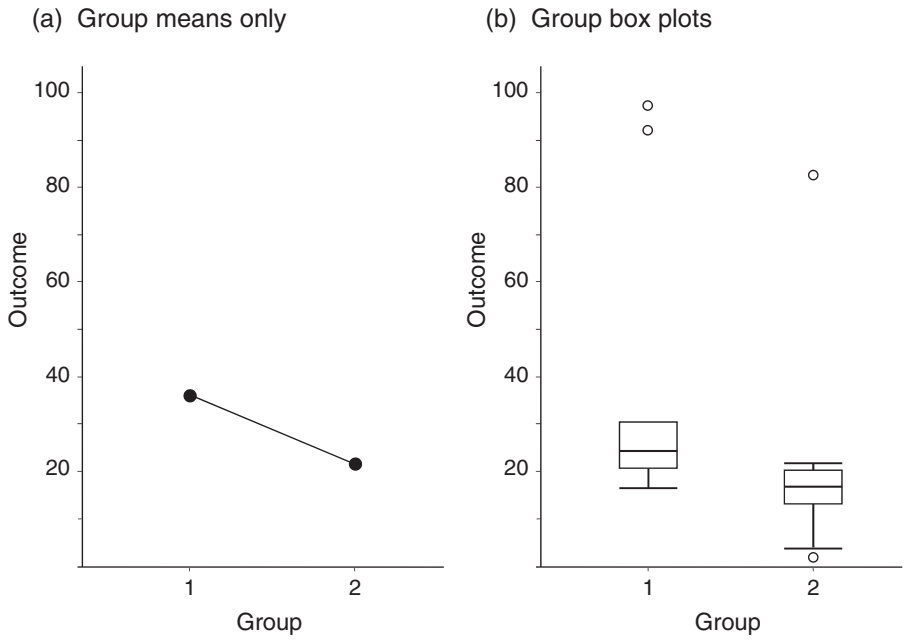

Figure 5.2. A graphical display of means only (a) versus one that shows box plots (b) for the scores from two groups with outliers in Table 2.4.

Tail Ratios

A **right-tail ratio** (RTR) is the relative proportion of scores from two different groups that fall beyond a cutting point in the upper tails of both distributions. Such thresholds may be established based on merit, such as a minimum score on an admissions test. Likewise, a **left-tail ratio** (LTR) is the relative proportion of scores that fall below a cutting point in the lower extremes of both distributions, which may be established based on need. An example of a needs-based classification is a remedial (compensatory) program available only for students with low reading test scores. Students with higher scores would not be eligible.

Because tail ratios are computed with the largest proportion in the numerator, their values are ≥1.00. For example, RTR = 2.00 says that cases from the group represented in the numerator are twice as likely as cases in the other group to have scores above a threshold in the upper tail. Presented in Figure 5.3(a) are two frequency distributions where $M_1 > M_2$. The shaded areas in each distribution represent the proportion of scores in group 1 (p_1) and group 2 (p_2) that exceed a cutting point in the right tails. Because relatively more scores from group 1 exceed the threshold, $p_1 > p_2$ and RTR = $p_1/p_2 > 1.00$.

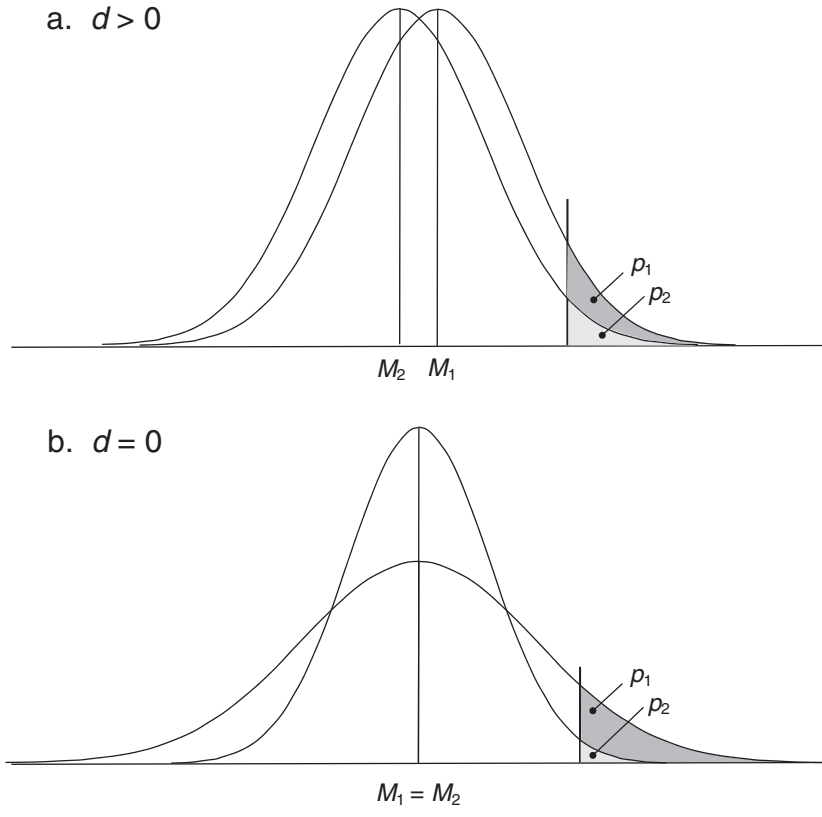

Figure 5.3. The right tail ratio p_1/p_2 relative to cutting point when $d > 0$ for $M_1 > M_2$ (a) and $d = 0$ for $M_1 = M_2$ (b).

Maccoby and Jacklin (1974) reported the following descriptive statistics for large samples of women and men on a verbal ability test for which $\mu = 100.00$, $\sigma = 15.00$:

$$M_W = 103.00,\ s_W = 14.80 \quad \text{and} \quad M_M = 100.00,\ s_M = 14.21$$

Suppose that job applicants will be considered only if their scores exceed 130, or two standard deviations above the mean. Normal deviate equivalents of this threshold in the separate distributions for women and men are, respectively,

$$z_W = \frac{130 - 103.00}{14.80} = 1.82 \quad \text{and} \quad z_M = \frac{130 - 100.00}{14.21} = 2.11$$

Assuming normality, $z_W = 1.82$ falls at the 96.56th percentile in the distribution for women, so .0344 of their scores exceed 130. For men, $z_M = 2.11$

falls at the 98.26th percentile, which implies that .0174 of their scores exceed the cutting point. (You can use a normal curve table or a web calculator for a normal curve to generate these proportions.) Given these results,

$$\text{RTR} = \frac{.0344}{.0174} = 1.98$$

Thus, women are about twice as likely as men to have scores that exceed the cutting point. This method may not give accurate results if the distributions are not approximately normal. One should instead analyze the frequency distributions to find the exact proportions of scores beyond the cutting point. Tail ratios are often reported when gender differences at the extremes of distributions are studied.

Tail ratios generally increase as the threshold moves further to the right (RTR) or further to the left (LTR) when $M_1 \neq M_2$, assuming symmetrical distributions with equal variances. But it can happen that tail ratios are not zero even though $M_1 = M_2$ and $d = r_{pb} = 0$ when there is heteroscedasticity. For example, the two distributions in Figure 5.3(b) have the same means, but the tail ratios are not also generally 1.00 because group 1 is more variable than group 2. Thus, scores from group 1 are overrepresented at both extremes of the distributions. If the researcher wants only to compare central tendencies, this "disagreement" between the tail ratios and *d* may not matter. In a selection context, though, the tail ratios would be of critical interest.

Other Case-Level Proportions

McGraw and Wong's (1992) **common language effect size** (CL) is the predicted probability that a random score on a continuous outcome selected from the group with the higher mean exceeds a random score from the group with the lower mean. If two frequency distributions are identical, CL = .50, which says that it is just as likely that a random score from one group exceeds a random score from the other group. As the two frequency distributions become more distinct, the value of CL increases up to its theoretical maximum of 1.00. Vargha and Delaney (2000) described the **probability of (stochastic) superiority**, which can be applied to ordinal outcome variables. Huberty and Lowman's (2000) **improvement over chance classification**, or *I*, is for the classification phase of logistic regression or discriminant function analysis. The *I* statistic measures the proportionate reduction in the error rate compared with random classification. If *I* = .35, for example, the observed classification error rate is 35% less than that expected in random classification.

TABLE 5.4
Relation of Selected Values of the Standardized Mean Difference to the Point-Biserial Correlation and Case-Level Proportions

Group or variable level		Case level			
d	r_{pb}	U_1	U_3	RTR + 1	RTR + 2
0	0	0	.500	1.00	1.00
.10	.05	.007	.540	1.16	1.27
.20	.10	.148	.579	1.36	1.61
.30	.15	.213	.618	1.58	2.05
.40	.20	.274	.655	1.85	2.62
.50	.24	.330	.691	2.17	3.37
.60	.29	.382	.726	2.55	4.36
.70	.33	.430	.758	3.01	5.68
.80	.37	.474	.788	3.57	7.46
.90	.41	.515	.816	4.25	9.90
1.00	.45	.554	.841	5.08	13.28
1.25	.53	.638	.894	8.14	29.13
1.50	.60	.707	.933	13.56	69.42
1.75	.66	.764	.960	23.60	—[a]
2.00	.71	.811	.977	43.04	—[a]
2.50	.78	.882	.994	—[a]	—[a]
3.00	.83	.928	.999	—[a]	—[a]

Note. RTR + 1 = right tail ratio for a cutting point one standard deviation above the mean of the combined distribution; RTR + 2 = right tail ratio for a cutting point two standard deviations above the mean of the combined distribution.
[a]>99.99.

Relation of Group- or Variable-Level Effect Size to Case-Level Proportions

Assuming normality, homoscedasticity, and large and equal group sizes, case-level proportions are functions of effect size at the group level. Listed in the first column in Table 5.4 are values of d in the range 0–3.0. (Here, d refers to d_{pool}, d_{s1}, or d_{s2}, which are asymptotically equal under these assumptions.) Listed in the remaining columns are values of r_{pb} and case-level proportions that correspond to d in each row. Reading each row in Table 5.4 gives a case-level perspective on mean differences of varying magnitudes. If $d = .50$, for example, the expected value of r_{pb} is .24. For the same effect size, we expect the following at the case level:

1. About one third of all scores are distinct across the two distributions ($U_1 = .33$). That is, about two thirds of the scores fall within the area of overlap.
2. The typical score in the group with the higher mean exceeds about 70% of the scores in the group with the lower mean ($U_3 = .69$).

3. The upper group will have about twice as many scores as the lower group that exceed a threshold one standard deviation above the mean of the combined distribution (RTR = 2.17). For a cutting point two standard deviations above the mean of the combined distribution, the upper group will have more than three times as many scores as the lower group that exceed the threshold (RTR = 3.37).

The relations summarized in Table 5.4 hold only under the assumptions of normality, homoscedasticity, and balanced designs with large samples. Otherwise, it can happen that the group statistics d or r_{pb} tell a different story than case-level effect sizes. Look back at Figure 5.3(b), for which $d = r_{pb} = 0$ but tail ratios are not generally 1.0 due to heteroscedasticity. In actual data sets, researchers should routinely evaluate effects at both the group and case levels, especially in treatment outcome studies; see McGrath and Meyer (2006) for more examples.

SUBSTANTIVE SIGNIFICANCE

This section provides interpretive guidelines for effect sizes. I also suggest how to avoid fooling yourself when estimating effect sizes.

Questions

As researchers learn about effect size, they often ask, what is a large effect? a small effect? a *substantive* (important) effect? Cohen (1962) devised what were probably the earliest guidelines for describing qualitative effect size magnitudes that seemed to address the first two questions. The descriptor *medium* corresponded to a subjective average effect size in nonexperimental studies. The other two descriptors were intended for situations where neither theory nor prior empirical findings distinguish between *small* and *large* effects. In particular, he suggested that $d = .50$ indicated a medium effect size, $d = .25$ corresponded to a small effect size, and $d = 1.00$ signified a large effect size. Cohen (1969) later revised his guidelines to $d = .20$, 50, and .80 as, respectively, small, medium, and large, and Cohen (1988) described similar benchmarks for correlation effect sizes.

Cohen never intended the descriptors small, medium, and large—T-shirt effect sizes—to be applied rigidly across different research areas. He also acknowledged that his conventions were an educated guess. This is why he encouraged researchers to look first to the empirical literature in their areas before using these descriptors. Unfortunately, too many researchers blindly

apply T-shirt effect sizes, claiming, for example, that $d = .49$ indicates a small effect but $d = .51$ is a medium-sized effect (first mistake), because Cohen said so (second mistake). It seems that the bad habit of dichotomous thinking in significance testing is contagious.

The best way for researchers to judge the relative magnitudes of their effects is to consult relevant meta-analytic studies. There are computer tools with the capabilities to calculate, record, and organize effect sizes for sets of related studies. An example is Comprehensive Meta-Analysis (CMA; Borenstein, Hedges, Higgins, & Rothstein, 2005), which accepts different forms of input data and computes various effect sizes.[8] Although intended for researchers who conduct meta-analyses, CMA and similar programs can be used by primary researchers to collect and analyze effect sizes from individual studies. The real benefit from reporting standardized effect sizes comes not from comparing them against arbitrary guidelines but instead from comparing effect sizes directly with those reported in previous studies.

What Is a Substantive Effect?

The third question about effect size—what is a substantive result?—is the toughest. This is because demonstration of an effect's significance—whether theoretical, practical, or clinical—calls for more discipline-specific expertise than the estimation of its magnitude (Kirk, 1996). For example, the magnitude of the gender difference in height among adults is about $d = 2.00$. Whether this difference is substantive depends on the context. In terms of general life adjustment, this gender difference is probably irrelevant. But in the context of automobile safety, the gender difference in height may be critical. Remember a problem with the front air bags in automobiles manufactured before the late 1990s: Their deployment force could injure or kill a small-stature driver or passenger, which presented a greater risk to women. Nowadays, most cars have front seat air bags that vary deployment force according to driver weight. But even smart air bags can injure or kill children, so the even greater height difference between adults and children still has substantive significance in this domain.

By the same logic, results gauged to be small in a T-shirt metric are not necessarily unimportant. Bellinger (2007), Fern and Monroe (1996), Prentice and Miller (1992), and B. Thompson (2006b) described the contexts for when small effects may be noteworthy summarized next:

1. Minimal manipulation of the independent variable results in some change in the outcome variable; that is, a small cause size nevertheless produces an observable effect size.

[8]http://www.meta-analysis.com/index.html

2. An effect operates in a domain where theoretically no effect is expected. An example is the finding that the physical attractiveness of defendants in courtroom trials predicts to some degree sentence severity.
3. The outcome variable is very important, such as human life.
4. An effect is observed on an outcome variable that is difficult to influence. Finding that a treatment alters the course of a severe, degenerative disease is an example.
5. Small shifts in mean values of health-related variables can sometimes lead to big effects when spread over the whole population. The aforementioned finding that taking small doses of aspirin can reduce the risk for heart attack is an example.
6. It can also happen that effects in early stage research are larger than those in later research. This can occur as researchers shift their attention from determining whether an effect exists to studying its more subtle mechanisms at boundary conditions.

In general, effect sizes that are "unimportant" may be ones that fall within the margins of measurement error. But even this general definition does not always hold. The difference in vote totals for the two major candidates in the 2000 presidential election in the United States was within the margin of error for vote counting in certain precincts, but these small differences determined the outcome. Effect sizes that are "important" should also have theoretical, practical, or clinical implications, given the research context. Just as there is no absolute standard for discriminating between small and large effects, however, there is no absolute standard for determining effect size importance. Part of the challenge in a particular research area is to develop benchmarks for substantive significance. Awareness of effect sizes in one's own area helps, as does an appreciation of the need to examine effects at both group and case levels.

If practitioners such as therapists, teachers, or managers are the intended audience, researchers should describe substantive significance in terms relevant to these groups. As part of their customer-centric science model, Aguinis et al. (2010) described the use of **ethnographic techniques** to elucidate frames of reference among practitioners. The goal is to discover relevance from the perspective of stakeholders, who may use different meanings or language than researchers. This process may involve study of phenomena in their natural settings or use of semistructured interviews, diaries, or other qualitative methods. A **conversation analysis** involves the identification of key words or phrases that signal affirmation of relevance when practitioners discuss a problem. A **narrative analysis** in which practitioners describe personal experiences using notes, photographs, or case studies has a similar aim.

Other ethnographic methods include focus groups, historical research, and field notes analysis. The point is to facilitate communication between producers and consumers of research. Similar techniques are used in computer science to discover requirements of those who will use a computer tool and to design program interfaces for specific type of users (e.g., Sutcliffe, 2002).

Clinical Significance

Comparing treated with untreated patients is the basis for discerning clinical significance. One should observe at the case level that a typical treated case is distinguishable from a typical untreated case. The case-level effect sizes U_1 and U_3 speak directly to this issue (see Figure 5.1). For example, the treatment effect size should generally exceed one full standard deviation ($d > 1.0$) in order for most treated and untreated patients to be distinct ($U_1 > .50$; see Table 5.4). At the group level, **criterion contrasts** involve comparisons between groups that represent a familiar and recognizable difference on a relevant outcome. Some possibilities include contrasts between patients with debilitating symptoms and those with less severe symptoms or between patients who require inpatient versus outpatient treatment.

If the metric of the outcome variable is meaningful, it is easier to interpret unstandardized criterion contrasts in terms of clinical significance, but judgment is still required. An example by Blanton and Jaccard (2006) illustrates this point: Suppose a new treatment reduces the mean number of migraines from 10 to 4 per month. A change of this magnitude is likely to be seen by both patients and practitioners as clinically important. But what if the mean reduction were smaller, say, from 10 migraines per month to 9? Here, the researcher must explain how this result makes a difference in patients' lives, such as their overall quality of life, balanced against risk such as side effects when evaluating clinical significance. Again, this is a matter of judgment based on the researcher's domain knowledge, not solely on statistics.

When metrics of outcome variables are arbitrary, it is common for researchers to report **standardized criterion contrast effect sizes** as d statistics. Because metrics of standardized effect sizes are just as arbitrary as the original units, they do not directly convert to implications for clinical significance. This includes effects deemed large in some T-shirt metrics (Blanton & Jaccard, 2006). The realization that neither statistical significance nor standardized effect sizes are sufficient to establish clinical significance means that some so-called empirically validated therapies, such as cognitive behavioral therapy for depression, are not really empirically validated in terms of clinical significance (Kazdin, 2006).

For measures with arbitrary metrics, researchers should also try to identify real-world referents of changes in scores of different magnitudes. A

critical issue is whether effects observed in university laboratories generalize to real-world settings. Another is whether participants in such studies are representative of patients in the general population. For example, patients with multiple diagnoses are often excluded in treatment outcome studies, but many patients in real clinical settings are assigned more than one diagnosis. Some psychological tests have been used so extensively as outcome measures that guidelines about clinical significance are available. For example, Seggar, Lambert, and Hansen (2002) used multiple methods, including analysis of distribution overlap, to recommend thresholds for clinical significance on the Beck Depression Inventory (Beck, Rush, Shaw, & Emory, 1979). Fournier et al. (2010) applied a similar definition in a meta-analysis of the effects of anti-depressant medication. They found that such effects were not clinically significant for patients with mild or moderate levels of depression, but larger and clinically meaningful changes were observed among patients who were severely depressed.

HOW TO FOOL YOURSELF WITH EFFECT SIZE ESTIMATION

Some ways to mislead yourself with effect size estimation mentioned earlier are summarized next. There are probably other paths to folly, but I hope that the major ones are included below:

1. Measure effect size magnitude only at the group level (ignore the case level).
2. Apply T-shirt definitions of effect size without first looking to the empirical literature in one's area.
3. Believe that an effect size judged as large according to T-shirt conventions must be an important result and that a small effect is unimportant.
4. Ignore the question of how to establish substantive significance in one's research area.
5. Estimate effect size only for results that are statistically significant.
6. Believe that effect size estimation somehow lessens the need for replication.
7. Report values of effect sizes only as point estimates; that is, forget that effect sizes are subject to sampling error, too.
8. Forget that effect size for fixed factors is specific to the particular levels selected for study. Also forget that effect size is in part a function of study design.
9. Forget that standardized effect sizes encapsulate other quantities or characteristics, including the unstandardized effect size,

error variance, sample base rates, and experimental design. These are all crucial aspects in study planning and must not be overlooked.

10. As a journal editor or reviewer, substitute effect size for statistical significance as a criterion for whether a work is published.

RESEARCH EXAMPLE

This example illustrates effect size estimation at both the group and case levels in an actual data set. You can download the raw data file in SPSS format for this example from the web page for this book. At the beginning of courses in introductory statistics, 667 psychology students (M age = 23.3 years, s = 6.63; 77.1% women) were administered the original 15-item test of basic math skills reproduced in Table 5.5. The items are dichotomously scored as either 0 (wrong) or 1 (correct), so total scores range from 0 to 17. These

TABLE 5.5
Items of a Basic Math Skills Test

1. $6\frac{1}{4}$ $1\frac{5}{8}$ $+\ 4\frac{1}{2}$	2. Write as a fraction .0031 =	3. Write as a decimal $52\frac{1}{2}\%$ =	4. $\frac{7}{17} = \frac{6}{x}$ x =
5. Multiply 15% by 175	6. Multiply $\frac{1}{4}\%$ by 60	7. $1\frac{1}{2} \div .25$ =	8. $\sqrt{16} \div \sqrt{4}$ =
9. Find average: 34, 16, 45, 27	10. $\frac{2}{x-4} = \frac{4}{x}$ x =	11. If $a = -1$ and $b = 5$ then ab =	
12. $\underline{x}$ $\underline{y}$ 3 4 2 3 4 5 $\sum xy + \sum y$ =	13. If $a = -1$ and $b = 4$ then $a + b$ =	14. $\frac{7-(6+8)}{2}$ =	
15. Where does the line $3x - 2y = 12$ cross the x-axis? the y-axis? what is the slope?			

TABLE 5.6
Descriptive Statistics and Effect Sizes for the Contrast of Students With Satisfactory Versus Unsatisfactory Outcomes in Introductory Statistics on a Basic Math Skills Test

				Group or variable level				Case level	
Group	*n*	*M*	s^2	d_{pool}	$\hat{d}_{pool}$	r_{pb}	$\hat{r}_{pb}$	U_1	U_3
Satisfactory	511	10.96	9.45	.46[a]	.54	.19[b]	.22	.05	.67
Unsatisfactory	156	9.52	10.45						

Note. For a nil hypothesis, t (665) = 5.08. Corrections for measurement error based on r_{XX} = .72.
[a]Noncentral 95% confidence interval for δ [.28, 65]. [b]Noncentral 95% confidence interval for η^2 [.014, .069].

scores had no bearing on subsequent course grades. The internal consistency reliability (Cronbach's alpha) in this sample is .72.

Reported in the left side of Table 5.6 are descriptive statistics for students with satisfactory outcomes in statistics (a final course letter grade of C– or better) versus those with unsatisfactory outcomes (final letter grade of D+ or lower or withdrew). Effect sizes for the group contrast are listed in the right side of the table. Students with satisfactory outcomes had higher mean math test scores (M_1 = 10.96, 64.5% correct) than those with unsatisfactory outcomes (M_2 = 9.52, 56.0% correct) by .46 standard deviations, noncentral 95% CI for δ [.28, .65]. Adjusted for measurement error, this effect size is .54 standard deviations.

The observed correlation between math test scores and satisfactory versus unsatisfactory outcomes in statistics is .19, so the proportion of explained variance is $.19^2$, or .037, noncentral 95% CI for η^2 [.014, .069] (see Table 5.6). The correlation effect size adjusted for measurement error is .22. Because the range of math test scores is relatively narrow, it is not surprising that only about 5% fall outside the area of overlap of the two distributions. There were no math test scores of 16 or 17 among students with unsatisfactory outcomes, but a total of 34 students with satisfactory outcomes achieved scores this high (U_1 = 34/667 = .05). The median score of students with satisfactory outcomes in statistics (11) exceeded about two thirds of the scores among students with unsatisfactory outcomes (U_3 = .67). Exercise 7 asks you to reproduce some of the results in Table 5.6.

Presented in Table 5.7 are results of an alternative case-level analysis that makes implications of the results more obvious for general audiences. Scores on the math test were partitioned into four categories, 0–39, 40–59, 60–79, and 80–100% correct. This was done to find the level of performance (if any) that distinguished students at risk for having difficulties in statistics. Percentages in rows of Table 5.7 indicate proportions of students with satisfactory versus unsatisfactory outcomes for each of the four levels of math test

TABLE 5.7
Relation Between Outcomes in Introductory Statistics and Level of Performance on a Basic Math Skills Test

		Course outcome[a]	
Math score (%)	*n*	Satisfactory	Unsatisfactory
80–100	133	117 (88.0)	16 (12.0)
60–79	237	181 (76.4)	56 (23.6)
40–50	222	172 (77.5)	50 (22.5)
<40	75	41 (54.7)	34 (45.3)

[a]Frequency (row percentage).

performance. The risk for a negative outcome in statistics increases from about 12% among students who correctly solved at least 80% of the math test items to the point where nearly half (45.3%) of the students who correctly solved fewer than 40% of test items had unsatisfactory outcomes.

CONCLUSION

This chapter introduced basic principles of effect size estimation and two families of group- or variable-level standardized effect sizes for continuous outcomes, standardized mean differences and correlation effect sizes. Denominators of standardized mean differences for contrasts between independent means are standard deviations in the metric of the original scores, but it is critical to report which particular standard deviation was specified as the standardizer. There are robust standardized mean differences that may be less affected by nonnormality, heteroscedasticity, or outliers. Descriptive correlation effect sizes considered are all forms of $\hat{\eta}$, or the square root of the sum of squares for the contrast over the total sum of squares. Case-level analysis of proportions of scores from one group versus another group that fall above or below certain reference points can illuminate practical implications of group-level differences. Estimating effect size is part of determining substantive significance, but the two are not synonymous. The next chapter deals with effect sizes for categorical outcomes.

LEARN MORE

Breaugh (2003) reviews mistakes to avoid, and books about effect size estimation by Ellis (2010) and Grissom and Kim (2011) are good resources for applied researchers. Kelley and Preacher (2012) give an excellent overview of the concept of effect size.

Breaugh, J. A. (2003). Effect size estimation: Factors to consider and mistakes to avoid. *Journal of Management, 29,* 79–97. doi:10.1016/S0149-2063(02)00221-0

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results.* New York, NY: Cambridge University Press.

Grissom, R. J., & Kim, J. J. (2011). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York, NY: Routledge.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17,* 137–152. doi: 10.1037/a0028086

EXERCISES

1. Calculate the d statistics in Table 5.3 for $n = 30$.
2. For the scores in Table 2.2, verify that $d_{pool} = .80$ and $d_{diff} = 1.07$.
3. For the data in Table 2.4, calculate the robust effect sizes d_{Win1} and d_{Win2}.
4. Calculate r_{pb} for the data in Table 5.3 for $n = 30$.
5. Verify that $\hat{\eta}^2 = .177$ and partial $\hat{\eta}^2 = .588$ in a dependent samples analysis of the data in Table 2.2.
6. Given $d_{pool} = .80$, $n = 30$, and $r_{12} = .75$ in a dependent samples analysis, construct the approximate 95% confidence interval for δ.
7. Calculate an approximate 95% confidence interval for δ based on the results in Table 5.6.

6

CATEGORICAL OUTCOMES

Change is not merely necessary to life—it is life.
—Alvin Toffler (1970, p. 304)

Some outcomes are categorical instead of continuous. The levels of a categorical outcome are mutually exclusive, and each case is classified into just one level. Widely used effect sizes for categorical outcomes in areas such as medicine, epidemiology, and education are introduced in this chapter. Some of the effect sizes described next can also be estimated in logistic regression or log-linear analysis. Doing so bases effect sizes for categorical outcomes on an underlying statistical model and also corrects for intercorrelations among multiple predictors. In contrast, the same effect sizes computed with the methods described next should be considered descriptive statistics. Exercises for this chapter provide additional practice in estimating effect size for categorical outcomes.

DOI: 10.1037/14136-006
Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, Second Edition, by R. B. Kline

TYPES OF CATEGORICAL OUTCOMES

The simplest categorical outcomes are binary (dichotomous) variables with only two levels, such as relapsed or not relapsed. When two groups are compared on a dichotomy, the data are frequencies that are represented in a 2 × 2 contingency table, also called a **fourfold table**. Categorical variables can also have more than 2 levels, such as *agree*, *disagree*, and *uncertain*. The size of the contingency table is larger than 2 × 2 if two groups are contrasted across > 2 categories. Only some effect sizes for 2 × 2 tables can be extended to larger tables. The same statistics can also be used when > 2 groups are compared on a categorical outcome.

Levels of categorical variables are either unordered or ordered. **Unordered categories**, such as those for ethnicity, marital status, or occupational type, imply no rank order. The technique of binary logistic regression is for dichotomous outcomes, and its extension for outcomes with ≥ 3 unordered categories is multinomial logistic regression. **Ordered categories (multilevel ordinal categories)** have ≥ 3 levels that imply a rank order. An example is the Likert scale *strongly agree*, *agree*, *disagree*, or *strongly disagree*. There are specialized techniques for ordered categories such as **ordinal logistic regression**, which analyzes the relative frequencies of each level of an ordinal criterion and all outcomes that are ordered before it. Methods for ordered categories are not as familiar as those for unordered categories, so the former are not considered further. One alternative is to collapse multilevel categories into two substantively meaningful, mutually exclusive outcomes. Estimation of effect size magnitude is then conducted with methods for fourfold tables.

Another framework that analyzes data from fourfold tables is the **sensitivity, specificity, and predictive value model**. Better known in medicine as a way to evaluate the accuracy of screening tests, this approach can be fruitfully applied to psychological tests that screen for problems such as depression or learning disabilities (e.g., Kennedy, Willis, & Faust, 1997). Because screening tests are not as accurate as more individualized and costly diagnostic methods, not all persons with a positive test result will actually have the target condition. Likewise, not everyone with a negative result is actually free of that condition. The 2 × 2 table analyzed is the cross-tabulation of screening test results (positive–negative) and true status (disorder–no disorder).

The sensitivity, specificity, and predictive value model is a special case of the **receiver operating characteristic (ROC) model** based on **signal detection theory**. The latter was developed in the 1950s by researchers who studied the ability of radar operators to correctly determine whether a screen blip was or was not a threat as a function of thresholds for classifying radar readings as indicating signal versus noise. The ratio of the rate of true positives over false negatives is plotted in ROC curves, which can be studied over different cutting

points. Analysis of ROC curves has been applied in the behavioral sciences to the study of sensory thresholds in psychophysics and decision making under conditions of uncertainty (Swets, 1996). Screening test accuracy can also be analyzed in Bayesian estimation controlling for base rates when estimating the probability of a disorder given positive versus negative test results.

EFFECT SIZES FOR 2 × 2 TABLES

Effect sizes, considered next, estimate the degree of relative risk for an undesirable outcome, such as relapsed–not relapsed, across different populations, such as treatment versus control. The same estimators and their corresponding parameters can also be defined when neither level of the outcome dichotomy corresponds to something undesirable, such as agree–disagree. In this case, the idea of risk is replaced by that of comparing relative proportions for binary outcomes.

Presented in Table 6.1 is a fourfold table for comparing treatment and control groups on the outcome relapsed–not relapsed. The letters in the table represent observed frequencies in each cell. For example, the size of the control group is $n_C = A + B$, where A and B, respectively, stand for the number of untreated cases that relapsed or did not relapse. The size of the treatment group is $n_T = C + D$, where C and D, respectively, symbolize the number of treated cases that relapsed or did not relapse. The total sample size is the sum of A, B, C, and D. Listed in Table 6.2 are the effect sizes, the equation for each effect size based on the cell frequencies represented in Table 6.1, and the corresponding parameter.

Risk Rates

With reference to Table 6.1, the proportion of cases in the control group (C) and the treatment group (T) that relapsed are respectively defined as $p_C = A/(A + B)$ and $p_T = C/(C + D)$. The complements of these ratios,

TABLE 6.1
A Fourfold Table for a Contrast on a Dichotomy

	Relapsed	Not relapsed
Control	*A*	*B*
Treatment	*C*	*D*

Note. The letters *A*–*D* represent observed cell frequencies.

TABLE 6.2
Risk Effect Sizes for Fourfold Tables

Statistic	Equation	Parameter
Risk rates		
p_C	$\frac{A}{A+B}$	π_C
p_T	$\frac{C}{C+D}$	π_T
Comparative risk		
RD	$p_C - p_T$	$\pi_C - \pi_T$
RR	$\frac{p_C}{p_T}$	$\frac{\pi_C}{\pi_T}$
OR	$\frac{odds_C}{odds_T} = \frac{p_C/(1-p_C)}{p_T/(1-p_T)}$	$\omega = \frac{\Omega_C}{\Omega_T}$
Correlation		
$\hat{\varphi}$	$\left\lvert \frac{AD-BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} \right\rvert = \sqrt{\frac{\chi^2_{2\times2}}{N}}$	φ

Note. The letters *A–D* represent observed cell frequencies in Table 6.1. If *A, B, C,* or *D* = 0 in computation of OR, add .5 to the observed frequencies in all cells. RD = risk difference; RR = risk ratio; OR = odds ratio; $\chi^2_{2\times2}$ = contingency table chi-square with a single degree of freedom.

or $1 - p_C$ and $1 - p_T$, are the proportions of cases in each group that did not relapse. The statistic p_C estimates π_C, the proportion of cases in the control population that relapsed, and p_T estimates the corresponding parameter π_T in the treatment population.

Comparative Risk

The **risk difference** (RD) is defined as $p_C - p_T$, and it estimates the parameter $\pi_C - \pi_T$. The result $p_C - p_T = .10$ indicates a relapse rate 10% higher in the control sample than in the treatment sample. Likewise, $p_C - p_T = -.20$ says that the relapse rate among treated cases is 20% higher than that among control cases. The **risk ratio** (RR) is the ratio of the risk rates. It is defined here as p_C/p_T, but which rate appears in the numerator versus the denominator is arbitrary, so one should always explain how RR is computed. If $p_C/p_T = 1.30$, for example, the relapse risk is 1.3 times higher among control than treated cases. Similarly, if RR = .80, the relapse risk in the control group is 80% as high as that in the treatment group. The statistic RR estimates π_C/π_T.

The **odds ratio** (OR) is the ratio of the within-groups odds for the undesirable event. It is defined in Table 6.1 as the ratio of the odds for relapse in the control group, $odds_C$, over the odds for relapse in the treatment group, $odds_T$. These odds are defined as

$$odds_C = \frac{p_C}{1 - p_C} \quad \text{and} \quad odds_T = \frac{p_T}{1 - p_T} \tag{6.1}$$

Suppose $p_C = .60$ and $p_T = .40$ are, respectively, the relapse rates among control and treated cases. The relapse odds in the control group are .60/.40 = 1.50, so the odds of relapse are 3:2. In the treatment group, the odds for relapse are lower, .40/.60 = .67; that is, the odds of relapse are 2:3. The odds ratio is OR = 1.50/.67 = 2.25, which says that the relapse odds are 2¼ times higher among control cases than treated cases. Likewise, OR = .75 would say that the relapse odds in the control group are only 75% as high as the odds in the treatment group. In fourfold tables where all margin totals are equal, $OR = RR^2$. The parameter for OR is $\omega = \Omega_C/\Omega_T$, the ratio of the within-populations odds where

$$\Omega_C = \frac{\pi_C}{1 - \pi_C} \quad \text{and} \quad \Omega_T = \frac{\pi_T}{1 - \pi_T} \tag{6.2}$$

A convenient property of OR is that it can be converted to a kind of standardized mean difference known as **logit *d*** (Chinn, 2000). Here, a logit is ln (OR), the natural log of OR. This logistic distribution is approximately normal with a standard deviation that equals $pi/3^{1/2}$, or about 1.8138. The ratio of ln (OR) over $pi/3^{1/2}$ is a logit *d* that is comparable to a standardized mean difference for the same contrast but on a continuous outcome. The logit *d* can also be expressed in basically the same form as a conventional standardized mean difference:

$$\text{logit } d = \frac{\ln (\text{OR})}{\text{pi}/\sqrt{3}} = \frac{\ln (odds_C) - \ln (odds_T)}{\text{pi}/\sqrt{3}} \tag{6.3}$$

Reporting logit *d* may be of interest when the hypothetical variable that underlies the observed dichotomy is continuous. For example, there may be degrees of recovery that underlie the binary classification of recovered–not covered. Suppose that $p_C = .60$ and $p_T = .40$, which implies $odds_C = 1.50$, $odds_T = .67$, and OR = 2.25. The value of logit *d* is

$$\text{logit } d = \frac{\ln (2.25)}{\text{pi}/\sqrt{3}} = \frac{\ln (1.50) - \ln (.67)}{\text{pi}/\sqrt{3}} = .45$$

Thus, the finding that the odds for relapse are 2¼ times higher among control cases corresponds to a treatment effect size magnitude of about .45 standard deviations in logistic units. Hunter and Schmidt (2004) described other ways to adjust for dichotomization of continuous outcomes.

Correlation

The Pearson correlation between two dichotomous variables is $\hat{\varphi}$ (Greek lowercase phi). It can be calculated with the standard equation for the Pearson *r* if the levels of both variables are coded as 0 or 1. It may be more convenient to calculate $\hat{\varphi}$ directly from the cell and margin frequencies using the equation in Table 6.2. The theoretical range of $\hat{\varphi}$ derived this way is −1.0 to 1.0, but the sign of $\hat{\varphi}$ is arbitrary because it depends on the particular arrangement of the cells. But remember that effects in 2 × 2 tables are directional. For example, either treated or untreated cases will have a higher relapse rate (if there is a difference). The absolute value of $\hat{\varphi}$ also equals the square root of the ratio of χ^2 (1) statistic for the fourfold table over the sample size (see Table 6.2). In a squared metric, $\hat{\varphi}^2$ estimates the proportion of explained variance. In fourfold tables where the row and column marginal totals are all equal, $|RD| = |\hat{\varphi}|$.

The parameter estimated by $\hat{\varphi}$ is

$$\varphi = \frac{\pi_{CR}\,\pi_{TNR} - \pi_{CNR}\,\pi_{TR}}{\sqrt{\pi_{C\bullet}\,\pi_{T\bullet}\,\pi_{\bullet R}\,\pi_{\bullet NR}}} \tag{6.4}$$

where subscripts C, T, R, and NR mean, respectively, control, treatment, relapsed, and not relapsed. The proportions in the numerator represent the four possible outcomes and sum to 1.0. For example, π_{CR} is the probability of being in the control population and relapsing. The subscript $\bullet$ indicates a marginal proportion. For example, $\pi_{C\bullet}$ and $\pi_{T\bullet}$ are, respectively, the relative proportions of cases in the control and treatment populations, and they sum to 1.0.

Evaluation

The risk difference RD is easy to interpret but has a drawback: Its range depends on the values of the population proportions π_C and π_T. That is, the range of RD is greater when both π_C and π_T are closer to .50 than when they are closer to either 0 or 1.00. The implication is that RD values may not be comparable across different studies when the corresponding parameters π_C and π_T are quite different. The risk ratio RR is also easy to interpret. It has the shortcoming that only the finite interval from 0 to < 1.0 indicates lower

risk in the group represented in the numerator, but the interval from > 1.00 to infinity is theoretically available for describing higher risk in the same group. The range of RR varies according to its denominator. For example, the range of p_C/p_T is 0–2.50 for $p_T = .40$, but for $p_T = .60$ its range is 0–1.67. This property limits the value of RR for comparing results across different studies. This problem is dealt with by analyzing natural log transformations of RR, a point elaborated momentarily.

The odds ratio OR shares the limitation that the finite interval from 0 to < 1.0 indicates lower risk in the group represented in the numerator, but the interval from > 1.0 to infinity describes higher risk for the same group. Analyzing natural log transformations of OR and then taking antilogs of the results deals with this problem, just as for RR. The odds ratio may be the least intuitive of the comparative risk effect sizes, but it probably has the best overall statistical properties. This is because OR can be estimated in prospective studies, in studies that randomly sample from exposed and unexposed populations, and in retrospective studies where groups are first formed based on the presence or absence of a disease before their exposure to a putative risk factor is determined (Fleiss & Berlin, 2009). Other effect sizes may not be valid in retrospective studies (RR) or in studies without random sampling ($\hat{\varphi}$).

Do not lose sight of absolute risk rates when reporting risk or odds ratios for rare events. Suppose that the rate of a serious side effect among treated patients is 1/1,000. The base rate of the same complication in the general public is 1/10,000. These results imply

$$\text{RR} = \frac{.001}{.0001} = 10.00 \quad \text{and} \quad \text{OR} = \frac{.001/(1-.001)}{.0001/(1-.0001)} = 10.01$$

but these tenfold increases in relative risk or odds among treated cases refer to a rare outcome. Ten times the likelihood of rare event still makes for a low base rate. Only the risk difference makes it clear that the absolute increase in risk is slight, RD = .0009, or .09%. King and Zeng (2001) discussed challenges in estimating rare events in logistic regression.

The correlation $\hat{\varphi}$ can reach its maximum absolute value (1.0) only if the marginal proportions for rows and columns in a fourfold table are equal. As the row and column marginal proportions diverge, the maximum absolute value of $\hat{\varphi}$ approaches zero. This implies that the value of $\hat{\varphi}$ will change if the cell frequencies in any row or column are multiplied by an arbitrary constant. This makes $\hat{\varphi}$ a **margin-bound effect** size; the correlation r_{pb} is also margin bound because it is affected by group base rates (see Equation 5.12). Exercise 1 asks you to demonstrate this property of $\hat{\varphi}$. Grissom and Kim (2011, Chapters 8–9) described additional effect sizes for categorical outcomes.

Interval Estimation

Sample proportions follow binomial distributions. The **Wald method** for constructing approximate 100 $(1 - \alpha)$% confidence intervals for π depends on normal approximations. Widths of confidence intervals centered on p_C or p_T in this method are calculated as products of the standard errors listed in Table 6.3 and $z_{2\text{-tail},\,\alpha}$. A potential problem is **overshoot**, which happens when the lower bound of an interval based on a very low sample proportion, such as .02, is less than zero. It can also happen that the upper bound based on a very high proportion, such as .97, exceeds 1.0. Overshoot is dealt with by truncating the interval to lie within the range 0–1.0. Another problem is **degeneracy**, which refers to confidence intervals with zero widths when the sample proportion is either 0 or 1.0. A continuity correction avoids degenerate intervals by adding the constant $1/2n$ where n is the group size, but this correction can cause overshoot. Newcombe (1998) described additional approximate methods.

Approximate standard errors for RD, RR, and OR are also listed in Table 6.3. Distributions of RR and OR are not generally normal, but natural log transformations of these statistics are approximately normal. Consequently, the lower and upper bounds of confidence intervals in natural log units are converted back to their respective original units by taking their antilogs. The

TABLE 6.3
Asymptotic Standard Errors of Risk Effect Sizes

Statistic	Standard error
p_C	$\sqrt{\frac{p_C(1-p_C)}{n_C}}$
p_T	$\sqrt{\frac{p_T(1-p_T)}{n_T}}$
RD	$\sqrt{\frac{p_C(1-p_C)}{n_C}+\frac{p_T(1-p_T)}{n_T}}$
ln (RR)	$\sqrt{\frac{1-p_C}{n_C p_C}+\frac{1-p_T}{n_T p_T}}$
ln (OR)	$\sqrt{\frac{1}{n_C p_C(1-p_C)}+\frac{1}{n_T p_T(1-p_T)}}$

Note. Four-decimal accuracy is recommended in computations. RD = risk difference; RR = risk ratio; ln = natural log; OR = odds ratio.

equation for the asymptotic standard error of $\hat{\varphi}$ is complicated and is not presented here (see Fleiss & Berlin, 2009, p. 242).

The results $p_C = .60$ and $p_T = .40$ imply RD = .20, RR = 1.50, and OR = 2.25. Assume group sizes of $n_C = n_T = 100$. Next, we calculate the approximate 95% confidence interval for the population risk difference $\pi_C - \pi_T$. When the third equation in Table 6.3 is used, the estimated standard error is

$$s_{RD} = \sqrt{\frac{.40\,(1-.40)}{100} + \frac{.60\,(1-.60)}{100}} = .0693$$

The value of $z_{2\text{-tail},\,.05}$ is 1.96, so the approximate 95% confidence interval for $\pi_C - \pi_T$ is

$$.20 \pm .0693\ (1.96)$$

which defines the interval [.06, .34]. Thus, the sample result RD = .20 is just as consistent with a population risk difference as low as .06 as it is with a population risk difference as high as .34, with 95% confidence.

This time, I construct the approximate 95% confidence interval for the population odds ratio ω based on OR = 2.25:

$$\ln(2.25) = .8109$$

$$s_{\ln(OR)} = \sqrt{\frac{1}{100\,(.40)\,(1-.40)} + \frac{1}{100\,(.60)\,(1-.60)}} = .2887$$

The approximate 95% confidence interval for ln (ω) is

$$.8109 \pm .2887\ (1.96)$$

which defines the interval [.2450, 1.3768]. To convert the lower and upper bounds of this interval back to OR units, I take their antilogs:

$$\ln^{-1}(.2450) = e^{.2450} = 1.2776 \quad \text{and} \quad \ln^{-1}(1.3768) = e^{1.3768} = 3.9622$$

The approximate 95% confidence interval for ω is [1.28, 3.96] at two-decimal accuracy. I can say that OR = 2.25 is just as consistent with a population odds ratio as low as $\omega = 1.28$ as it is with a population odds ratio as high as $\omega = 3.96$, with 95% confidence. Exercise 2 asks you to calculate the approximate 95% confidence interval for π_C / π_T given RR = 1.50 in this example. There are some calculating web pages that derive approximate confidence

intervals for ω.[1] Herbert's (2011) Confidence Interval Calculator is a Microsoft Excel spreadsheet for means, proportions, and odds.[2] The R package EpiTools by T. Aragón supports point and interval estimation with risk effect sizes in applied epidemiological studies.[3]

EFFECT SIZES FOR LARGER TABLES

If the categorical outcome has more than two levels or there are more than two groups, the contingency table is larger than 2 × 2. Measures of comparative risk (RD, RR, OR) can be computed for such a table only if it is reduced to a 2 × 2 table by collapsing or excluding rows or columns. What is probably the best known measure of association for contingency tables with more than two rows or columns is Cramer's *V,* an extension of the $\hat{\varphi}$ coefficient. Its equation is

$$V = \sqrt{\frac{\chi^2_{r \times c}}{\min\,(r-1, c-1) \times N}} \tag{6.5}$$

where the numerator under the radical is the contingency table chi-square with degrees of freedom equal to the number of rows (*r*) minus one times the number of columns (*c*) minus one (see Equation 3.15). The denominator under the radical is the product of the sample size and the smallest dimension of the table minus one. For example, if the table is 3 × 4 in size, then

$$\min\,(3-1, 4-1) = 2$$

For a 2 × 2 table, the equation for Cramer's *V* reduces to that for $|\hat{\varphi}|$. For larger tables, Cramer's *V* is not a correlation, although its range is 0 to 1.00. Thus, one cannot generally interpret the square of Cramer's *V* as a proportion of explained variance. Exercise 3 asks you to calculate Cramer's *V* for the 4 × 2 cross-tabulation in Table 5.7.

SENSITIVITY, SPECIFICITY, AND PREDICTIVE VALUE

Suppose for a disorder there is a gold standard diagnostic method that is individualized and expensive. A screening test is not as accurate as the gold standard, but it costs less and is practical for group administration. Screening

[1]http://www.hutchon.net/ConfidOR.htm
[2]http://www.pedro.org.au/english/downloads/confidence-interval-calculator/
[3]http://cran.r-project.org/web/packages/epitools/index.html

tests are often continuous measures, such as the blood concentration of a particular substance. They also typically have a threshold that differentiates between a positive result that predicts the presence of the disorder (clinical) and a negative result that predicts its absence (normal).

Distributions of clinical and normal groups on continuous screening tests tend to overlap. This situation is illustrated in Figure 6.1, where the clinical group has a higher mean than the normal group. Also represented in the figure is a threshold that separates positive and negative test results. Some clinical cases have negative results; these are false negative test outcomes. Likewise, some normal cases have positive results, which are false positive outcomes. Such results represent potential diagnostic or prediction errors. Analyzing ROC curves can help to determine optimal thresholds, ones that balance costs of a decision error (false positive or negative) against benefits of correct prediction (Perkins & Schisterman, 2006). The discussion that follows assumes a threshold is already established.

The relation between screening test results (positive–negative) and actual status as determined by a gold standard method (clinical–normal) is represented in the top part of Table 6.4. The letters in the table stand for cell frequencies. For example, *A* represents the number of clinical cases with a positive result, and *D* represents the number of normal cases with negative results. Both cells just described correspond to correct predictions. Cells *B* and C in the table represent, respectively, false positive and false negative results.

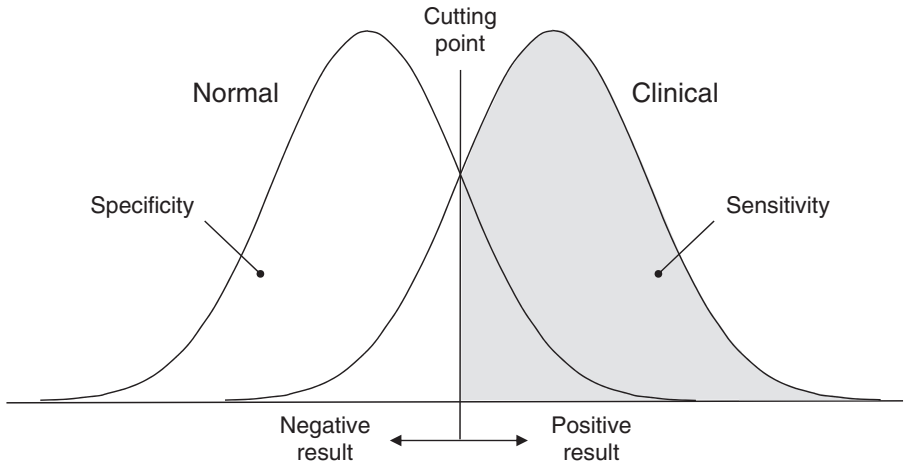

Figure 6.1. Distributions of clinical and normal groups on a continuous screening test with a cutting point. A positive result means predict clinical; a negative result means predict normal.

TABLE 6.4
Definitions of Sensitivity, Specificity, Predictive Value, and Base Rate

		True status	
Screening test result	Prediction	Clinical	Normal
Positive	Clinical	*A*	*B*
Negative	Normal	*C*	*D*
Statistic		Equation	
Sensitivity		$\frac{A}{A+C}$	
Specificity		$\frac{D}{B+D}$	
BR		$\frac{A+C}{A+B+C+D}$	
PPV		$\frac{A}{A+B}$	
NPV		$\frac{D}{C+D}$	

Note. The letters *A*–*D* represent observed cell frequencies. BR = base rate; PPV = positive predictive value; NPV = negative predictive value. The total number of cases is $N = A + B + C + D$.

Sensitivity and Specificity

Sensitivity, specificity, base rate, and predictive value are defined in the bottom part of Table 6.4 based on cell frequencies in the top part of the table. **Sensitivity** is the proportion of screening test results from clinical cases that are correct, or $A/(A + C)$. If sensitivity is .80, then 80% of test results in the clinical group are valid positives and the rest, 20%, are false negatives. **Specificity** is the proportion of results from normal cases that are correct, or $D/(B + D)$. If specificity is .70, then 70% of the results in the normal group are valid negatives and the rest, 30%, are false positives. The ideal screening test is 100% sensitive and 100% specific. Given overlap of distributions such as that illustrated in Figure 6.1, this ideal is not within reach.

Sensitivity and specificity are determined by the threshold on a screening test. This means that different thresholds on the same test will generate different sets of sensitivity and specificity values in the same sample. But both sensitivity and specificity are independent of population base rate and sample size. For example, a test that is 80% sensitive for a disorder should correctly

detect 80 of 100 or 400 of 500 clinical cases. Likewise, a test that is 70% specific should correctly classify 140 of 200 or 700 of 1,000 normal cases.

Predictive Value and Base Rate

Sensitivity and specificity affect **predictive value**, the proportion of test results that are correct, and in this sense predictive value reflects the confidence that diagnosticians can place in test results. **Positive predictive value** (PPV) is the proportion of all positive results that are correct—that is, obtained by clinical cases, or $A/(A + B)$ in Table 6.4. **Negative predictive value** (NPV) is the proportion of negative test results that are correct, that belong to normal cases, or $D/(C + D)$ in the table. In general, predictive values increase as sensitivity and specificity increase.

Predictive value is also influenced by the **base rate** (BR), the proportion of all cases with the disorder, or $(A + C)/N$ in Table 6.4. The effect of BR on predictive value is demonstrated in Table 6.5. Two different fourfold tables are presented there for hypothetical populations of 1,000 cases and a test where sensitivity = .80 and specificity = .70. In the first table, BR = .10 because 100/1,000 cases have the disorder, and 80% of them (80) have a correct (positive) test result. A total of 90% of the cases do not have the disorder (900), and 70% of them (630) have a correct (negative) result. But of all positive results, only 80/350 = 23% are correct, so PPV = .23. But most negative test results are correct, or 630/650 = 97%, so NPV = .97. These predictive values say that the test is quite accurate in ruling out the disorder but not in detecting its presence.

TABLE 6.5
Positive and Negative Predictive Values at Two Different Base Rates for a Screening Test 80% Sensitive and 70% Specific

Screening test result	Prediction	True status			Predictive value	
		Clinical	Normal	Total	Positive	Negative
		Base rate = .10				
Positive	Clinical	80	270	350	.23	.97
Negative	Normal	20	630	650		
Total		100	900	1,000		
		Base rate = .75				
Positive	Clinical	600	75	675	.89	.54
Negative	Normal	150	175	325		
Total		750	250	1,000		

If BR is not .10, both predictive values change. This is shown in the second 2 × 2 cross-tabulation in Table 6.5 for the same test but now for BR = .75. (Base rates of certain parasitic diseases in some parts of the world are this high.) Of all 675 cases with positive test results, a total of 600 belong to clinical cases, so PPV = 600/675, or .89. Likewise, of all 325 negative results, a total of 175 are from normal cases, so NPV = 175/325, or .54. Now more confidence is warranted in positive test results than in negative results, which is just the opposite of the case for BR = .10. Exercise 4 asks you to calculate PPV and NPV for sensitivity = .80, specificity = .70, and BR = .375.

In general, PPV decreases and NPV increases as BR approaches zero. This means that screening tests tend to be more useful for ruling out rare disorders than correctly predicting their presence. It also means that most positive results may be false positives under low base rate conditions. This is why it is difficult for researchers or social policy makers to screen large populations for rare conditions without many false positives. It also explains why the No-Fly List, created by national intelligence services after the 2001 terrorist attacks in New York, seems to mainly prevent innocent passengers from boarding commercial aircraft for travel in or out of the United States. For example, it has happened that prospective passengers have been matched to the list based only on surnames, some of which are common in certain ethnic groups. Once they have been flagged as a risk, innocent passengers find it difficult to clear their names. This problem is expected given the very low base rate of the intention to commit terrorist acts when relying on imperfect screening measures.

Depicted in Figure 6.2 are the expected relations between base rate and predictive values for a screening test where sensitivity = .80 and specificity = .70. The figure makes apparent that PPV heads toward zero as BR is increasingly lower. If PPV < .50, most positive test results are false positives, but BR must exceed roughly .30 before it is even possible for PPV to exceed .50. As BR increases, the value of PPV gets progressively higher, but NPV gradually declines toward zero as BR approaches 1.00. That is, the screening test is not very accurate in ruling out some conditions when most of the population is afflicted.

The effect of BR on predictive values is striking but often overlooked, even by professionals (Grimes & Schulz, 2002). One misunderstanding involves confusing sensitivity and specificity, which are invariant to BR, with PPV and NPV, which are not. This means that diagnosticians fail to adjust their estimates of test accuracy for changes in base rates, which exemplifies the **base rate fallacy**. This type of question presented by Casscells,

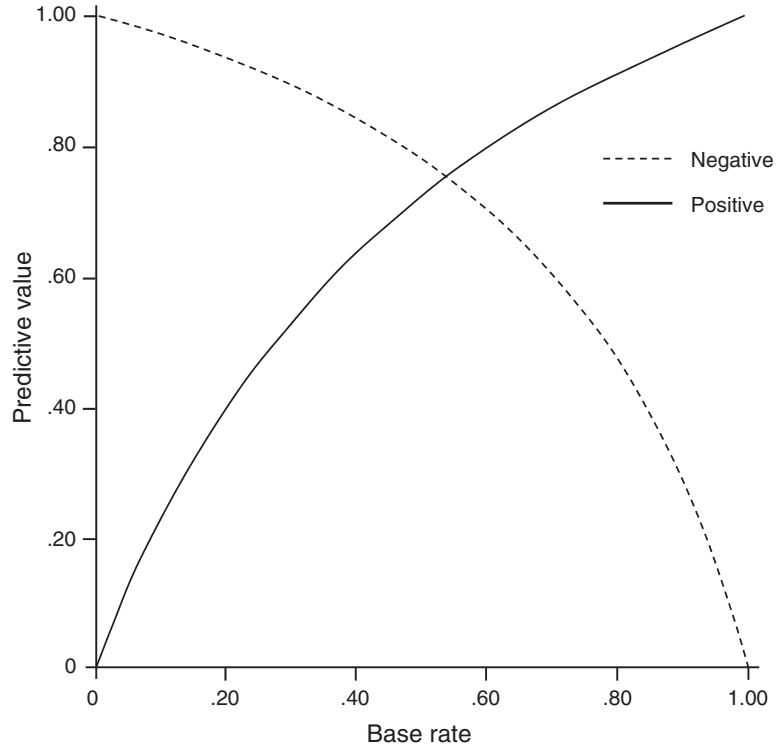

Figure 6.2. Expected predictive values as functions of base rate for a screening test that is 80% sensitive and 70% specific.

Schoenberger, and Graboys (1978) to senior medical students and physicians illustrates this fallacy:

> The prevalence (base rate) of a disease is .10. A test is used to detect the disease. The test is useful but not perfect. Patients with the disease are identified by the test 80% of the time. Patients without the disease are correctly identified by the test 70% of the time. A patient seen recently at a clinic was diagnosed by the test as having the disease. How likely is it that this result is correct? (p. 999)

The correct answer to this question is the PPV of the test, which you now know is .23 (see Table 6.5). But most respondents in Casscells et al.'s (1978) study gave answers close to .80, which is sensitivity. In this case, confusing PPV with sensitivity means that the former is overestimated by a factor of almost four. The base rate fallacy is still a concern in medical education (Maserejian, Lutfey, & McKinlay, 2009).

Likelihood Ratio and Posttest Odds

The likelihood ratio is the probability that a screening test result—positive or negative—would be expected among clinical cases compared with the probability of the same result among normal cases. The **positive likelihood ratio** (PLR) is

$$\text{PLR} = \frac{\text{sensitivity}}{1 - \text{specificity}} \tag{6.6}$$

which indicates the number of times more likely that a positive result comes from clinical cases (numerator) than from normal cases (denominator). The **negative likelihood ratio** (NLR) is

$$\text{NLR} = \frac{1 - \text{sensitivity}}{\text{specificity}} \tag{6.7}$$

and it measures the degree to which a negative result is more likely to come from clinical cases than from normal cases.

Using Bayesian methods, diagnosticians can estimate how much the odds of having a disorder will change given a positive versus negative test result and the disorder base rate expressed as odds. The relation is

$$odds_{\text{post}} = odds_{\text{pre}} \times \text{LR} \tag{6.8}$$

where $odds_{\text{pre}}$ are the odds of having the disorder before administering the screening test calculated as BR/(1 – BR). The term LR in Equation 6.8 refers to the likelihood ratio, either positive or negative, and it is the factor by which the pretest odds will *change* given the corresponding test result. The expression $odds_{\text{post}}$ is the odds of having the disorder, given both the test result and the pretest odds. Equation 6.8 is Bayesian because it updates the old belief (pretest odds) by a factor that estimates the likelihood of either a positive or a negative result.

A screening test is 80% sensitive and 70% specific, and BR = .15. The value of PLR is

$$\text{PLR} = \frac{.80}{1 - .70} = 2.667$$

so positive test results are about 2⅔ times more likely among clinical cases than among normal cases. The value of NLR is

$$\text{NLR} = \frac{.20}{.70} = .286$$

so we can conclude that a negative test result among clinical cases is only about .29 times as likely as among normal cases. The pretest odds are .15/(1 − .15) = .176. The posttest odds of the disorder following a positive result are

$$odds_{\text{post}+} = .176 \times 2.667 = .471$$

which, as expected, are higher than the pretest odds (.176). To convert odds to probability, we calculate $p = odds/(1 + odds)$. So the probability of the disorder increases from BR = .15 before testing to .471/(1 + .471), or about .32, after a positive test result. The posttest odds of the disorder after a negative result are

$$odds_{\text{post}-} = .176 \times .286 = .050$$

which are lower than the pretest odds (.176). Thus, the probability of the disorder decreases from BR = .15 to .050/(1 + .050), or about .048, after observing a negative result.

A negative test result in this example has a greater relative impact than a positive result on the odds of having the disorder. The factor by which the pretest odds are increased, given a positive result, is PLR = 2.667. But the factor by which the pretest odds are reduced, given a negative result, is NLR = .286, which is same as dividing the pretest odds by a factor of 1/.286, or about 3.50. This pattern is consistent with this screening test, where sensitivity = .80, specificity = .70, and BR = .15. That is, the test will be better at ruling out a disorder than at detecting it under this base rate (see Figure 6.2). In general, test results have greater impact on changing the pretest odds when the base rate is moderate, neither extremely low (close to 0) nor extremely high (close to 1.0). But if the target disorder is either very rare or very common, only a result from a highly accurate screening test will change things much. There are web pages that calculate likelihood ratios.[4]

The method just described to estimate posttest odds can be applied when base rate or test characteristics vary over populations. Moons, van Es, Deckers, Habbema, and Grobbee (1997) found that sensitivity, specificity, and likelihood ratios for the exercise (stress) test for coronary disease varied by gender and systolic blood pressure at baseline. In this case, no single set of estimates was adequate for all groups. Exercise 5 concerns the capability to tailor estimates for different groups, in this case the disorder base rate. It is also possible to combine results from multiple screening tests, which may further improve prediction accuracy.

[4]http://www.medcalc.org/calc/diagnostic_test.php

These issues have great relevance from a public health perspective. As I was writing this chapter, the Canadian Task Force on Preventive Health Care (CTFPHC; 2011) released new guidelines for breast cancer screening. The main change is that clinical breast exams are no longer routinely recommended for women at average risk. For women in this group who are 40 to 49 years old, the task force also recommended that mammography should not be routinely conducted. Part of the rationale concerned the relatively high rate of false positives from mammograms for healthy women in this age range. The CTFPHC estimated that for every 2,100 women age 40–49 years routinely screened every 2–3 years, about 690 will have a false positive mammogram leading to unnecessary stress and follow-up testing. About 75 of these women are expected to undergo an unnecessary breast biopsy. These guidelines were controversial at the time of their publication (e.g., The Canadian Press, 2011), in part due to the false belief that the CTFPHC recommended that mammograms should be denied to women who request them. There is a similar controversy about the value of routine prostate-specific antigen (PSA) screening for prostate cancer (Neal, Donovan, Martin, & Hamdy, 2009).

In summary, sensitivity and specificity describe how the presence versus absence of a disorder affects screening test results, and they are not affected by base rates. In contrast, predictive values estimate the probabilities of abnormality versus normality for, respectively, positive versus negative test results, and they are subject to base rates. Likelihood ratios can be used to estimate the odds (or probability) of having the disorder while taking account of the prior odds (base rate) of the disorder, given a positive versus negative result. Base rate estimates can be adjusted for different patient groups or contexts (Deeks & Altman, 2004).

Estimating Base Rates

Estimates of disorder base rates are not always readily available. Without large-scale epidemiological studies, other sources of information, such as case records or tabulations of the frequencies of certain diagnoses, may provide reasonable approximations. The possibility of estimating base rates from such sources prompted Meehl and Rosen (1955) to say that "our ignorance of base rates is nothing more subtle than our failure to compute them" (p. 213). One can also calculate predictive values for a range of base rates. The use of imprecise (but not grossly inaccurate) estimates may not have a large impact on predictive values, especially for tests with high sensitivities and specificities. But lower sensitivity and specificity values magnify errors due to imprecise base rate estimates.

Interval Estimation

Sensitivity, specificity, and predictive values are proportions that are typically calculated in samples, so they are subject to sampling error. Base rates are subject to sampling error, too, if they are empirically estimated. It is possible to construct confidence intervals based on any of these proportions using the Wald method. Just use the equation for either p_C or p_T in Table 6.3 to estimate the standard error for the proportion of interest. Another option is to use one of the other methods described by Newcombe (1998) for sample proportions.

Likelihood ratios are affected by sampling error in estimates of sensitivity and specificity. Simel, Samsa, and Matchar (1991) described a method to construct approximate confidence intervals based on observed likelihood ratios. Their method analyzes natural log transformations of likelihood ratios, and it is implemented in some web calculating pages that also derive confidence intervals based on sample sensitivity, specificity, and predictive values.[5] Herbert's (2011) Confidence Interval Calculator spreadsheet for Excel also uses the Simel et al. (1991) method to calculate confidence intervals based on observed likelihood ratios and values of sensitivity and specificity (see footnote 2).

Posttest odds of the disorder are affected by sampling error in estimates of specificity, sensitivity, likelihood ratios, and base rates. Mossman and Berger (2001) described five different methods of interval estimation for the posttest odds following a positive test result. Two of these methods are calculable by hand and based on natural log transformations, but other methods, such as Bayesian interval estimation, require computer support. Results of computer simulations indicated that results across the five methods were generally comparable for group sizes > 80 and sample proportions not very close to either 0 or 1.00.

Crawford, Garthwaite, and Betkowska (2009) extended the Bayesian method described by Mossman and Berger (2001) for constructing confidence intervals based on posttest probabilities following a positive result. Their method accepts either empirical estimates of base rates or subjective estimates (guesses). It also constructs either two-sided or one-sided confidence intervals. One-sided intervals may be of interest if, for example, the diagnostician is interested in whether the posttest probability following a positive result is lower than the point estimate suggests but not in whether it is higher. A computer program that implements the Crawford et al. (2009) method can be freely downloaded.[6] It requires input of observed frequencies, not proportions, and it does not calculate intervals for posttest probabilities after negative results.

Suppose that results from an epidemiological study indicate that a total of 150 people in a sample of 1,000 have the target disorder (BR = .15). In

[5]http://ktclearinghouse.ca/cebm/practise/ca/calculators/statscalc
[6]http://www.abdn.ac.uk/~psy086/dept/BayesPTP.htm

another study of 150 patients with the disorder and 300 control cases, the sensitivity of a screening test is .80 (i.e., 120 valid positives, 30 false negatives), and the specificity is .70 (i.e., 210 valid negatives, 90 false positives). The positive and negative likelihood ratios of the test are, respectively, PLR = 2.667 and NLR = .286. Given these results, the posttest odds for the disorder after a positive test result are .471, which converts to a posttest probability of .320. I used the Wald method to calculate approximate 95% confidence intervals based on the observed values for sensitivity, specificity, and base rate with sample sizes of, respectively, N = 150, 300, and 1,000. I used Herbert's (2011) Excel spreadsheet to generate 95% approximate intervals based on the likelihood ratios. Finally, I used the Crawford et al. (2009) computer tool to construct an approximate 95% confidence interval based on the posttest probability of the disorder after a positive test result. You should verify that the confidence intervals reported next are correct:

Sensitivity = .80, 95% CI [.736, .864] Specificity = .70, 95% CI [.648, .752]

BR = .15, 95% CI [.128, .172]

PLR = 2.667, 95% CI [2.204, 3.226]

NLR = .286, 95% CI [.206, .397]

$odds_{post+}$ = .471, 95% CI [.364, .610]

p_{post+} = .320, 95% CI [.267, .379]

RESEARCH EXAMPLES

The examples presented next demonstrate effect size estimation with binary outcomes.

Math Skills and Statistics Course Outcome

The data for this example are described in Chapter 5. Briefly, a total of 667 students in introductory statistics courses completed a basic math skills test at the beginning of the semester. Analysis of scores at the case level indicated that students with test scores < 40% correct had higher rates of unsatisfactory course outcomes (see Table 5.7). These data are summarized in the fourfold table on the left side of Table 6.6. Among the 75 students with math test scores < 40%, a total of 34 had unsatisfactory outcomes, so $p_{< 40\%}$ = 34/75, or .453, 95% CI [.340, .566]. A total of 122 of 592 students

TABLE 6.6
Fourfold Table for the Relation Between Outcomes in Introductory Statistics and Level of Performance on a Basic Math Skills Test for the Data in Table 5.7

		Course outcome		Effect size						
Math score (%)	*n*	Satisfactory	Unsatisfactory	$p_{< 40\%}$	$p_{\geq 40\%}$	RD	RR	OR	logit *d*	$\hat{\varphi}$
40–100	592	470	122	.453[a]	.206[b]	.247[c]	2.199[d]	3.192[e]	.640	.185
< 40	75	41	34							

Note. All confidence intervals are approximate. χ^2 (1) = 22.71. RD = risk difference; RR = risk ratio; OR = odds ratio; CI = confidence interval.
[a]95% CI for $\pi_{< 40\%}$ [.340, .566].
[b]95% CI for $\pi_{\geq 40\%}$ [.173, .239].
[c]95% CI for $\pi_{< 40\%} - \pi_{\geq 40\%}$ [.129, .364].
[d]95% CI for $\pi_{< 40\%} / \pi_{\geq 40\%}$ [1.638, 2.953].
[e]95% CI for ω [1.943, 5.244].

with math test scores ≥ 40% had unsatisfactory outcomes, so their risk rate is $p_{\geq 40\%} = 122/592$, or .206, 95% CI [.173, .239]. The observed risk difference is RD = .453 − .206, or .247, 95% CI [.129, .364], so students with the lowest math test scores have about a 25% greater risk for poor outcomes in introductory statistics.

The risk ratio is RR = .453/.206, or 2.199, 95% CI [1.638, 2.953], which says that the rate of unsatisfactory course outcomes is about 2.2 times higher among the students with the lowest math scores (Table 6.6). The odds ratio is

$$\text{OR} = \frac{.453/(1-.453)}{.206/(1-.206)} = 3.192$$

with 95% CI [1.943, 5.244], so the odds of doing poorly in statistics are about 3.2 times higher among students with math scores < 40% correct than among their classmates with higher scores. The correlation between the dichotomies of < 40% versus ≥ 40% correct on the math test and unsatisfactory–satisfactory course outcome is $\hat{\varphi} = .185$, so the former explains about 3.4% of the variance in the latter.

Screening for Urinary Incontinence

About one third of women at least 40 years of age experience urinary incontinence. There are two major types, urge (leakage happens with the urge to urinate) and stress (urine leaks when stretching, straining, or coughing). Although both urge and stress incontinence may be reduced through behavioral intervention, such as bladder control, urge incontinence can be effectively treated with antimuscarinic or anticholinergic medications, and stress incontinence is treated with pelvic muscle exercises and surgery (Holroyd-Leduc & Straus, 2004). Comprehensive differential diagnosis of urge versus stress incontinence involves neurologic and pelvic examination, measurement of residual urine volume after voiding, a test for urinary tract infection, a cough stress test, and keeping a voiding diary. This diagnostic regimen is expensive, invasive, and impractical in many primary care settings.

J. S. Brown et al. (2006) devised a questionnaire intended to categorize types of urinary incontinence. Based on women's responses to the 3 Incontinence Questions (3IQ) scale, their pattern of urinary incontinence is classified as urge, stress, mixed (both urge and stress), or other. The 3IQ was administered in a sample of 301 women with untreated incontinence. The same women were also individually examined by urologists or urogynecologists. The final diagnoses from these examinations were the criterion for the 3IQ. There were no control samples. Instead, J. S. Brown et al. (2006) estimated the sensitivity, specificity,

likelihood ratios, and posttest probabilities of the 3IQ in the differentiation of urge versus stress incontinence. Their reporting of the results is a model in that confidence intervals were calculated for all point estimates and significance testing played little substantive role in the analysis. They also explicitly compared their results with those from other published studies in the same area. But they did not estimate score reliabilities of the 3IQ, which is a drawback.

J. S. Brown et al. (2006) reported separate results for the classification of urge versus stress incontinence based on the 3IQ, but just the former set of findings is summarized next. The sensitivity of the 3IQ in the prediction of urge incontinence is .75, 95% CI [.68, .81], so 75% of women diagnosed by clinical examination as having urge incontinence were correctly classified. The specificity is .77, 95% CI [.69, .84], which says that 77% of women diagnosed as having a non-urge type of incontinence were correctly identified. The positive and negative likelihood ratios are, respectively, PLR = 3.29, 95% CI [2.39, 4.51], and NLR = .32, 95% CI [.24, .43]. Thus, a classification of urge incontinence is about 3.3 times more likely among women who actually have this condition based on clinical examinations than among women who have other types of urinary incontinence. Also, a classification of a non-urge type of incontinence is only about one third as likely among women who actually manifest urge incontinence than among women who have other, non-urge varieties of incontinence.

Given the sensitivity, specificity, and likelihood ratios just summarized, J. S. Brown et al. (2006) estimated posttest probabilities of urge urinary incontinence given a classification of such on the 3IQ at the three different base rates, 25, .50, and .75. These rates are estimated proportions of urge incontinence among women who are, respectively, < 40, 40–60, and > 60 years old. For a base rate of .25, the posttest probability of urge incontinence given a positive test result is .52, 95% CI [.44, .60], or about double the pretest probability. For base rates of .50 and .75, the estimated posttest probabilities of urge incontinence given positive test results are, respectively, .77, 95% CI [.70, .82] and .91, 95% CI [.88, .93]. J. S. Brown et al. (2006) described these results as indicating a modest level of accuracy, but they noted that the use of the 3IQ in combination with urine analysis to rule out urinary tract infection and hematuria (blood in the urine) may help to triage the treatment of women with urinary incontinence.

CONCLUSION

Effect sizes for categorical outcomes were introduced in this chapter. Some measure comparative risk across two groups for a less desirable versus a more desirable outcome where the data are summarized in a fourfold

table. The risk effect size with the best overall properties is the odds ratio, which can be converted to logistic *d* statistics that are comparable to *d*-type effect sizes for continuous outcomes. The sensitivity, specificity, and predictive value framework takes direct account of base rates when estimating the decision accuracy of screening tests. Likelihood ratios estimate the degree to which the odds of having the target disorder change given either a positive or a negative result. Considered in the next chapter is effect size estimation in single-factor designs with ≥3 conditions and continuous outcomes.

LEARN MORE

The text by Agresti (2007) is a good resource about statistical techniques for categorical data. Drobatz (2009) and Halkin, Reichman, Schwaber, Paltiel, and Brezis (1998) discuss evaluation of screening test accuracy and the role of likelihood ratios in diagnosis.

Agresti, A. (2007). *An introduction to categorical data analysis*. Hoboken, NJ: Wiley. doi:10.1002/0470114754

Drobatz, K. J. (2009). Measures of accuracy and performance of diagnostic tests. *Journal of Veterinary Cardiology, 11*(Suppl. 1), S33–S40. doi:10.1016/j.jvc.2009.03.004

Halkin, A., Reichman, J., Schwaber, M., Paltiel, O., & Brezis, M. (1998). Likelihood ratios: Getting diagnostic testing into perspective. *Quarterly Journal of Medicine, 91,* 247–258. doi:10.1093/qjmed/91.4.247

EXERCISES

1. Show that $\hat{\varphi}$ is margin bound using the following fourfold table:

	Relapsed	Not relapsed
Control	60	40
Treatment	40	60

2. For RR = 1.50 and $n_C = n_T = 100$, construct the approximate 95% confidence interval for π_C/π_T.
3. Calculate Cramer's *V* for the 4×2 cross-tabulation in Table 5.7.
4. Calculate positive and negative predictive values for a base rate of .375 for a screening test where sensitivity = .80 and specificity = .70.
5. For sensitivity = .80, specificity = .70, and base rate = .15, I earlier calculated these results: $odds_{post+} = .471$, $p_{post+} = .176$, $odds_{post-} = .050$, and $p_{post-} = .048$. Now assume that the base rate is .50 among high-risk persons. Recalculate the posttest odds and probabilities.

This page intentionally left blank

7

SINGLE-FACTOR DESIGNS

> Most of us are far more comfortable with the pseudo-objectivity of null hypothesis significance testing than we are with making subjective yet informed judgments about the meaning of our results.
>
> —Paul D. Ellis (2010, p. 43)

Effect size estimation in single-factor designs with ≥3 conditions and continuous outcomes is covered next. Because the omnibus comparison of all means is often uninformative, greater emphasis is placed on focused comparisons (contrasts), each with a single degree of freedom. A relatively large omnibus effect can also be misleading if it is due to a single discrepant mean not of substantive interest. Contrast specification and effect size estimation with standardized mean differences are described next. Later sections deal with measures of association for fixed or random factors and special issues for effect size estimation in covariate analyses. Chapter exercises build computational skills for effect size estimation in these designs.

DOI: 10.1037/14136-007
Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, Second Edition, by R. B. Kline

CONTRAST SPECIFICATION AND TESTS

A contrast is a directional effect that corresponds to a particular facet of the omnibus effect. It is often represented with the symbols ψ or $\hat{\psi}$. The former is a parameter that represents a weighted sum of population means:

$$\psi = \sum_{i=1}^{a} c_i \mu_i \tag{7.1}$$

where $(c_1, c_2, \ldots, c_a)$ is the set of **contrast weights (coefficients)** that specify the comparison. Application of the same weights to sample means estimates ψ:

$$\hat{\psi} = \sum_{i=1}^{a} c_i M_i \tag{7.2}$$

Contrast weights should respect a few rules: They must sum to zero, and weights for at least two different means should not equal zero. Means assigned a weight of zero are excluded, and means with positive weights are contrasted with means given negative weights. Suppose factor A has $a = 3$ levels. The weights $(1, 0, -1)$ meet the requirements just stated and specify

$$\hat{\psi}_1 = (1)\, M_1 + (0)\, M_2 + (-1)\, M_3 = M_1 - M_3$$

which is the **pairwise comparison** of M_1 with M_3 excluding M_2. The weights $(-1, 0, 1)$ just change the sign of $\hat{\psi}_1$. Thus, a contrast's sign is arbitrary, but one should always explain the meaning of positive versus negative contrasts. By the same logic, the sets of weights

$$(\tfrac{1}{2}, 0, -\tfrac{1}{2}), (5, 0, -5), \text{and } (1.7, 0, -1.7)$$

among innumerable others with the same pattern of coefficients, all specify the same pairwise comparison as the set $(1, 0, -1)$. The scale of $\hat{\psi}_1$ depends on which sets of weights are applied to the means. This does not affect statistical tests or measures of association for contrasts because their equations correct for the scale of the weights.

But the scale of contrast weights is critical if a comparison should be interpreted as the difference between the averages of two subsets of means. If so, the weights should make up a **standard set** and satisfy what Bird (2002) called **mean difference scaling**: The sum of the absolute values of the coefficients in a standard set is 2.0. This implies for a pairwise comparison that one weight must be +1, another must be −1, and the rest are all zero. For example,

the coefficients (1, 0, −1) are a standard set for comparing M_1 with M_3, but the set (½, 0, −½) is not.

At least three means contribute to a **complex comparison**. An example is when a control condition is compared with the average of two treatment conditions. A complex comparison is still a single-*df* effect because only two means are compared, at least one of which is averaged over ≥2 conditions. A standard set of weights for a complex comparison is specified as follows: The coefficients for one subset of conditions to be averaged together each equal +1 divided by the number of conditions in that subset; the coefficients for the other subset of conditions to be averaged together each equal −1 divided by the number of conditions in that subset; weights for any excluded condition are zero. For example, the coefficients (½, −1, ½) form a standard set for comparing M_2 with the average of M_1 and M_3:

$$\hat{\psi}_2 = (\tfrac{1}{2})M_1 + (-1)M_2 + (\tfrac{1}{2})M_3 = \frac{M_1 + M_3}{2} - M_2$$

But the weights (1, −2, 1) for the same pattern are not a standard set because the sum of their absolute values is 4.0, not 2.0.

Two contrasts are **orthogonal** if they each reflect an independent aspect of the omnibus effect; that is, the result in one comparison says nothing about what may be found in the other. In balanced designs (i.e., equal group sizes), a pair of contrasts is orthogonal if the sum of the products of their corresponding weights is zero, or

$$\sum_{i=1}^{a} c_{1_i} c_{2_i} = 0 \tag{7.3}$$

But in unbalanced designs, two contrasts are orthogonal if

$$\sum_{i=1}^{a} \frac{c_{1_i} c_{2_i}}{n_i} = 0 \tag{7.4}$$

where n_i is the number of cases in the *i*th condition. Otherwise, a pair of contrasts is **nonorthogonal**, and such contrasts describe overlapping facets of the omnibus effect.

Presented next is a pair of weights for contrasts in a balanced design with $a = 3$ levels:

$$\hat{\psi}_1: \quad (1, \quad 0, \quad -1)$$

$$\hat{\psi}_2: \quad (\tfrac{1}{2}, \quad -1, \quad \tfrac{1}{2})$$

This pair is orthogonal because the sum of the cross-products of their weights is zero, or

$$\sum_{i=1}^{a} c_{1_i} c_{2_i} = (1)\ (\tfrac{1}{2}) + (0)\ (-1) + (-1)(\tfrac{1}{2}) = 0$$

Intuitively, these contrasts are unrelated because the two means compared in $\hat{\psi}_1$, M_1 and M_3, are combined in $\hat{\psi}_2$ and contrasted against the third mean, M_2. The weights for a second pair of contrasts in the same design are listed next:

$$\hat{\psi}_2: \quad (\tfrac{1}{2}, \quad -1, \quad -\tfrac{1}{2})$$

$$\hat{\psi}_3: \quad (1, \quad -1, \quad 0)$$

This second pair is not orthogonal because the sum of the weight cross-products is not zero:

$$\sum_{i=1}^{a} c_{2_i} c_{3_i} = (\tfrac{1}{2})(1) + (-1)(-1) + (\tfrac{1}{2})(0) = 1.5$$

Contrasts $\hat{\psi}_2$ and $\hat{\psi}_3$ are correlated because M_2 is one of the two means compared in both.

If every pair in a set of contrasts is orthogonal, the entire set is orthogonal. The maximum number of orthogonal contrasts is limited by the degrees of freedom for the omnibus effect, df_A. Thus, the omnibus effect can theoretically be broken down into $a - 1$ independent directional effects, where a is the number of groups. Expressed in terms of sums of squares, this is

$$SS_A = \sum_{i=1}^{a-1} SS_{\hat{\psi}_i} \tag{7.5}$$

where SS_A and $SS_{\hat{\psi}_i}$ are, respectively, the sum of squares for the omnibus effect and the ith contrast in a set of $a - 1$ orthogonal comparisons. The same idea can be expressed in terms of the correlation ratio

$$\hat{\eta}_A^2 = \sum_{i=1}^{a-1} \hat{\eta}_{\psi_i}^2 \tag{7.6}$$

where $\hat{\eta}_A^2$ and $\hat{\eta}_{\psi_i}^2$ are, respectively, estimated eta-squared for the omnibus effect and the ith contrast in a set of all possible orthogonal comparisons.

There is generally more than one set of orthogonal comparisons that could be specified for $a \geq 3$ conditions. For example, the contrasts defined by the weights listed next

$$\hat{\psi}_1: \quad (1, \quad 0, \quad -1)$$

$$\hat{\psi}_2: \quad (\tfrac{1}{2}, \quad -1, \quad \tfrac{1}{2})$$

make up an orthogonal set. A different pair of orthogonal contrasts for the same design is

$$\hat{\psi}_3: \quad (1, \quad -1, \quad 0)$$

$$\hat{\psi}_4: \quad (\tfrac{1}{2}, \quad \tfrac{1}{2}, \quad -1)$$

Any set of $a - 1$ orthogonal contrasts satisfies Equations 7.5–7.6. Different sets of orthogonal contrasts for the same factor just specify different patterns of directional effects among its levels (groups), but altogether any set of orthogonal contrasts will capture all possible ways that the groups could differ. Statisticians like sets of orthogonal contrasts because of the independence of the directional effects specified by them. But it is better to specify a set of nonorthogonal contrasts that addresses substantive questions than a set of orthogonal contrasts that does not.

If levels of the factor are equally spaced along a quantitative scale, such as the drug dosages 3, 6, and 9 mg $\cdot$ kg^{-1}, special contrasts called **trends** or **polynomials** can be specified. A trend describes a particular shape of the relation between a continuous factor and outcome. There are as many possible trend components as there are degrees of freedom for the omnibus effect. For example, if a continuous factor has three equally spaced levels, there are only two possible trends, linear and quadratic. If there are four levels, an additional polynomial, a cubic trend, may be present. But it is relatively rare in behavioral research to observe nonlinear trends beyond cubic effects (e.g., dose–response or learning curves). Because trends in balanced designs are orthogonal, they are called **orthogonal polynomials**.

There are conventional weights for polynomials, and some computer procedures that analyze trends automatically generate them. Tables of polynomial weights are also available in several sources (e.g., Winer et al., 1991, p. 982). For example, the weights (–1, 0, 1) define a linear trend (positive or negative) for continuous factors with three equally spaced levels, and the set (1, –2, 1) specifies a quadratic trend (*U*-shaped or inverted-*U*-shaped). Most trend weights for larger designs are not standard sets, but this is not a problem because magnitudes of trends are usually estimated with measures of association, not standardized mean differences. A research example presented later concerns trend analysis.

Test Statistics

The general form of the test statistic for contrasts is

$$t_{\hat{\psi}}(df) = \frac{\hat{\psi} - \psi_0}{s_{\hat{\psi}}} \tag{7.7}$$

where ψ_0 is the value of contrast specified in H_0 and $s_{\hat{\psi}}$ is the standard error. For a nil hypothesis, $\psi_0 = 0$ and this term drops out of the equation. If the means are independent, $df = df_W = N - a$, the pooled within-groups degrees of freedom, and the standard error is

$$s_{\hat{\psi}\ \text{ind}} = \sqrt{MS_W \left(\sum_{i=1}^{a} \frac{c_i^2}{n_i} \right)} \tag{7.8}$$

where MS_W is the pooled within-groups variance (see Equation 3.11).

In a correlated design, the degrees of freedom for $t_{\hat{\psi}}$ are $n - 1$, where n is the group size, and the standard error takes the form

$$s_{\hat{\psi}\ \text{dep}} = \sqrt{\frac{s_{D_{\hat{\psi}}}^2}{n}} \tag{7.9}$$

where the term in the numerator under the radical is the variance of the contrast difference scores. Suppose the weights (1, 0, −1) define $\hat{\psi}_1$ in a correlated design with three conditions. If Y_1, Y_2, and Y_3 are scores from these conditions, the difference score is computed for each case as

$$D_{\hat{\psi}_1} = (1)\ Y_1 + (0)\ Y_2 + (-1)\ Y_3 = Y_1 - Y_3$$

The variance of these difference scores reflects the cross-conditions correlation r_{13}, or the subjects effect for this contrast (see Equation 2.21). The error term $s_{\hat{\psi}\ \text{dep}}$ does not assume sphericity because just two means are compared in any contrast.

Some computer programs, such as SPSS, print $F_{\hat{\psi}}$ for contrasts instead of $t_{\hat{\psi}}$. For a nil hypothesis, $t_{\hat{\psi}}^2 = F_{\hat{\psi}}$, but $t_{\hat{\psi}}$ preserves the sign of the contrast and can test non-nil hypotheses, too. The form for a contrast between independent means is $F_{\hat{\psi}}(1, df_W) = SS_{\hat{\psi}}/MS_W$, where the numerator equals

$$SS_{\hat{\psi}} = \frac{\hat{\psi}^2}{\sum_{i=1}^{a} \frac{c_i^2}{n_i}} \tag{7.10}$$

The error term for $F_{\hat{\psi}}$ in designs with dependent samples is typically not the omnibus error term (e.g., $MS_{A \times S}$ for a nonadditive model) that assumes sphericity when $a \geq 3$. Instead, it is based on scores from just the two subsets of conditions involved in the contrast. The degrees of freedom for $F_{\hat{\psi}}$ in this case are 1, $n - 1$. If the samples are independent, researchers can compute a standardized mean difference or a correlation effect size for a contrast from either $t_{\hat{\psi}}$ or $F_{\hat{\psi}}$ for a nil hypothesis. This makes these test statistics useful even when a nil hypothesis is likely false.

Controlling Type I Error and ANOVA Rituals

Methods for **planned comparisons** assume a relatively small number of a priori contrasts, but those for **unplanned comparisons** anticipate a larger number of post hoc tests, such as all possible pairwise contrasts. A partial list of methods is presented in Table 7.1 in ascending order by degree of protection against experimentwise Type I error and in descending order by power. These methods generally use $t_{\hat{\psi}}$ or $F_{\hat{\psi}}$ as test statistics but compare them against critical values higher than those from standard tables (i.e., it is more difficult to reject H_0). For example, the adjusted level of statistical significance for an individual contrast (α_{Bon}) in the Bonferroni–Dunn method equals α_{EW}/c, where the numerator is the target experimentwise error rate and c is the number of contrasts. Methods in Table 7.1 are also associated with the construction of simultaneous confidence intervals for ψ. See Hsu (1996)

TABLE 7.1
Methods for Controlling Type I Error Over Multiple Comparisons

Method	Nature of protection against α_{EW}
Planned comparisons	
Unprotected	None; uses standard critical values for $t_{\hat{\psi}}$ or $F_{\hat{\psi}}$
Dunnett	Across pairwise comparisons of a single control group with each of $a - 1$ treatment groups
Bechhofer–Dunnett	Across a maximum of $a - 1$ orthogonal a priori contrasts
Bonferroni–Dunn	Bonferroni correction applied across total number of either orthogonal or correlated contrasts
Unplanned comparisons	
Newman–Keuls	Across pairwise comparisons within sets of means ordered by differences in rank order
Tukey HSD	Across all possible pairwise comparisons
Scheffé	Across all possible pairwise or complex comparisons

Note. α_{EW} = experimentwise Type I error; HSD = honestly significant difference. Tukey HSD is also called Tukey A.

for information about additional methods to control for Type I error across multiple comparisons.

Controlling Type I error over contrast tests may not be desirable if power of the unprotected tests is already low. There is also little need to worry about experimentwise Type I error if few contrasts are specified. Wilkinson and the TFSI (1999) noted that the ANOVA ritual of routinely testing all pairwise comparisons following rejection of H_0 for the omnibus effect is typically wrong. This approach makes individual comparisons unnecessarily conservative when a classical post hoc method (e.g., Scheffé) is used, and it is rare that all such contrasts are interesting. The cost for reducing Type I error in this case is reduced power for the specific tests the researcher really cares about. This ritual is also a blind search for statistical significance, somewhere, anywhere, among pairs of means. It should be avoided in favor of analyzing contrasts of substantive interest with emphasis on effect sizes and confidence intervals.

Confidence Intervals for ψ

The general form of a 100 $(1 - \alpha)$% confidence interval for ψ is

$$\hat{\psi} \pm s_{\hat{\psi}} \left[t_{\hat{\psi}\,2\text{-tail},\,\alpha}(df_{\text{error}}) \right] \tag{7.11}$$

where the standard error is defined by Equation 7.8 for independent samples and by Equation 7.9 for dependent samples. The degrees of freedom for the critical value of $t_{\hat{\psi}}$ equal those of the corresponding error term. **Simultaneous (joint) confidence intervals** are based on sets of contrasts, and they are generally wider than confidence intervals for individual contrasts defined by Equation 7.11. This is because the former control for multiple comparisons. Suppose in the Bonferroni–Dunn method that $\alpha_{EW} = .05$ and $c = 10$, which implies $\alpha_{Bon} = .05/10$, or .005 for each comparison. The resulting 10 simultaneous 99.5% confidence intervals for ψ are each based on $t_{\hat{\psi}\,2\text{-tail},\,.005}$, and these intervals are wider than the corresponding 95% confidence interval based on $t_{\hat{\psi}\,2\text{-tail},\,.05}$ for any individual contrast; see Bird (2002) for more information.

STANDARDIZED CONTRASTS

Contrast weights that are standard sets are assumed next. A standardized mean difference for a contrast is **standardized contrast**. It estimates the parameter $\delta_\psi = \psi/\sigma^*$, where the numerator is the unstandardized population contrast and the denominator is a population standard deviation. The general form of the sample estimator is $d_{\hat{\psi}} = \hat{\psi}/\hat{\sigma}^*$, where the denominator (standardizer) is an estimator of σ^* that is not the same in all kinds of standardized contrasts.

Independent Samples

Two methods for standardizing contrasts in designs with ≥3 independent samples are described next. The first method may be suitable for pairwise comparisons, but the second method is good for simple or complex comparisons.

1. Calculate a standardized contrast using one of the methods for two-group designs described in Chapter 5. These methods differ according to specification of the standardizer and whether the estimators are robust or not robust (see Table 5.2). In designs with ≥3 groups, the standardizer is based on data from just the two groups involved in a particular comparison. For example, the standardizer s_{pool} for comparing M_1 and M_2 in a three-group design is the pooled within-groups standard deviation that excludes s_3, the standard deviation from the third group. It assumes homoscedasticity, but selection of either s_1 or s_2 as the standardizer does not. Robust alternatives include $s_{Win\,p}$ based on the pooled within-groups 20% Winsorized variances from groups 1 and 2 or either s_{Win1} or s_{Win2} from just one of these groups. Keselman et al. (2008) described SAS/STAT syntax that calculates robust standardized contrasts with confidence intervals in between-subjects single-factor designs that can be downloaded.[1] A drawback of these methods is that each contrast is based on a different standardizer, and each standardizer ignores information about variability in groups not involved in a particular contrast.
2. Select a common standardizer for any comparison, pairwise or complex, based on all information about within-groups variability. An example is the square root of MS_W, the pooled within-groups variance for the whole design. Contrasts standardized against $(MS_W)^{1/2}$ are designated next as d_{with}. In two-group designs, $d_{with} = d_{pool}$ for the sole contrast. But in larger designs, values of d_{with} and d_{pool} may differ for the same contrast. The effect size d_{with} assumes homoscedasticity over all groups. An alternative standardizer is the unbiased standard deviation for the total data set, $s_T = (SS_T/df_T)^{1/2}$. This option estimates the full range of variation for nonexperimental factors.

The value of $d_{with} = \hat{\psi}/(MS_W)^{1/2}$ can also computed from $t_{\hat{\psi}}$ for a nil hypothesis, the contrast weights, and the group sizes:

$$d_{with} = t_{\hat{\psi}} \sqrt{\sum_{i=1}^{a} \frac{c_i^2}{n_i}} \tag{7.12}$$

[1]http://supp.apa.org/psycarticles/supplemental/met_13_2_110/met_13_2_110_supp.html

In balanced designs, this equation reduces to $d_{\text{with}} = t_{\hat{\psi}}(2/n)^{1/2}$ for pairwise contrasts.

Look back at Table 3.4, which lists raw scores for a balanced three-sample design where $n = 5$, $M_1 = 13.00$, $M_2 = 11.00$, and $M_3 = 15.00$. Reported in Table 7.2 are results of an independent samples ANOVA for the omnibus effect and two orthogonal contrasts defined by the sets of weights (1, 0, –1) and (½, –1, ½) where

$$\hat{\psi}_1 = M_1 - M_3 = 13.00 - 15.00 = -2.00$$

$$\hat{\psi}_2 = \frac{M_1 + M_3}{2} - M_2 = \frac{13.00 + 15.00}{2} - 11.00 = 3.00$$

Note in Table 7.2 that

$$SS_A = SS_{\hat{\psi}1} + SS_{\hat{\psi}2} = 10.00 + 30.00 = 40.00$$

which is just as predicted by Equation 7.5 for a set of $a - 1 = 2$ orthogonal contrasts. Given $MS_W = 5.50$, values of the corresponding standardized contrasts are

$$d_{\text{with}1} = \frac{-2.00}{\sqrt{5.50}} = -.85 \quad \text{and} \quad d_{\text{with}2} = \frac{3.00}{\sqrt{5.50}} = 1.28$$

In words, M_1 is .85 standard deviations lower than M_3, and the average of M_1 and M_3 is 1.28 standard deviations higher than M_2.

TABLE 7.2
Independent Samples Analysis of the Data in Table 3.4

Source	*SS*		*df*	*MS*	*F*	d_{with}	$\hat{\eta}^2$		Partial $\hat{\eta}^2$
Between (*A*)	40.00		2	20.00	3.64[c]	—	.377[h]		.377
$\hat{\psi}_1 = -2.00$[a]		10.00	1	10.00	1.82[d]	–.85[f]		.094	.132[i]
$\hat{\psi}_2 = 3.00$[b]		30.00	1	30.00	5.45[e]	1.28[g]		.283	.313[j]
Within (error)	66.00		12	5.50					
Total	106.00		14						

Note. The contrast weights for $\hat{\psi}_1$ are (1, 0, –1) and those for $\hat{\psi}_2$ are (½, –1, ½). A dash (—) indicates that it is not possible to calculate the statistic indicated in the column heading for the effect listed in that row of the table. CI = confidence interval.

[a]95% CI for ψ_1 [–5.23, 1.23]. [b]95% CI for ψ_2 [.20, 5.80]. [c]$p = .058$. [d]$p = .202$. [e]$p = .038$. [f]Approximate 95% CI for $\delta_{\psi 1}$ [–2.23, .53]. [g]Approximate 95% CI for $\delta_{\psi 2}$ [.09, 2.47]. [h]Noncentral 95% CI for η_A^2 [0, .601]. [i]Noncentral 95% CI for partial $\eta_{\hat{\psi}1}^2$ [0, .446]. [j]Noncentral 95% CI for partial $\eta_{\hat{\psi}2}^2$ [0, .587].

Dependent Samples

There are two basic ways to standardize mean changes when the samples are dependent:

1. With one exception, use any of the methods described in the previous section for contrasts between unrelated means. These methods estimate population standard deviations in the metric of the original scores, but they ignore the subjects effect in correlated designs. The exception is Equation 7.12, which requires $t_{\hat{\psi}}$ for independent samples to compute d_{with}.
2. Standardize the mean change against the standard deviation of the difference scores for that particular contrast. This option takes account of the cross-conditions correlation, but it does not describe change in the metric of the original scores.

Reported in Table 7.3 are the results of a dependent samples analysis for an additive model of the data in Table 3.4 for the omnibus effect and the same two contrasts analyzed in Table 7.2. The F and p values differ for all effects across the independent samples analysis in Table 7.2 and the dependent samples analysis in Table 7.3. But $d_{\text{with}1} = -.85$ and $d_{\text{with}2} = 1.28$ in both analyses, because each is calculated the same way regardless of the design.

Approximate Confidence Intervals for δ_ψ

If the samples are independent and the effect size is d_{with}, an approximate confidence interval for δ_ψ can be obtained by dividing the endpoints of the

TABLE 7.3
Dependent Samples Analysis of the Data in Table 3.4

Source	*SS*		*df*	*MS*	*F*	d_{with}	$\hat{\eta}^2$	Partial $\hat{\eta}^2$
Between (*A*)	40.00		2	20.00	14.12[c]	—	.377	.779
$\hat{\psi}_1 = -2.00$[a]		10.00	1	10.00	5.71[d]	−.85[f]	.094	.588
$\hat{\psi}_2 = 3.00$[b]		30.00	1	30.00	27.69[e]	1.28[g]	.283	.874
Within	66.00		12	5.50				
Subjects		54.67	4	13.67				
Residual (*A*)		11.33	8	1.42				
Residual ($\hat{\psi}_1$)		7.00	4	1.75				
Residual ($\hat{\psi}_2$)		4.33	4	1.08				
Total	106.00		14					

Note. The contrast weights for $\hat{\psi}_1$ are (1, 0, −1) and those for $\hat{\psi}_2$ are (½, −1, ½). A dash (—) indicates that it is not possible to calculate the statistic indicated in the column heading for the effect listed in that row of the table.
[a]95% CI for ψ_1 [−4.32, .32]. [b]95% CI for ψ_2 [1.42, 4.58]. [c]$p = .002$. [d]$p = .075$. [e]$p = .006$. [f]Approximate 95% CI for $\delta_{\psi 1}$ [−1.84, .14]. [g]Approximate 95% CI for $\delta_{\psi 2}$ [.60, 1.95].

corresponding interval for ψ (Equation 7.11) by the square root of MS_W. The form of the resulting interval is

$$d_{\text{with}} \pm s_{d\ \text{with}} \left[t_{\hat{\psi}\,2\text{-tail},\,\alpha}(df_W) \right] \tag{7.13}$$

where the approximate standard error is

$$s_{d\ \text{with}} = \sqrt{\sum_{i=1}^{a} \frac{c_i^2}{n_i}} \tag{7.14}$$

Four-decimal accuracy is recommended for hand calculations.

Bird (2002) recommended the same method when the samples are dependent and the effect size is d_{with} except that the degrees of freedom for $t_{\hat{\psi}\,2\text{-tail},\,\alpha}$ are $n - 1$, not df_W. The resulting confidence interval for ψ controls for the subjects effect. Next, divide the lower and upper bounds of this confidence interval by the square root of MS_W, which standardizes the interval in the metric of the original scores. The result is the approximate confidence interval for δ_ψ for a dependent standardized mean change.

Refer back to Table 7.2 and look over the results of the independent samples analysis where $df_W = 12$ and $MS_W = 5.50$. The standard error of $\hat{\psi}_1 = -2.00$ is

$$s_{\hat{\psi}1\ \text{ind}} = \sqrt{5.50\left(\frac{1^2}{5} + \frac{0^2}{5} + \frac{-1^2}{5}\right)} = 1.4832$$

Given $t_{\hat{\psi}\,2\text{-tail},\,.05}\,(12) = 2.179$, the 95% confidence interval for ψ_1 is

$$-2.00 \pm 1.4832\ (2.776)$$

which defines the interval [–5.2319, 1.2319]. If we divide the endpoints of this interval by $5.50^{1/2} = 2.345$, we obtain the approximate 95% confidence interval for $\delta_{\psi 1}$. The lower and upper bounds of this interval based on $d_{\text{with}1} = -.85$ are [–2.2309, .5253]. Thus, we can say that $d_{\text{with}1} = -.85$ is just as consistent with a population effect size as small as $\delta_{\psi 1} = -2.23$ as it is with a population effect size as large as $\delta_{\psi 1} = .53$, with 95% confidence. The range of imprecision is this great due to the small group size ($n = 5$). Exercise 1 involves constructing the approximate 95% confidence interval for $\delta_{\psi 2}$ based on the results in Table 7.2.

Now look at the results of the dependent samples analysis in Table 7.3. Given $r_{13} = .7303$, $s_1^2 = 7.50$, and $s_3^2 = 4.00$ for $\hat{\psi}_1 = -2.00$ (see Table 3.4), the variance of the difference scores and the standard error of $\hat{\psi}_1$ are, respectively,

$$s^2_{D_{\hat{\psi}1}} = 7.50 + 4.00 - 2\sqrt{7.50(4.00)}(.7303) = 3.50$$

$$s_{\hat{\psi}1\text{ dep}} = \sqrt{\frac{3.50}{5}} = .8367$$

Given $t_{\hat{\psi}2\text{-tail},\,.05}\,(4) = 2.776$, the 95% confidence interval for ψ_1 is

$$-2.00 \pm .8367\,(2.776)$$

which defines the interval [–4.3226, .3226]. Dividing the endpoints of this interval by the square root of $MS_W = 5.50$ gives the lower and upper bounds of the approximate 95% confidence interval for $\delta_{\psi 1}$ based on $d_{\text{with}1} = -.85$ in the dependent samples analysis. The resulting interval is [–1.8432, .1376], or [–1.84, .14] at two-decimal accuracy. As expected, this interval for $\delta_{\psi 1}$ in the dependent samples analysis is narrower than the corresponding interval in the independent samples analysis of the same scores, or [–2.23, .53]. Exercise 2 asks you to construct the approximate 95% confidence interval for $\delta_{\psi 2}$ for the dependent samples analysis in Table 7.3.

The PSY computer program (Bird, Hadzi-Pavlovic, & Isaac, 2000) for Microsoft Windows calculates individual or simultaneous approximate confidence intervals for δ_ψ when the effect size is d_{with} in designs with one or more between-subjects or within-subjects factors.[2] It accepts only integer contrast weights, but it can automatically convert the weights to a standard set so that all contrasts are scaled as mean differences.

Results of computer simulations by Algina and Keselman (2003) indicated that Bird's (2002) approximate confidence intervals for δ_ψ were reasonably accurate in between-subjects designs except for larger population effect sizes, such as $\delta_\psi > 1.50$. But in correlated designs, their accuracies decreased as either the population effect size increased or the cross-conditions correlation increased. In both designs under the conditions just stated, approximate confidence intervals for δ_ψ were generally too narrow, which makes the results look falsely precise. Algina and Keselman (2003) also found that noncentral confidence intervals in between-subjects designs and approximate noncentral confidence intervals in within-subjects designs were generally more accurate than Bird's (2002) approximate intervals.

Noncentral Confidence Intervals for δ_ψ

When the means are independent, d_{with} follows noncentral $t_{\hat{\psi}}\,(df_W, \Delta)$ distributions. A computer tool calculates a noncentral 100 $(1 - \alpha)$% confidence

[2]http://www.psy.unsw.edu.au/research/research-tools/psy-statistical-program

interval for δ_ψ by first finding the corresponding confidence interval for Δ. Next, the endpoints of this interval, Δ_L and Δ_U, are converted to δ_ψ units based on the equation

$$\delta_\psi = \Delta \sqrt{\sum_{i=1}^{a} \frac{c_i^2}{n_i}} \tag{7.15}$$

For the independent samples analysis in Table 7.2 where $n = 5$ and $\hat{\psi}_1 = -2.00$ is defined by the weights (1, 0, −1), the standardized contrast is $d_{\text{with}1} = -.85$ and the test statistics are

$$F_{\hat{\psi}1}\ (1, 12) = 1.82 \quad \text{and} \quad t_{\hat{\psi}1}\ (12) = -1.35\ (\textit{i.e.}, t^2_{\hat{\psi}1} = F_{\hat{\psi}1})$$

I used J. H. Steiger's NDC calculator (see footnote 2, Chapter 5) to construct the noncentral 95% confidence interval for Δ, which is [−3.3538, .7088]. When Equation 7.15 is used, the endpoints of this interval in Δ units are transformed to δ_ψ units by multiplying each by

$$\sqrt{\frac{1^2}{5} + \frac{0^2}{5} + \frac{-1^2}{5}} = .6325$$

The result, [−2.1213, .4483], or [−2.12, .45] at two-decimal accuracy, is the noncentral 95% confidence interval for $\delta_{\psi 1}$. Earlier, the approximate 95% confidence interval for $\delta_{\psi 1}$ was calculated based on the same data as [−2.23, .53] (see also Table 7.2).

Two computer tools that construct noncentral confidence intervals for δ_ψ in designs with independent samples include MBESS for R (Kelley, 2007; see footnote 5, Chapter 5) and syntax for SAS/IML presented by Algina and Keselman (2003; see footnote 7, Chapter 5). Distributions of d_{with} in correlated designs are complex and may follow neither central nor noncentral $t_{\hat{\psi}}$ distributions. Algina and Keselman (2003) described a method to calculate approximate confidence intervals for δ_ψ in such designs based on noncentral $t_{\hat{\psi}}$ distributions, and SAS/IML syntax for this method can be downloaded from the source just mentioned. Steiger (2004) described additional methods of interval estimation for contrasts in ANOVA.

Bootstrapped Confidence Intervals Based on Robust Standardized Contrasts

Two software packages or scripts compute bootstrapped confidence intervals for the product of the scale factor .642 and $\delta_{\psi\,\text{rob}}$ (i.e., δ_ψ is estimated) based on trimmed means and Winsorized variances in single- or multiple-

factor designs. These include the SAS/IML script by Keselman et al. (2008) and Wilcox's (2012) WRS package for R (see footnote 11, Chapter 2).

CORRELATIONS AND MEASURES OF ASSOCIATION

Reviewed next are descriptive and inferential *r*-type effect sizes for contrasts and the omnibus effect. The descriptive effect sizes assume a fixed factor (as do standardized contrasts). The most general descriptive effect size is the correlation ratio. For contrasts it takes the form $\hat{\eta}^2_{\psi} = SS_{\psi}/SS_T$, and it measures the proportion of total observed explained by that contrast. The corresponding effect size for the omnibus effect is $\hat{\eta}^2_A = SS_A/SS_T$. In balanced designs with a fixed factor, the inferential measures of association $\hat{\omega}^2_{\psi}$ for contrasts and $\hat{\omega}^2_A$ for omnibus effects control for positive bias in, respectively, $\hat{\eta}^2_{\psi}$ and $\hat{\eta}^2_A$. But for random factors, contrast analysis is typically uninformative. This is because levels of such factors are randomly selected, so they wind up in a particular study by chance. In this case, the appropriate inferential measure of association is the intraclass correlation $\hat{\rho}_I$, which is already in a squared metric, for the omnibus effect in balanced designs.

Correlation Effect Sizes for Contrasts

The unsigned correlation between a contrast and the outcome variable is $\hat{\eta}_{\psi}$. Its signed counterpart in designs with independent samples is $r_{\hat{\psi}}$, which is calculated as follows:

$$r_{\hat{\psi}} = t_{\hat{\psi}} \sqrt{\frac{1}{F_{\hat{\psi}} + df_{\text{non-}\hat{\psi}}(F_{\text{non-}\hat{\psi}}) + df_W}} \tag{7.16}$$

where $df_{\text{non-}\hat{\psi}}$ and $F_{\text{non-}\hat{\psi}}$ are, respectively, the degrees of freedom and *F* statistic for all noncontrast sources of between-groups variability. The statistic $F_{\text{non-}\hat{\psi}} = MS_{\text{non-}\hat{\psi}}/MS_W$, where

$$MS\text{non-}\psi = SS\text{non-}\hat{\psi}/df\,\text{non-}\psi$$

$$SS_{\text{non-}\hat{\psi}} = SS_A - SS_{\hat{\psi}} \quad \text{and} \quad df_{\text{non-}\hat{\psi}} = df_A - 1$$

For the results in Table 7.2 where the weights (1, 0, −1) specify $\hat{\psi}_1$,

$$\hat{\psi}_1 = -2.00,\ SS_{\hat{\psi}1} = 10.00,\ MS_W = 5.50,\ F_{\hat{\psi}1}(1, 12) = 1.82$$

$$t_{\hat{\psi}_1}(12) = -1.35,\ SS_A = 40.00,\ df_A = 2$$

which imply that

$$SS_{\text{non-}\hat{\psi}_1} = 40.00 - 10.00 = 30.00,\ df_{\text{non-}\hat{\psi}_1} = 2 - 1 = 1$$

$$F_{\text{non-}\hat{\psi}_1} = 30.00/5.50 = 5.45$$

Now we calculate

$$r_{\hat{\psi}1} = -1.35\sqrt{\frac{1}{1.82 + (1)\,5.45 + 12}} = -.307$$

So we can say that the correlation between the dependent variable and the contrast between the first and third groups is –.307 and that this contrast explains $-.307^2$, or about .094 (9.4%), of the total observed variance in outcome.

The partial correlation effect size

$$\text{partial } r_{\hat{\psi}} = t_{\hat{\psi}}\sqrt{\frac{1}{F_{\hat{\psi}} + df_W}} \tag{7.17}$$

removes the effects of all other contrasts from total variance. For $\hat{\psi}_1$ in Table 7.2,

$$\text{partial } r_{\hat{\psi}_1} = -1.35\sqrt{\frac{1}{1.82 + 12}} = -.363$$

which says that correlation between $\hat{\psi}_1$ and outcome is –.363 controlling for $\hat{\psi}_2$ and that $\hat{\psi}_1$ explains $-.363^2$, or about .132 (13.2%), of the residual variance. The absolute value of partial $r_{\hat{\psi}}$ is usually greater than that of $r_{\hat{\psi}}$ for the same contrast, which is here true for $\hat{\psi}_1$ (respectively, .363 vs. .307). Also, partial $r^2_{\hat{\psi}}$ values are not generally additive over sets of contrasts, orthogonal or not. Exercise 3 asks you to calculate $r_{\hat{\psi}2}$ and partial $r_{\hat{\psi}2}$ for the results in Table 7.2.

The correlation $r_{\hat{\psi}}$ assumes independent samples. For dependent samples, we can compute instead the unsigned correlation of which $r_{\hat{\psi}}$ is a special case, $\hat{\eta}_{\psi}$. For the dependent contrast $\hat{\psi}_1$ in Table 7.3 where $SS_{\hat{\psi}_1} = 10.00$ and $SS_T = 106.00$, $\hat{\eta}^2_{\psi 1} = (10.00/106.00)^{1/2}$, or .307, which is also the absolute value of $r_{\hat{\psi}1}$ for the same contrast in the independent samples analysis of the same data (see Table 7.2). The proportion of total variance explained is also the same in both analyses, or $\hat{\eta}^2_{\psi_1} = r^2_{\hat{\psi}_1} = .094$ (i.e., 9.4% of total variance).

The general form of partial $\hat{\eta}^2_{\psi}$ is $SS_{\hat{\psi}}/(SS_{\hat{\psi}} + SS_{\text{error}})$, where SS_{error} is the error sum of squares for the contrast. If the samples are independent, partial $\hat{\eta}^2_{\psi}$ controls for all noncontrast sources of between-conditions variability; if the samples are dependent, it also controls for the subjects effect. For example, partial $\hat{\eta}^2_{\psi_1} = .132$ and $\hat{\eta}^2_{\psi_1} = .094$ for the independent samples analysis in Table 7.2.

But partial $\hat{\eta}^2_{\psi_1} = .588$ for the dependent samples analysis in Table 7.3 because it controls for both $\hat{\psi}_2$ and the subjects effect. Exercise 4 involves calculating partial $\hat{\eta}^2_{\psi_2}$ for the results in Table 7.3. Rosenthal et al. (2000) described many examples of contrast analyses in single- and multiple-factor designs.

Descriptive Measures of Association for Omnibus Effects

When $df_A \geq 2$, $\hat{\eta}^2_A$ is the squared multiple correlation (R^2) between the omnibus effect and outcome. If the samples are independent, $\hat{\eta}^2_A$ can also be computed as

$$\hat{\eta}^2_A = \sqrt{\frac{F_A}{F_A + \frac{df_W}{df_A}}} \tag{7.18}$$

where F_A is the test statistic for the omnibus effect with df_A, df_W degrees of freedom (see Equation 3.8). For example, $SS_A = 40.00$ and $SS_T = 106.00$ for the omnibus effect in Tables 7.2 and 7.3, so the omnibus effect explains 40.00/106.00 = .377, or about 37.7%, of the total variance in both analyses. But only for the independent samples analysis in Table 7.2 where F_A (2, 12) = 3.64 can we also calculate (using Equation 7.18) for the omnibus effect

$$\hat{\eta}^2_A = \sqrt{\frac{3.64}{3.64 + \frac{12}{2}}} = .377$$

The general form of partial $\hat{\eta}^2_A$ is $SS_A / (SS_A + SS_{error})$. In a between-subjects design where MS_W is the error term, $\hat{\eta}^2_A$ = partial $\hat{\eta}^2_A$ because there is no other source of systematic variation besides the omnibus effect. But in a within-subjects design, it is generally true that $\hat{\eta}^2_A \leq$ partial $\hat{\eta}^2_A$ because only the latter controls for the subjects effect. For example, $\hat{\eta}^2_A = .377$ but partial $\hat{\eta}^2_A = .779$ for the omnibus effect for the dependent samples analysis in Table 7.3.

Inferential Measures of Association

The inferential measures $\hat{\omega}^2$ for fixed factors and $\hat{\rho}_I$ for random factors in balanced designs are based on ratios of **variance components**, which involve the expression of expected sample mean squares as functions of population sources of systematic versus error variation. Extensive sets of equations for variance component estimators in sources such as Dodd and Schultz (1973), Kirk (2012), Vaughan and Corballis (1969), and Winer et al. (1991) provide the basis for computing inferential measures of association. Schuster and von

Eye (2001) showed that random effects models and repeated measures models are variations of each other because both control for scores dependencies. It is not always possible to estimate population variance components without bias, and certain components cannot be expressed as unique functions of sample data in some designs. But there are heuristic estimators that may let one get by in the latter case.

Both $\hat{\omega}^2$ and $\hat{\rho}_I$ have the general form $\hat{\sigma}^2_{\text{effect}}/\hat{\sigma}^2_{\text{total}}$, where the numerator estimates the variance component for the effect of interest and the denominator estimates total variance due to all sources in the design. For a fixed factor, $\hat{\sigma}^2_{\text{effect}}$ is the numerator of $\hat{\omega}^2_{\text{effect}}$. In designs with either independent or dependent samples, its general form is

$$\hat{\sigma}^2_{\text{effect}} = \frac{df_{\text{effect}}}{an}\left(MS_{\text{effect}} - MS_{\text{error}}\right) \tag{7.19}$$

where MS_{effect} and df_{effect} are, respectively, the effect mean square and its degrees of freedom, and MS_{error} is its error term. When Equation 7.20 is used, the estimator for the omnibus effect is

$$\hat{\sigma}^2_{\text{A fix}} = \frac{a-1}{an}\left(MS_{\text{A}} - MS_{\text{error}}\right) \tag{7.20}$$

and that for a contrast is

$$\hat{\sigma}^2_{\hat{\psi}} = \frac{1}{an}\left(MS_{\hat{\psi}} - MS_{\text{error}}\right) \tag{7.21}$$

(Recall that $MS_{\hat{\psi}} = SS_{\hat{\psi}}$ because $df_{\hat{\psi}} = 1$.) But for a random factor the estimator for the omnibus effect is

$$\hat{\sigma}^2_{\text{A ran}} = \frac{1}{n}\left(MS_{\text{A}} - MS_{\text{error}}\right) \tag{7.22}$$

Estimation of $\hat{\sigma}^2_{\text{total}}$ also depends on the design. The sole estimate of error variance when the samples are independent is MS_W, regardless of whether the factor is fixed or random. This means that total variance in either case is estimated as

$$\hat{\sigma}^2_{\text{total}} = \hat{\sigma}^2_{\text{A}} + MS_W \tag{7.23}$$

where $\hat{\sigma}^2_{\text{A}}$ is defined by Equation 7.20 for a fixed factor but by Equation 7.22 for a random factor. The composition of total variance for fixed versus random factors is thus not the same.

Estimation of total variance is more complicated in correlated designs. For a fixed factor assuming a nonadditive model, it is not possible to uniquely estimate variance components for the subjects effect, the person × treatment interaction, and random error (e.g., Winer et al., 1991, p. 276). Instead, combinations of these parameters are estimated heuristically by the equations for direct calculation of $\hat{\omega}^2_{\text{effect}}$ introduced below. These heuristic estimators are negatively biased, so they generally underestimate population proportions of explained variance. For a random factor, though, the combined effects of person × treatment interaction and random error are estimated with a single error mean square in a nonadditive model.

The equations presented in Table 7.4 allow you to directly calculate $\hat{\omega}^2_{\text{effect}}$ or $\hat{\rho}_I$ instead of working with all the variance component estimators that make up each effect size. For example, the numerator of $\hat{\omega}^2_{\text{effect}}$ for fixed factors in the method of direct calculation is always

$$df_{\text{effect}}\left(\text{MS}_{\text{effect}} - \text{MS}_{\text{error}}\right) \tag{7.24}$$

but computation of the denominator depends on the design. Likewise, the numerator of $\hat{\rho}_I$ for random factors in Table 7.4 is always

$$a\left(\text{MS}_A - \text{MS}_{\text{error}}\right) \tag{7.25}$$

but calculation of its denominator also depends on the design.

Outlined next is a method for direct computation of inferential measures of association based on residual variance. These statistics are partial

TABLE 7.4
Numerators and Denominators for Direct Calculation of Inferential Measures of Association Based on Total Variance for Single-Factor Designs

Sample	Model	Denominator
	Fixed factor[a]	
Independent	—	$SS_T + MS_W$
Dependent	Additive	$SS_T + MS_S$
Dependent	Nonadditive	$SS_T + MS_S + n\,MS_{A\times S}$
	Random factor[b]	
Independent	—	$SS_T + MS_A$
Dependent	Additive	$SS_T + MS_A + MS_S - MS_{\text{res}}$
Dependent	Nonadditive	$a\,MS_A + n\,MS_S + (an - a - n)\,MS_{A\times S}$

Note. The cell size is n, factor A has a levels, and MS_{error} is the ANOVA error term for the corresponding effect.
[a]Effect size is $\hat{\omega}^2_{\text{effect}}$, numerator = $df_{\text{effect}}\,(MS_{\text{effect}} - MS_{\text{error}})$. [b]Effect size is $\hat{\rho}_I$, numerator = $a\,(MS_A - MS_{\text{error}})$.

$\hat{\omega}^2_{\text{effect}}$ for effects of fixed factors and partial $\hat{\rho}_I$ for omnibus effects of random factors. The general form of both is $\hat{\sigma}^2_{\text{effect}}/(\hat{\sigma}^2_{\text{effect}} + \hat{\sigma}^2_{\text{error}})$:

1. *Independent samples.* The error variance estimator is $\hat{\sigma}^2_{\text{error}} = MS_W$.
   a) *Fixed factor.* The numerator of partial $\hat{\omega}^2_\psi$ is $\hat{\sigma}^2_\psi$ (Equation 7.21), and the denominator is $\hat{\sigma}^2_\psi + MS_W$. For the omnibus effect, partial $\hat{\omega}^2_A = \hat{\omega}^2_A$ because there is no other source of between-groups variability and MS_W is the sole error variance estimator.
   b) *Random factor.* Partial $\hat{\rho}_I = \hat{\rho}_I$ for the omnibus effect, for the reasons just stated.
2. *Dependent samples, additive model.* The error variance estimator is $\hat{\sigma}^2_{\text{error}} = MS_{\text{error}}$, the error term for the effect of interest.
   a) *Fixed factor.* The numerator of partial $\hat{\omega}^2_\psi$ is $\hat{\sigma}^2_\psi$ (Equation 7.21), and the denominator is $\hat{\sigma}^2_\psi + MS_{\text{error}}$. The numerator of partial $\hat{\omega}^2_A$ is $\hat{\sigma}^2_A$ (Equation 7.20), and the denominator is $\hat{\sigma}^2_A + MS_{\text{error}}$.
   b) *Random factor.* The numerator of partial $\hat{\rho}_I$ is $\hat{\sigma}^2_A$ (Equation 7.22), and the denominator is $\hat{\sigma}^2_A + MS_{\text{error}}$.
3. *Dependent samples, nonadditive model.* Because error variance cannot be uniquely estimated apart from that due to a true person × treatment interaction, there is no unbiased definition of partial $\hat{\omega}^2_{\text{effect}}$ or partial $\hat{\rho}_I$. The statistic partial $\hat{\eta}^2$ is an alternative in this case.

For the independent samples analysis in Table 7.2, assuming a fixed factor,

$$MS_A = 20.00,\ MS_{\hat{\psi}1} = 10.00,\ MS_{\hat{\psi}2} = 30.00,\ MS_W = 5.50,\ SS_T = 106.0$$

$$\hat{\eta}^2_A = \hat{\eta}^2_{\psi 1} + \hat{\eta}^2_{\psi 2} = .094 + .283 = .377$$

Using the method of direct calculation outlined in Table 7.4, we compute inferential measures of association based on total variance as follows:

$$\hat{\omega}^2_A = \frac{2\,(20.00 - 5.50)}{106.00 + 5.50} = .260$$

$$\hat{\omega}^2_{\psi_1} = \frac{10.00 - 5.50}{106.00 + 5.50} = .040 \quad \text{and} \quad \hat{\omega}^2_{\psi_2} = \frac{30.00 - 5.50}{106.00 + 5.50} = .220$$

As expected, values of the inferential measures just calculated are each smaller than the corresponding descriptive measure (e.g., $\hat{\eta}^2_A = .377$, $\hat{\omega}^2_A = .260$). Because $\hat{\psi}_1$ and $\hat{\psi}_2$ are orthogonal and $df_A = 2$, it is also true that

$$\hat{\omega}_A^2 = \hat{\omega}_{\psi 1}^2 + \hat{\omega}_{\psi 2}^2 = .040 + .220 = .260$$

Proportions of observed residual variance explained by the contrasts in Table 7.2 are

$$\text{partial } \hat{\eta}_{\psi 1}^2 = .132 \quad \text{and} \quad \text{partial } \hat{\eta}_{\psi 2}^2 = .303$$

The numerator of partial $\hat{\omega}_\psi^2$ for each contrast is calculated with Equation 7.21 as

$$\hat{\sigma}_{\psi 1}^2 = \frac{1}{15}(10.00 - 5.50) = .300 \quad \text{and} \quad \hat{\sigma}_{\psi 2}^2 = \frac{1}{15}(30.00 - 5.50) = 1.633$$

Given $\hat{\sigma}_{\text{error}}^2 = MS_W = 5.50$, next we calculate

$$\text{partial } \hat{\omega}_{\psi 1}^2 = \frac{.300}{.300 + 5.50} = .052 \quad \text{and} \quad \text{partial } \hat{\omega}_{\psi 2}^2 = \frac{1.633}{1.633 + 5.50} = .229$$

Again, each bias-adjusted proportion of explained residual variance is smaller than its observed counterpart for each contrast. Exercise 5 asks you to verify that $\hat{\rho}_I = .345$ assuming a random factor for the results in Table 7.2. Exercise 6 asks you to compute $\hat{\omega}_A^2$, partial $\hat{\omega}_{\psi 1}^2$, and partial $\hat{\omega}_{\psi 2}^2$ for the dependent samples analysis in Table 7.3 for an additive model and a fixed factor.

An alternative to ANOVA-based estimation of variance components in designs with random factors is **maximum likelihood estimation**, which iteratively improves the estimates until statistical criteria are satisfied. It also requires large samples; see Searle, Casella, and McCulloch (1992) for more information. A problem that arises in both ANOVA and maximum likelihood methods is negative variance estimates, which are most likely in small samples or when effect sizes are about zero. Estimates < 0 are usually interpreted as though the value were zero.

Effect Sizes for Power Analysis

Population effect size for fixed factors is represented in some computer tools for power analysis with the **f^2 parameter** (Cohen, 1988), which is related to η^2 as follows:

$$f^2 = \frac{\eta^2}{1 - \eta^2} \tag{7.26}$$

That is, f^2 is the ratio of explained variance over unexplained variance, and thus it is a kind of **population signal-to-noise ratio**. The f^2

parameter can also be expressed for balanced designs and assuming homoscedasticity as

$$f^2 = \frac{1}{a}\sum_{i=1}^{a}\left(\frac{\mu_i - \mu}{\sigma}\right)^2 \tag{7.27}$$

where the expression in parentheses is the contrast between each of the i population means and the grand mean, $\mu_i - \mu$, standardized by the common population standard deviation, σ. Thus, f^2 is the average squared standardized contrast over all populations. Researchers infrequently report sample estimators of f^2 as effect sizes, but see Winer et al. (1991, pp. 126–127) for an example. Steiger (2004) described additional forms of f^2.

Interval Estimation

Measures of association generally have complex distributions, and methods for obtaining approximate confidence intervals are not really amenable to hand calculation. Some of the computer tools described earlier that calculate noncentral confidence intervals for η^2 in two-group designs can also be used in designs with ≥3 independent samples. For example, I used Smithson's (2003) scripts (see footnotes 3–4, Chapter 5) to calculate for the results in Table 7.2 (a) the noncentral 95% confidence interval for η^2_A based on the omnibus effect and (b) the noncentral 95% confidence intervals for partial η^2_ψ based on the two contrasts, given the test statistics for these effects. The input data and results are summarized next:

$$\hat{\eta}^2_A = .377, 95\%\ \text{CI}\ [0, .601], F_A(2, 12) = 3.64$$

$$\text{partial}\ \hat{\eta}^2_{\psi 1} = .132, 95\%\ \text{CI}\ [0, .446], F_{\psi 1}(1, 12) = 1.82$$

$$\text{partial}\ \hat{\eta}^2_{\psi 2} = .313, 95\%\ \text{CI}\ [0, .587], F_{\psi 2}(1, 12) = 5.45$$

Fidler and Thompson (2001) gave a modification of an earlier version of Smithson's (2003) SPSS scripts that constructs noncentral confidence intervals for ω^2. They also reviewed methods to construct confidence intervals for ρ_I in designs with independent samples and random factors. They found that confidence intervals are typically wider (less precise) when the factor is random than when it is fixed. This is one of the costs of generalizing beyond the particular levels randomly selected for study. W. H. Finch and French (2012) compared in computer simulations different methods of interval estimation

for ω^2 in single-factor and two-way designs. No method performed well under conditions of nonnormality, especially for $N < 50$; otherwise, nonparametric bootstrapping with bias correction was generally accurate. They also reported that adding a second factor affected the precision of confidence intervals for the original factor, so interval estimation for one factor was affected by other variables in the design.

The MBESS package for R (Kelley, 2007) can calculate noncentral confidence intervals for f^2 or η^2 in designs with independent samples. Steiger (2004) outlined noncentrality interval estimation for ω^2 and a root-mean-squared standardized effect size related to the f^2 parameter in balanced, completely between-subjects designs. He also described ANOVA methods for testing non-nil hypotheses, such as whether observed effect sizes differ statistically from trivial levels or nontrivial (exceeds a threshold for substantive significance) levels. There is a paucity of computer tools that calculate confidence intervals based on measures of association in correlated designs, but this situation is likely to change.

EFFECT SIZES IN COVARIATE ANALYSES

A **covariate** is a variable that predicts outcome but is ideally unrelated to the independent variable (factor). The variance explained by the covariate is removed, which reduces error variance. With sufficient reduction in error variance, the power of the statistical test for the factor may be higher in ANCOVA than in ANOVA without the covariate. The ANCOVA also yields group means on the dependent variable adjusted for the covariate. These adjustments reflect (a) the pooled within-groups regression of the outcome variable on the covariate and (b) the amount of deviation of group covariate means from the grand mean. If this deviation is slight, there is little adjustment. Otherwise, the adjusted means can be substantially higher or lower than the corresponding unadjusted means.

In experimental designs where cases are randomly assigned to conditions, group means on the covariate vary only by chance. As a consequence, (a) adjusted group means on the outcome variable tend to be similar to the unadjusted means, and (b) it may be only the error term that differs appreciably across ANCOVA and ANOVA results for the same outcome. But groups in nonexperimental designs may differ systematically on the covariate. If so, the covariate is related to both the factor and the dependent variable, and both the ANCOVA error term and the adjusted means can differ substantially from their ANOVA counterparts. But unless the covariate reflects basically all sources of differences between intact groups, the adjusted means may be incorrect. This is why ANCOVA does not cure preexisting group

differences on confounding variables (i.e., covariates); see G. A. Miller and Chapman (2001) for a review.

The technique of ANCOVA has two more assumptions than ANOVA does. One is **homogeneity of regression**, which requires equal within-populations unstandardized regression coefficients for predicting outcome from the covariate. In nonexperimental designs where groups differ systematically on the covariate (and presumably also on other variables related to outcome), the homogeneity of regression assumption is rather likely to be violated. The second assumption is that the covariate is measured without error (its scores are perfectly reliable). Violation of either assumption may lead to inaccurate results. For example, an unreliable covariate in experimental designs causes loss of statistical power and in nonexperimental designs may also cause inaccurate adjustment of the means (Culpepper & Aguinis, 2011). In nonexperimental designs where groups differ systematically, these two extra assumptions are especially likely to be violated.

An alternative to ANCOVA is **propensity score analysis** (PSA). It involves the use of logistic regression to estimate the probability for each case of belonging to different groups, such as treatment versus control, in designs without randomization, given the covariate(s). These probabilities are the propensities, and they can be used to match cases from nonequivalent groups. But PSA offers no magic, because the accuracy of propensities requires that the covariates measure in large part the selection process whereby cases wound up in their respective nonequivalent groups; see Shadish et al. (2001) for more information.

The SPSS raw data file for this example can be downloaded from this book's web page. McWhaw and Abrami (2001) conducted a 30-minute workshop for Grade 11 students about finding main ideas in text. The same students were later randomly assigned to one of two incentive conditions. Students in the extrinsic condition were offered a monetary reward if they found 75% of the main ideas in a text passage, but students in the intrinsic condition were merely encouraged to see the task as a challenge. The outcome variable was the number of main ideas found, and the covariate was the students' grades in school expressed as percentages. Descriptive statistics are summarized in Table 7.5. Observed means on the reading task indicate that the extrinsic group found on average about one more main idea than the intrinsic group, respectively, 3.05 versus 2.08. Adjusted means controlling for grades in the extrinsic and intrinsic groups are, respectively, 3.02 and 2.11. These means are similar to the observed means because the factor and covariate are essentially unrelated in this experimental design.

Reported at the top of Table 7.6 is the ANOVA source table for the data in Table 7.5 except for the covariate. The intrinsic–extrinsic factor in this

TABLE 7.5
Descriptive Statistics on the Outcome Variable and Covariate for Two Learning-Incentive Conditions

Variable	Incentive condition[a]	
	Intrinsic	Extrinsic
School grades (covariate)	74.59 (7.37)[b]	75.13 (10.69)
Main ideas found (outcome)		
Observed	2.08 (2.09)	3.05 (2.42)
Adjusted	2.11[c]	3.02

Note. These data are from K. McWhaw (personal communication, January 23, 2012) and are used with permission. The unstandardized pooled within-groups coefficient for the regression of outcome on the covariate is .117.
[a]Intrinsic condition, $n_1 = 55$; extrinsic condition, $n_2 = 37$. [b]Mean (standard deviation). [c]Mean.

analysis explains about 4.5% of total variance on the reading task ($\hat{\eta}^2 = .045$). Other key ANOVA results are

$$MS_W = 4.96,\ F(1, 90) = 4.21,\ p = .043$$

Presented in the middle of Table 7.6 is the traditional ANCOVA source table. The sums of squares are Type III, which reflect the unique explanatory

TABLE 7.6
Analysis of Variance (ANOVA) and Analysis of Covariance (ANCOVA) Results for the Data in Table 7.5

Source	*SS*	*df*	*MS*	*F*	$\hat{\eta}^2$
ANOVA					
Between (incentive)	20.91	1	20.91	4.21[b]	.045
Within (error)	446.77	90	4.96		
Total	467.68	91			
Traditional ANCOVA[a]					
Total effects	117.05	2	58.52	14.86[c]	**.250**
Covariate (grades)	96.14	1	96.14	24.40[c]	.206
Between (incentive)	18.25	1	18.25	**4.63**[d]	**.039**
Within (error)	350.63	89	3.94		
Total	467.68	91			

ANCOVA-as-regression						
Step	Predictors	R^2	R^2 change	*F* change	df_1	df_2
1	Grades	.211	.211	24.11[c]	1	90
2	Grades, incentive	**.250**	**.039**	**4.63**[d]	1	89

Note. Entries in boldface emphasize common results across traditional ANCOVA and ANCOVA-as-regression analyses of the same data and are discussed in the text.
[a]Type III sums of squares. [b]$p = .043$. [c]$p < .001$. [d]$p = .034$.

power of individual predictors. The factor and covariate together explain about 25.0% of the total variance in reading scores ($\hat{\eta}^2 = .250$). Of the two, the factor and covariate each uniquely explain, respectively, about 3.9% and 20.6% of total variance. The prime symbol "′" designates results adjusted for the covariate, and the results of the *F* test presented next are for incentive condition:

$$MS'_W = 3.94, F(1, 89) = 4.63, \ p = .034$$

The ANCOVA error term is related to the ANOVA error term as follows:

$$MS'_W = MS_W \left(\frac{df_W \left(1 - r^2_{\text{pool}}\right)}{df_W - 1} \right) \tag{7.28}$$

where r^2_{pool} is the squared pooled within-groups correlation between the covariate and the dependent variable. For this example, where $df_W = 90$ and $r_{\text{pool}} = .464$,

$$MS'_W = 4.96 \left(\frac{90\left(1 - .464^2\right)}{90 - 1} \right) = 3.94$$

Results at the bottom of Table 7.6 represent a third perspective. They are from a hierarchical multiple regression analysis where grade (covariate) is the sole predictor of reading at step 1 and incentive condition is entered as the second predictor at step 2. At step 1, the result $R^2_1 = .212$ is just the squared Pearson correlation between grades and reading. At step 2, $R^2_2 = .250$ is the squared multiple correlation where the predictors are grades and incentive condition. This result (.250) is presented in boldface in the table to emphasize its equivalence with $\hat{\eta}^2 = .250$ for the total effects in the traditional ANCOVA. The statistic $F\ (1, 89) = 4.63$ reported in boldface for the regression results tests whether

$$R^2_2 - R^2_1 = .250 - 210 = .039$$

differs statistically from zero. The *F* statistic and R^2 change values just stated are each identical to their counterparts in the traditional ANCOVA results (e.g., $\hat{\eta}^2 = .039$ for incentive condition; see Table 7.6). Thus, ANCOVA from a regression perspective is nothing more than a test of the incremental valid ity of the factor over the covariate in predicting outcome.

Let us now estimate the magnitude of the standardized contrast between the extrinsic and intrinsic conditions on the reading task. There are two possibilities for the numerator: the contrast of the two unadjusted means, $M_1 - M_2$, or means adjusted for the covariate, $M'_1 - M'_2$. There are also at least

two possibilities for the standardizer: a standard deviation in the metric of the original scores or a standard deviation in the metric of the adjusted scores. This makes a total of at least four possible forms of a standardized mean contrast for these ANCOVA results. In an experimental design, there should be little difference between $M_1 - M_2$ and $M'_1 - M'_2$, which is true for this example (see Table 7.5). Unless the highly restrictive assumptions described earlier hold in a nonexperimental design, the value of $M'_1 - M'_2$ may be inaccurate, so $M_1 - M_2$ as the numerator may be the best option. The most general choice for the standardizer in the metric of the original scores is the square root of the ANOVA error term, MS_W (i.e., the effect size is d_{pool} when there are two groups). This term reflects variability due to the covariate, but the ANCOVA error term MS'_W holds the covariate constant (statistically controls for it).

Cortina and Nouri (2000) suggested that if the covariate varies naturally in the population to which the results should generalize, selection of the square root of MS_W as the standardizer may be the best choice. This would be true even if MS'_W is substantially less than MS_W. The grades covariate for the present example varies naturally among students, so the standardized mean difference for the comparison of the extrinsic and intrinsic conditions, given $MS_W = 4.96$ and the group descriptive statistics in Table 7.5, is calculated as

$$d_{pool} = \frac{3.05 - 2.08}{\sqrt{4.96}} = .44$$

That is, the students in the extrinsic reward condition outperformed their peers in the intrinsic reward condition on the reading task by about .44 standard deviations. See Colliver and Markwell (2006) and Rutherford (2011, Chapter 4) for more information.

RESEARCH EXAMPLES

Raw data files for the two examples considered next are not available, but you can download from this book's web page SPSS syntax that analyzes the summary statistics for each.

Relative Cognitive Status of Recreational Ecstasy Users

Ecstasy (MDMA) and related stimulant compounds (MDA, MDEA) make up a class of recreational drugs used by some adolescents and young adults. Results of animal studies from the early 1990s indicated neurotoxic effects of high doses, but whether lower doses of ecstasy impair cognitive functioning in humans was not well understood during the ensuing decade.

Gouzoulis-Mayfrank et al. (2000) recruited 28 ecstasy users who also smoked cannabis and compared their performance on tasks of attention, learning, and abstract thinking with that of two different control groups of similar age (M = 23 years) and educational backgrounds, cannabis-only users and nonusers of either substance. The ecstasy users agreed to abstain from the drug for at least seven days, which was confirmed by urine analysis on the day of testing.

Gouzoulis-Mayfrank et al. (2000) found many statistically significant differences but did not report effect sizes. Presented in Table 7.7 are representative results for five tasks where d_{pool} is reported for each pairwise comparison. The three groups performed about the same on an attention task of simple reaction time. On a more demanding selective attention task and also on measures of learning and abstract thinking, ecstasy users performed worse than both other groups by about .80 standard deviations. Sizes of differences between cannabis users and nonusers of either substance were generally smaller—about .10 standard deviations—except on the verbal learning task, where the nonusers had an advantage of about .40 standard deviations. Exercise 7 asks you to calculate $\hat{\eta}^2$ for the two attention tasks in Table 7.7.

TABLE 7.7
Cognitive Test Scores for Ecstasy (MDMA) Users, Cannabis Users, and Nonusers

	User group[a]				d_{with} for pairwise contrasts		
Task	1 Ecstasy	2 Cannabis	3 Nonuser	F (2, 81)	1 vs. 2	1 vs. 3	2 vs. 3
Attention[b]							
Simple	218.9[e] (28.2)	221.1 (26.3)	218.7 (27.5)	.07[f]	−.08	.01	.09
Selective	532.0 (65.4)	484.4 (57.9)	478.6 (48.4)	7.23[g]	.83	.93	.10
Learning and abstract thinking							
Verbal[c]	4.46 (.79)	3.71 (1.15)	3.29 (1.12)	9.22[h]	.73	1.13	.41
Visual[c]	4.61 (.96)	4.00 (1.41)	4.11 (1.13)	2.12[i]	.52	.42	−.09
Abstract thinking[d]	25.96 (4.10)	29.46 (4.19)	29.50 (3.64)	7.29[g]	−.88	−.89	−.01

Note. From "Impaired Cognitive Performance in Drug Free Users of Recreational Ecstasy (MDMA)," by E. Gouzoulis-Mayfrank, J. Daumann, F. Tuchtenhagen, S. Pelz, S. Becker, H.-J. Kunert, B. Fimm, and H. Sass, 2000. *Journal of Neurology, Neurosurgery & Psychiatry, 68,* p. 723. Copyright 2000 by BMJ Publishing Group Limited. Adapted with permission.

[a] $n = 28$ for all groups. [b] Scores are in milliseconds; higher scores indicate worse performance. [c] Number of trials; higher scores indicate worse performance. [d] Higher scores indicate better performance. [e] Mean (standard deviation). [f] $p = .933$. [g] $p = .001$. [h] $p < .001$. [i] $p = .127$.

Gouzoulis-Mayfrank et al. (2000) discussed the possibility that preexisting differences in cognitive ability or neurological status may explain their findings. There has been subsequent evidence that, compared with nonusers, chronic ecstasy users may be susceptible to hippocampal damage (den Hollander et al., 2012) and aberrant visual cortical excitability in transcranial magnetic stimulation studies (Oliveri & Calvo, 2003). Turning back to the original example, let us consider the practical significance of group differences in cognitive functioning that are about .80 standard deviations in magnitude. Assuming normal distributions and homoscedasticity, it is expected that the typical nonecstasy user will outperform about 80% of the ecstasy users ($U_3 = .79$; see Table 5.4). It is also expected that ecstasy users will be underrepresented by a factor of about 3½ among those young adults who are more than one standard deviation above the mean in the combined distribution for learning and abstract thinking ability (RTR = 3.57).

Analysis of Learning Curve Data

Kanfer and Ackerman (1989) administered to 137 U.S. Air Force personnel a computerized air traffic controller task, presented over six 10-minute trials, where the outcome variable was the number of successful landings. Summarized in Table 7.8 are the means, standard deviations, and correlations across all trials. The last show a typical pattern for learning data in that correlations between adjacent trials are higher than those between nonadjacent trials. This pattern may violate the sphericity assumption of statistical tests for comparing dependent means for equality. Accordingly, *p* values for effects with more than a single degree of freedom are based on the Geisser–Greenhouse conservative test. Task means over trials exhibit both linear and quadratic trends, which are apparent in Figure 7.1.

Reported in Table 7.9 are the results of repeated measures analyses of variance of the omnibus trials effect, the linear and quadratic trends, and all other higher order trends combined (cubic, quartic, quintic; i.e., respectively, 2, 3, or 4 bends in the curve). Effect size is estimated with $\hat{\omega}^2$ for an additive model. All effects have *p* values <.001, but their magnitudes are clearly different. The omnibus trials effect explains about 43% of the total variance corrected for capitalization on chance. Because the linear trend itself accounts for about 38% of the total variance, it is plain to see that this polynomial is the most important aspect of the learning curve. The quadratic trend explains an additional 5% of the total variance, and all higher order trends together explain <.1% of the total variance. The orthogonal linear and quadratic trends together thus account for virtually all of the explained variance.

In their analysis of the same learning curve data, Kanfer and Ackerman (1989) reduced unexplained variance even more by incorporating a cognitive

TABLE 7.8
Descriptive Statistics for a Computerized Air Traffic Controller Task

	Trial					
	1	2	3	4	5	6
M	11.77	21.39	27.50	31.02	32.58	34.20
s	7.60	8.44	8.95	9.21	9.49	9.62
r	1.00					
	.77	1.00				
	.59	.81	1.00			
	.50	.72	.89	1.00		
	.48	.69	.84	.91	1.00	
	.46	.68	.80	.88	.93	1.00

Note. The group size is $n = 137$, and the correlation matrix is in lower-diagonal form. Adapted from "Models for Learning Data," by M. W. Browne and S. H. C. Du Toit, 1991. In L. M. Collins and J. L. Horn (Eds.), *Best Methods for Analysis of Change,* p. 49, Washington, DC: American Psychological Association. Copyright 1991 by the American Psychological Association.

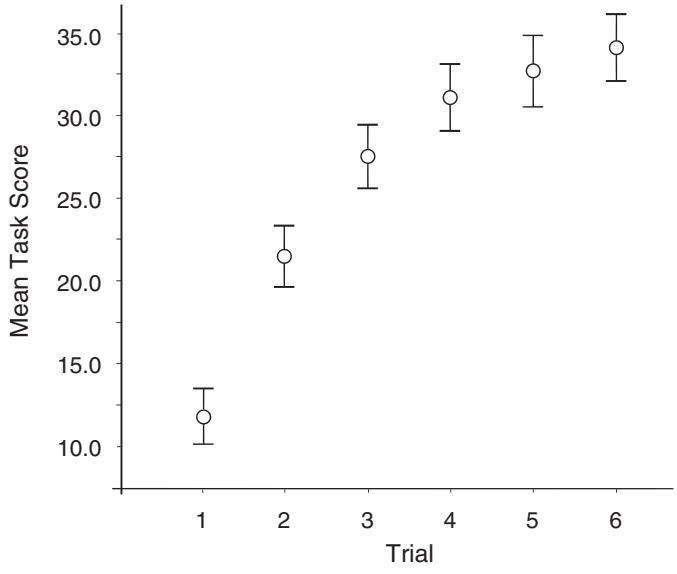

Figure 7.1. Means and 95% confidence intervals for μ for the learning trial data in Table 7.8.

TABLE 7.9
Analysis of the Learning Curve Data in Table 7.8

Source	*SS*			*df*	*MS*	*F*	$\hat{\omega}^2$
Between (trials)	49,419.08			5	9,883.82	470.88[a]	.43
Linear			43,590.62	1	43,590.62	742.84[a]	.38
Quadratic			5,524.43	1	5,524.43	269.58[a]	.05
All other trends			304.04	3	101.35	11.79[a]	<.001
Within	64,807.36			816	79.42		
Subjects (*S*)		50,531.23		136	371.55		
Residual (trials)		14,276.13		680	20.99		
Residual (linear)			7,980.64	136	58.68		
Residual (quadratic)			2,787.00	136	20.49		
Residual (all other)			3,508.49	408	8.60		
Total		114,226.44		821			

Note. The contrast weights for linear are (–5, –3, –1, 1, 3, 5) and those for quadratic are (5, –1, –4, –4, –1, 5).
[a]$p < .001$.

ability test as a predictor of learning in addition to the trials factor. This approach was elaborated by Browne and Du Toit (1991), who specified and tested various latent variable models of Kanfer and Ackerman's (1989) data. These models attempted to predict not only the mean level of performance over trials but also the shapes and variabilities of the learning curves of individual participants and whether parameters of these curves covary with overall cognitive ability. In this approach, the proportions of explained variance were >50%, which is better than the results reported in Table 7.9. Browne and Du Toit's (1991) analyses highlight the potential value of a model-fitting approach for analyzing learning curve data.

CONCLUSION

It is often more informative to analyze contrasts than the omnibus effect in single-factor designs. The most general standardized contrast is d_{with}, where the standardizer is the square root of MS_W, the pooled within-conditions variance. There are also robust standardized contrasts for data sets with outliers, nonnormal distributions, or heteroscedasticity. Descriptive measures of association are all forms of the correlation ratio $\hat{\eta}^2$. The inferential measure of association $\hat{\omega}^2$ corrects for positive bias in $\hat{\eta}^2$ but requires balanced designs. Both statistics just

mentioned are proportions of total variance explained by an effect. There are versions of these effect sizes that control for other effects, namely, partial $\hat{\eta}^2$ and partial $\hat{\omega}^2$, but they are proportions of explained residualized variance. The intraclass correlation $\hat{\rho}_I$ is an appropriate measure of association for omnibus effects in balanced designs with random factors. How to estimate effect sizes in designs with multiple factors and continuous outcomes is considered in the next chapter.

LEARN MORE

Miller and Chapman (2001) consider misuses of ANCOVA, and Rosenthal et al. (2000) describe many examples of contrast analysis. Steiger (2004) reviews effect size confidence intervals and tests of non-nil hypotheses in ANOVA for balanced designs with fixed factors.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110,* 40–48. doi:10.1037/0021-843X.110.1.40

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research*. New York, NY: Cambridge University Press.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9,* 164–182. doi: 10.1037/1082-989X.9.2.164

EXERCISES

1. Calculate the approximate 95% confidence interval for $\delta_{\psi 2}$ for the independent samples analysis in Table 7.2.
2. Calculate the approximate 95% confidence interval for $\delta_{\psi 2}$ for the dependent samples analysis in Table 7.3.
3. Calculate and interpret $r_{\hat{\psi}}$ and partial $r_{\hat{\psi}}$ for $\hat{\psi}_2$ in Table 7.2.
4. Calculate and interpret partial $\hat{\eta}^2$ for each contrast in Table 7.3.
5. Verify that $\hat{\rho}_I = .345$ for the independent samples analysis in Table 7.2 for a random factor.
6. Calculate and interpret $\hat{\omega}^2_A$, partial $\hat{\omega}^2_{\psi_1}$, and partial $\hat{\omega}^2_{\psi_2}$ for the dependent samples analysis in Table 7.3 assuming a fixed factor and an additive model.
7. Calculate $\hat{\eta}^2$ for the omnibus effects for each of the two attention tasks in Table 7.7.

8

MULTIFACTOR DESIGNS

> Fools ignore complexity. Pragmatists suffer it. Some can avoid it. Geniuses remove it.
>
> —Alan J. Perlis (1982, p. 10)

Designs with multiple factors and continuous outcomes require special considerations for effect size estimation. This is because some methods for single-factor designs may not give the best results in multifactor designs, and ignoring this problem may introduce variation across studies because of statistical artifacts rather than real differences in effect sizes. Described in the next two sections are various kinds of multifactor designs and the basic logic of factorial ANOVA. Effect size estimation with standardized contrasts and measures of association in factorial designs are then considered. Exercises for this chapter help you to consolidate skills about effect size estimation in multifactor designs.

TYPES OF MULTIFACTOR DESIGNS

The most common type of multifactor design is a **factorial design** where every pair of factors is **crossed**, which means that the levels of each factor are studied in all combinations with levels of other factors. If factor *A* has

DOI: 10.1037/14136-008
Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, Second Edition, by R. B. Kline

$a = 2$ levels and factor B has $b = 3$ levels, for instance, a full $A \times B$ factorial design would have a total of 2×3, or 6, conditions, including

$$A_1B_1, A_1B_2, A_1B_3, A_2B_1, A_2B_2, \text{and } A_2B_3$$

The conditions are studied with independent samples in a **completely between-subjects factorial design**. That is, the subjects factor is **nested** under these combinations, which means that there is a separate group for each one. If cases are randomly assigned to conditions, the design is a **randomized groups factorial design**, but if at least one factor is an individual difference (nonexperimental) variable and the rest are manipulated (experimental) variables, it is a **randomized blocks design**. An example is where samples of women and men (A) are randomly assigned to one of three different incentive conditions (B) in a 2×3 randomized blocks design.

A **mixed within-subjects factorial design**—also called a **split-plot design**—has both between-subjects and within-subjects factors. An example is where samples of women and men are each tested twice on the same outcome, such as before and after an intervention. In this case, the subjects factor is nested under gender, but it is crossed with measurement occasion because every person is tested twice in this 2×2 design. In a **completely within-subjects factorial design**, each case in a single sample is tested under all combinations of two or more crossed factors. That is, the subjects factor is crossed with all independent variables.

This chapter deals with factorial designs, but their basic principles generalize to the variations mentioned next. In a **hierarchical design**, at least one factor is nested under another. Suppose factor A is drug versus placebo and factor B represents patient groups from four separate clinics. Patients from the first two clinics are randomly assigned to receive the drug, and the rest get a placebo. The four combinations are

$$A_1B_1, A_1B_2, A_2B_3, \text{and } A_2B_4$$

which are not all possible permutations (8) due to nesting of B under A. Nested factors are typically considered random. Levels of factors in **fractional (partial, incomplete) factorial designs** may not be studied in every combination with levels of other factors. This reduces the number of conditions, but certain main and interaction effects may be confounded. A **Latin square design**, which also counterbalances order effects for repeated measures factors, is perhaps the best known example; see Kirk (2012, Chapters 11–16) for more information.

FACTORIAL ANALYSIS OF VARIANCE

To understand effect size estimation in factorial designs, you need to know about factorial ANOVA. This is a broad topic, and its coverage in applied textbooks is often lengthy. These facts preclude a detailed review, so this presentation emphasizes common principles of factorial ANOVA that also inform effect size estimation. It deals first with concepts in balanced two-way designs and then extends them to larger or unbalanced designs (i.e., with unequal cell sizes). It also encourages you to pay more attention to the sums of squares and mean squares in a factorial ANOVA source table than to *F* ratios and *p* values. See Kirk (2012); Myers, Well, and Lorch (2010); Rutherford (2011); or Winer et al. (1991) for more information.

Basic Distinctions

Factorial ANOVA models are distinguished by (a) whether the factors are between-subjects versus within-subjects, (b) whether the factors are fixed versus random, and (c) whether the cell sizes are equal or unequal. The first two distinctions affect the denominators (error terms) of *F* tests and their statistical assumptions, but they do not influence the numerators. That is, effect sums of squares and mean squares are derived the same way regardless of whether the factor is between-subjects versus within-subjects or fixed versus random. In balanced factorial designs (i.e., all cell sizes are equal), the main and interaction effects are independent, which means that they can occur in any pattern. Accordingly, balanced factorial designs are called **orthogonal designs**. Independence of effects in such designs simplifies their analysis, but many real-world factorial designs are not balanced, especially in applied research (e.g., Keselman et al., 1998).

We must differentiate between unbalanced factorial designs with proportional versus disproportional cell sizes. Consider the two 2×3 factorial layouts represented in the table that follows, where the numbers are cell sizes (n):

	B_1	B_2	B_3
A_1	5	10	20
A_2	10	20	40

	B_1	B_2	B_3
A_1	5	20	10
A_2	10	10	50

The cell sizes in the upper left matrix are proportional because the ratios of their relative values are constant across all rows (1:2:4) and columns (1:2). The cell sizes are disproportional in the upper right matrix because their relative ratios are not constant over rows or columns. This distinction is crucial because factorial designs with unequal-but-proportional cell sizes can be analyzed as orthogonal (balanced) designs (e.g., Winer et al., 1991, pp. 402–404). This is true because equal cell sizes are a special case of proportional cell sizes.

Disproportional cell sizes cause the factors to be correlated, which implies that their main effects overlap. The greater the departure from proportional cell sizes, the greater is this overlap. Thus, factorial designs with disproportional cell sizes are called **nonorthogonal designs**, and they require special methods that try to disentangle correlated effects. Unfortunately, there are different methods for nonorthogonal designs, and it is not always clear which one is best for a particular study. The choice among alternative methods for nonorthogonal designs affects both statistical tests and effect size estimation for main effects.

Factorial designs tend to have equal or at least proportional cell sizes if all or all but one of the factors is experimental. If at least two factors are nonexperimental and cases are drawn from a population where these variables are correlated, the cell sizes may be disproportional. But this nonorthogonality may actually be an asset if it reflects disproportional population group sizes. It may be possible to force equal cell sizes by dropping cases from the larger cells or recruiting additional participants for the smaller cells, but the resulting **pseudo-orthogonal design** may not be representative. This is why nonorthogonal designs are sometimes intentional—that is, based on a specific sampling plan.

A critical issue is missing data. For example, a nonorthogonal design may arise due to randomly missing data from a factorial design intended as balanced, such as when equipment fails and scores are not recorded. A handful of missing observations is probably of no great concern, such as $n = 50$ in all cells except for $n = 47$ in a single cell as a result of three randomly missing scores. A more serious problem occurs when nonorthogonal designs result from nonrandomly missing data, such as when higher proportions of participants drop out of the study under one combination of treatments than others. Systematic data loss may cause a bias: Patients who withdrew may differ from those who remain, and the results may not generalize to the intended population(s). There is no simple statistical fix for bias because of nonrandomly missing data. About all that can be done is to understand the nature of the data loss and then accordingly qualify your interpretation of the results. McKnight, McKnight, Sidani, and Figueredo (2007) reviewed contemporary options for dealing with missing data.

Basic Sources of Variability

Just as in single-factor designs, there are two basic sources of variability in factorial designs, between and within conditions (cells). In both kinds of ANOVA, differences in cell means (between variation) correspond to effects of the factors on the outcome variable. If all cell means are equal, there are no effects; otherwise, the factors may affect the dependent variable to some extent. Variation within the cells is not attributed to the factors. The estimate of overall (pooled) within-cells variation has the same general form in any type of ANOVA, or $MS_W = SS_W/df_W$, which is the weighted average of the cell variances. For example, the following equation generates MS_W in any factorial design with two factors, *A* and *B*, where none of the cells are empty:

$$MS_W = \frac{SS_W}{df_W} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} df_{ij}(s_{ij}^2)}{\sum_{i=1}^{a}\sum_{j=1}^{b} df_{ij}} \tag{8.1}$$

where df_{ij} and s_{ij}^2 are, respectively, the degrees of freedom ($n_{ij} - 1$) and variance of the cell at the *i*th level of *A* and the *j*th level of *B*. Only in a balanced design can MS_W also be computed as the simple arithmetic average of the cell variances.

The between-conditions variance in a single-factor design, MS_A, reflects the effects of factor *A*, sampling error, and cell size (Equation 3.10). In a factorial design, the overall between-conditions variance reflects the main and interactive effects of all factors, sampling error, and cell size. For example, the between-cells variance in a two-way design is designated below as

$$MS_{A,B,AB} = \frac{SS_{A,B,AB}}{df_{A,B,AB}} \tag{8.2}$$

where the subscript indicates the main and interaction effects analyzed together (total effects), and the degrees of freedom equal the number of cells minus one, or $ab - 1$. It is only in balanced two-way designs that the sum of squares for the total effects can be computed directly as

$$SS_{A,B,AB} = \sum_{i=1}^{a}\sum_{j=1}^{b} n\left(M_{ij} - M_T\right)^2 \tag{8.3}$$

where *n* is the size of all cells, M_{ij} is the mean for the cell at the *i*th level of *A* and the *j*th level of *B*, and M_T is the grand mean for the whole design. That is,

$SS_{A, B, AB}$ is estimated as the total of the squared deviations of each cell mean from the grand mean weighted by the cell size. It is also only in balanced designs that the total effects sum of squares can be broken down into unique (orthogonal) and additive values for the individual effects:

$$SS_{A, B, AB} = SS_A + SS_B + SS_{AB} \tag{8.4}$$

This relation can also be expressed in terms of the correlation ratio in balanced designs:

$$\hat{\eta}^2_{A, B, AB} = \hat{\eta}^2_A + \hat{\eta}^2_B + \hat{\eta}^2_{AB} \tag{8.5}$$

(Recall that the general form of $\hat{\eta}^2_{\text{effect}}$ is SS_{effect}/SS_T.) Equations 8.4 and 8.5 define effect orthogonality in two-way designs. Orthogonality in factorial designs of any size means that the main and interaction effects can appear in any combination. This means that observing one type of effect, such as a main effect of factor *A*, says nothing about whether any other effect will be found, such as a main effect of *B* or the interaction effect *AB*.

EFFECTS IN BALANCED TWO-WAY DESIGNS

A balanced design where factor *A* has $a = 2$ levels and factor *B* has $b = 3$ levels is represented in Table 8.1. Cell means and variances, marginal (row or column) means, and the grand mean are shown. Because the cell sizes are equal, the marginal means are just the arithmetic averages of the corresponding row or column cell means. Each marginal mean can also be computed as the average of the individual scores (not shown in the table) in the corresponding row or column. That is, the marginal means for each factor are calculated by

TABLE 8.1
General Descriptive Statistics for a Balanced 2×3 Factorial Design

	B_1	B_2	B_3	Row means
A_1	M_{11} (s^2_{11})	M_{12} (s^2_{12})	M_{13} (s^2_{13})	M_{A1}
A_2	M_{21} (s^2_{21})	M_{22} (s^2_{22})	M_{23} (s^2_{23})	M_{A2}
Column means	M_{B1}	M_{B2}	M_{B3}	M_T

Note. The size of all cells is *n*.

collapsing across the levels of the other factor. The grand mean for the whole design is the arithmetic average of all six cell means. It can also be computed as the average of the row or column means or as the average of the *abn* individual scores.

Main Effects and Main Comparisons

Conceptual equations for sample main and interaction sums of squares in balanced two-way designs are presented in Table 8.2. A **main effect** is estimated by the differences among the observed marginal means for the same factor, and the sample sum of squares for that effect is the total of the weighted squared deviations of the associated marginal means from the grand mean. For example, if $M_{A1} = M_{A2}$ in Table 8.1, the estimated main effect of *A* is zero and $SS_A = 0$; otherwise, $SS_A > 0$, as is the estimated main effect of this factor. The main effect of *B* is estimated in a similar way but concerns weighted variation of the marginal means M_{B1}, M_{B2}, and M_{B3} around M_T. Because main effects are estimated by collapsing over the levels of the other factor, they are single-factor effects.

The main effect of *A* in Table 8.1 is a contrast because $df_A = 1$, but the *B* main effect is omnibus because $df_B = 2$ (i.e., > 1). This implies that up to two orthogonal contrasts could be specified among the levels of this factor. Keppel and Wickens (2004) referred to such contrasts as **main comparisons**. Such contrasts can be either planned or unplanned, but analyzing main comparisons would make sense only if the magnitude of the overall main effect were appreciable and the magnitude of the interaction effect were negligible.

TABLE 8.2
Equations for Main and Interaction Effect Sums of Squares in Balanced Two-Way Factorial Designs

Source	*SS*	*df*
A	$\sum_{i=1}^{a} bn\,(M_{A_i} - M_T)^2$	$a - 1$
B	$\sum_{j=1}^{b} an\,(M_{B_j} - M_T)^2$	$b - 1$
AB	$\sum_{i=1}^{a}\sum_{j=1}^{b} n\,[M_{ij} - (M_{A_i} - M_T) - (M_{B_j} - M_T) - M_T]^2$	$(a - 1)\,(b - 1)$
	$= \sum_{i=1}^{a}\sum_{j=1}^{b} n\,(M_{ij} - M_{A_i} - M_{B_j} + M_T)^2$	

Note. The size of all cells is *n*.

Interaction Effects, Simple Effects, and Simple Comparisons

An **interaction effect** can be understood in several ways. It is a combined or joint effect of the factors on the outcome variable above and beyond their main effects. It is also a conditional effect that, if present, says that effects of each factor on outcome change over the levels of the other factor and vice versa (i.e., interaction is symmetrical).[1] Interaction effects are also called **moderator effects**, and the factors involved in them are **moderator variables**. Both terms emphasize the fact that each factor's relation with outcome depends on the other factor when there is interaction. Do not confuse a moderator effect with a **mediator effect**, which refers to the indirect effect of one variable on another through a third (mediator) variable. Mediator effects can be estimated in structural equation modeling and meta-analysis but not in the ANOVA models discussed in this chapter; see Kline (2010) for more information.

Sums of squares for the two-way interaction, *AB*, reflect variability of cell means around the grand mean controlling for main effects (if any) and weighted by cell size (see Table 8.2). The pattern of this variation is not exactly predictable from that among marginal means when there is interaction. That is, interaction is related to sets of conditional **simple effects (simple main effects)**. They correspond to cell means in the same row or column, and there are as many simple effects of each factor as there are levels of the other factor. For example, there are two estimated simple effects of factor *B* represented in Table 8.1. One is the simple effect of *B* at A_1, and it corresponds to the three cell means in the first row, M_{11}, M_{12}, and M_{13}. If any two of these means are different, the estimate of the *B* at A_1 simple effect is not zero. The other estimated simple effect of this factor, *B* at A_2, corresponds to the cell means in the second row, M_{21}, M_{22}, and M_{23}. Because $df = 2$ for each simple effect of factor *B*, they are omnibus.

Simple comparisons are contrasts within omnibus simple effects. They are analogous to main comparisons, but simple comparisons concern rows or columns of cell means, not marginal means. Estimated simple effects of factor *A* in Table 8.1 correspond to the pair of cell means in each of the three columns, such as M_{11} versus M_{21} for the simple effect of *A* at B_1. Because $df = 1$ for each of the three simple effects of *A* at *B*, they are also simple comparisons. But the two simple effects of factor *B* in Table 8.1 each concern three cell means, such as M_{21}, M_{22}, and M_{23} for the simple effect of *B* at A_2. Thus, they are omnibus effects, each with 2 degrees of freedom. The contrast of M_{21} with M_{22} in Table 8.1 is an example of a simple comparison within the omnibus simple effect of *B* at A_2.

[1]A common but incorrect description of interaction is that "the factors affect each other." Factors may affect the dependent variable individually (main effects) or jointly (interaction), but factors do not affect each other in factorial designs.

Sums of squares for simple effects have the same general form as those for main effects (see Table 8.2) except that the former are the total of the weighted squared deviations of row or column cell means from the corresponding marginal mean, not the grand mean. It is true in balanced two-way designs that

$$\sum_{j=1}^{b} SS_{A \text{ at } B_j} = SS_A + SS_{AB} \quad \text{and} \quad \sum_{i=1}^{a} SS_{B \text{ at } A_i} = SS_B + SS_{AB} \tag{8.6}$$

In words, the total sum of squares for all simple effects of each factor equals the total sum of squares for the main effect of that factor and the interaction. When all simple effects of a factor are analyzed, it is actually the main and interactive effects of that factor that are analyzed. Given their overlap in sums of squares, it is usually not necessary to analyze both sets of simple effects, *A* at *B* and *B* at *A*. The choice between them should be made on a rational basis, depending on the perspective from which the researcher wishes to describe interaction.

An **ordinal interaction** occurs when simple effects vary in magnitude but not in direction. Look at the cell means for the 2 (drug) × 2 (gender) layout in the left side of Table 8.3, where higher scores indicate a better result. The interaction is ordinal because (a) women respond better to both drugs, but the size of this effect is greater for drug 2 than drug 1 (gender at drug simple effects). Also, (b) mean response is always better for drug 2 than drug 1, but this is even more true for women than for men (drug at gender simple effects). Both sets of simple effects just mentioned vary in magnitude but do not change direction.

The cell means in the right side of Table 8.3 indicate a **disordinal (crossover) interaction** where at least one set of simple effects reverses direction. These results indicate that drug 2 is better for women, but just the opposite is true for men. That is, simple effects of drug change direction for women

TABLE 8.3
Cell Means and Marginal Means for Two-Way Designs With Ordinal Versus Disordinal Interaction

	Ordinal interaction				Disordinal interaction		
	Drug 1	Drug 2			Drug 1	Drug 2	
Women	45.00	70.00	57.50	Women	60.00	70.00	65.00
Men	25.00	30.00	27.50	Men	25.00	15.00	20.00
	35.00	50.00	42.50		42.50	42.50	42.50

Note. The size of all cells is *n*.

versus men. The same data also indicate that simple effects of gender do not reverse because women respond better than men under both drugs, but the whole interaction is nevertheless disordinal.

Disordinal interaction is also indicated whenever lines that represent simple effects cross in graphical representations, but this may depend on how cell means are plotted. Presented in Figure 8.1(a) are two line graphics for cell means from the left side of Table 8.3. The graphics differ only in their representation of gender versus drug on the abscissa (*x*-axis). Lines for both sets of simple effects are not parallel, which indicates interaction, but they do

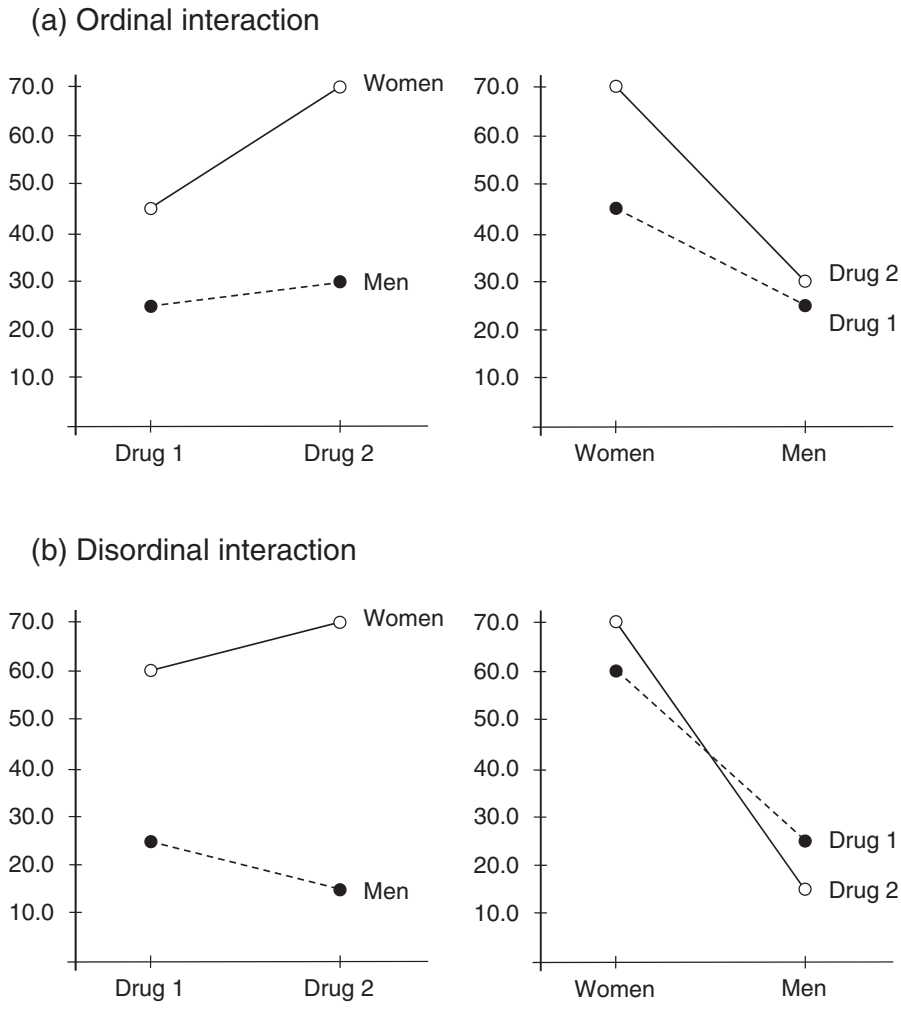

Figure 8.1. Cell mean plots for the data in Table 8.3 for (a) ordinal interaction and (b) disordinal interaction.

not cross, so the interaction is ordinal. Presented in Figure 8.1(b) are the line graphics for the cell means in the right side of Table 8.3, for which the interaction is disordinal. This fact is apparent by the crossing of the lines for at least one set of simple effects, those of drug for women versus men; see the right side of Figure 8.1(b). Exercise 1 asks you to calculate effects sum of squares given the means in the left side of Table 8.3 for $n = 10$ and $MS_W = 125.00$.

Stevens (1999) noted that ordinal interactions can arise from floor or ceiling effects in measurement. A ceiling effect occurs when cases with somewhat high versus very high levels on the construct cannot be distinguished by the outcome measure, which results in underestimation of some cell means. Floor effects imply the opposite: Certain means may be overestimated if the outcome measure cannot distinguish among cases with the lowest levels on the construct. Ceiling or floor effects basically convert interval data to ordinal data. Although ordinal interaction sometimes occurs due to measurement artifacts, disordinal interaction cannot be explained by such artifacts.

Interaction Contrasts and Interaction Trends

An interaction contrast is specified by a matrix of coefficients the same size as the original design that are **doubly centered**, which means that the coefficients sum to zero in every row and column. This property makes the resulting single-*df* interaction effect independent of the main effects. The weights for a two-way interaction contrast can be assigned directly or taken as the product of the corresponding weights of two single-factor comparisons, one for each factor. If the interaction contrast should be interpreted as the difference between a pair of simple comparisons (i.e., mean difference scaling), the sum of the absolute values of the coefficients must be 4.0 (Bird, 2002). This can be accomplished by selecting coefficients for the comparison on each factor that are a standard set (i.e., their absolute values sum to 2.0) and taking their corresponding products.

In a 2×2 design where all effects are contrasts, weights that define the interaction as the difference between two simple effects are presented in the cells of the leftmost matrix:

	B_1	B_2
A_1	1	−1
A_2	−1	1

	B_1	B_2
A_1	M_{11}	M_{12}
A_2	M_{21}	M_{22}

These weights are doubly centered, and the sum of their absolute values is 4.0. We can get the same set of weights for this interaction contrast by taking the corresponding products of the weights (1, –1) for factor *A* and the weights (1, –1) for factor *B*. After applying these weights to the corresponding cell means in the above rightmost matrix, we get

$$\hat{\psi}_{AB} = M_{11} - M_{12} - M_{21} + M_{22} \tag{8.7}$$

Rearranging the terms shows that $\hat{\psi}_{AB}$ equals (a) the difference between the two simple effects of *A* and (b) the difference between the two simple effects of *B*:

$$\begin{aligned}\hat{\psi}_{AB} &= \hat{\psi}_{\text{A at B1}} - \hat{\psi}_{\text{A at B2}} = (M_{11} - M_{21}) - (M_{12} - M_{22}) \\ &= \hat{\psi}_{\text{B at A1}} - \hat{\psi}_{\text{B at A2}} = (M_{11} - M_{12}) - (M_{21} - M_{22})\end{aligned} \tag{8.8}$$

In two-way designs where at least one factor has ≥ 3 levels, an interaction contrast may be formed by ignoring or collapsing across at least two levels of that factor. For example, the following coefficients define a **pairwise interaction contrast** in a 2 × 3 design (**I**):

(I)

	B_1	B_2	B_3
A_1	1	0	–1
A_2	–1	0	1

In the contrast specified, the simple effect of *A* at B_1 is compared with the simple effect of *A* at B_3. It is equivalent to say that these weights specify the contrast of the simple comparison of B_1 with B_3 across the levels of *A*.

The weights for the **complex interaction contrast** represented next compare B_2 with the average of B_1 and B_3 across the levels of *A* (**II**):

(II)

	B_1	B_2	B_3
A_1	½	–1	½
A_2	–½	1	–½

Exercise 2 asks you to prove that the interaction contrasts defined in (**I**) and (**II**) are orthogonal in a balanced design, and Exercise 3 involves showing that the sums of squares for the omnibus interaction can be uniquely decomposed into the sums of squares for each interaction contrast. The following equation for a balanced design will be helpful for this exercise:

$$SS_{\hat{\psi}_{AB}} = \frac{n\left(\hat{\psi}_{AB}\right)^2}{\left(\sum_{i=1}^{a} c_i^2\right)\left(\sum_{j=1}^{b} c_j^2\right)} \tag{8.9}$$

If at least one factor is quantitative with equally spaced levels, contrast weights for an **interaction trend** may be specified. Suppose in a 2×3 design that factor A represents two different groups of patients and the levels of factor B are three equally spaced dosages of a drug. The weights for the interaction contrast presented below

	B_1	B_2	B_3
A_1	1	−2	1
A_2	−1	2	−1

compare the quadratic effect of the drug across the groups. That the sum of the absolute values of the weights is not 4.0 is not a problem because magnitudes of differential trends are usually estimated with measures of association.

Unlike simple effects, interaction contrasts and main effects are not confounded. For this reason, some researchers prefer to analyze interaction contrasts instead of simple effects when the main effects are relatively large. It is also possible to test a priori hypotheses about specific facets of an omnibus interaction through the specification of interaction contrasts. It is not usually necessary to analyze both simple effects and interaction contrasts in the same design, so either one or the other should be chosen as a way to describe interaction.

TESTS IN BALANCED TWO-WAY DESIGNS

Presented in Table 8.4 are raw scores and descriptive statistics for balanced 2×3 designs, where $n = 3$. The data in the top part of the table are arranged in a layout consistent with a completely between-subjects design,

TABLE 8.4
Raw Scores and Descriptive Statistics for Balanced 2 × 3 Factorial Designs

Completely between-subjects or split-plot layout[a]

	B_1	B_2	B_3	
A_1	8 7 12 9.00 (7.00)[b]	10 11 15 12.00 (7.00)	9 7 11 9.00 (4.00)	10.00
A_2	3 5 7 5.00 (4.00)	5 5 8 6.00 (3.00)	10 14 15 13.00 (7.00)	8.00
	7.00	9.00	11.00	

Completely within-subjects layout

A_1B_1	A_1B_2	A_1B_3	A_2B_1	A_2B_2	A_2B_3
8 7 12 9.00 (7.00)	10 11 15 12.00 (7.00)	9 7 11 9.00 (4.00)	3 5 7 5.00 (4.00)	5 5 8 6.00 (3.00)	10 14 15 13.00 (7.00)

[a]Assumes *A* is the between-subjects factor and *B* is the repeated measures factor. [b]Cell mean (variance).

where each score comes from a different case. The same layout is also consistent with a split-plot design, where the three scores in each row are from the same case (e.g., *A* is a group factor, *B* is a repeated measures factor). The same basic data are presented in the bottom part of the table in a completely within-subjects layout, where the six scores in each row are from the same case. You should verify the following results by applying Equations 8.1 and 8.3 and those in Table 8.2 to the data in Table 8.4 in either layout:

$$SS_W = 64.00$$

$$SS_{A,B,AB} = SS_A + SS_B + SS_{AB} = 18.00 + 48.00 + 84.00 = 150.00$$

$$SS_T = 64.00 + 150.00 = 214.00$$

The results of three different factorial analyses of variance for the data in Table 8.4 assuming fixed factors are reported in Table 8.5. Results in the top of Table 8.5 are from a completely between-subjects analysis, results in

TABLE 8.5
Analysis of Variance Results for the Data in Table 8.4 for Balanced Factorial Designs

Source	*SS*		*df*	*MS*	*F*	*p*	$\hat{\eta}^2$
Completely between-subjects analysis							
Between-subjects effects							
A	18.00		1	18.00	3.38	.091	.084
B	48.00		2	24.00	4.50	.035	.224
AB	84.00		2	42.00	7.88	.007	.393
Within cells (error)	64.00		12	5.33			
Mixed within-subjects analysis							
Between-subjects effects							
A	18.00		1	18.00	1.27	.323	.084
Within-subjects effects							
B	48.00		2	24.00	26.18	<.001	.224
AB	84.00		2	42.00	45.82	<.001	.393
Within cells	64.00		12	5.33			
S/*A* (error for *A*)		56.67	4	14.17			
B × *S*/*A* (error for *B, AB*)		7.33	8	.92			
Completely within-subjects analysis							
Within-subjects effects							
A	18.00		1	18.00	5.68	.140	.084
B	48.00		2	24.00	144.00	<.001	.224
AB	84.00		2	42.00	25.20	.005	.393
Within cells	64.00		12	5.33			
Subjects (*S*)		50.33	2	25.17			
A × *S* (error for *A*)		6.33	2	3.17			
B × *S* (error for *B*)		.67	4	.17			
AB × *S* (error for *AB*)		6.67	4	1.67			

Note. For all analyses, $SS_T = 214.00$ and $df_T = 17$.

the middle of the table are from a split-plot analysis, and results in the bottom of the table are from a completely within-subjects analysis. Note that only the error terms, *F* ratios, and *p* values depend on the design. The sole error term in the completely between-subjects analysis is MS_W, and the statistical assumptions for tests with it are described in Chapter 3. In the split-plot analysis, SS_W and df_W are partitioned to form two different error terms, one for between-subjects effects (*A*) and another for repeated measures effects (*B*, *AB*). Tests with the former error term, designated in the table as *S*/*A* for "subjects within groups under *A*," assume homogeneity of variance for case average scores across the levels of the repeated measures factor. The within-subjects error

term for a nonadditive model, $B \times S/A$, assumes that both within-populations covariance matrices on the repeated measures factor are not only spherical but equal; see Kirk (2012, Chapter 12) and Winer et al. (1991, pp. 512–526) for more information. A different partition of SS_W in the completely within-subjects analysis results in sums of squares for three different error terms for repeated measures effects, $A \times S$, $B \times S$, and $AB \times S$, and a sum of squares for the subjects effect, S (see Table 8.5).

Factorial ANOVA generates the same basic source tables when either one or both factors are considered random instead of fixed. The only difference is that main effects may not have the same error terms in a random effects model as in a fixed effects model. For example, the error term for the main effects in a completely between-subjects design with two random factors is MS_{AB}, not MS_W, but the latter is still the error term for the AB effect. Tabachnick and Fidell (2001) gave a succinct, nontechnical explanation: Because levels of both factors are randomly selected, it is possible that a special interaction occurred with these particular levels. This special interaction may confound the main effects, but dividing the mean squares for the main effects by MS_{AB} is expected to cancel out these confounds in statistical tests of the former. Maximum likelihood estimation can also be used to estimate effects of random factors in large samples.

Extensions to Designs With Three or More Factors

All of the principles discussed to this point extend to balanced factorial designs with more than two factors. For example, there are a total of seven estimated main and interaction effects in a three-way design, including three main effects (A, B, and C) each averaged over the other two factors, three two-way interactions (AB, AC, and BC) each averaged over the third factor, and the highest order interaction, ABC. A three-way interaction means that the effect of each factor on the outcome variable changes across the levels of the other two factors. It also means that the **simple interactions** of any two factors are not the same across the levels of the third factor (e.g., the AB effect changes across C_1, C_2, and so on). Omnibus ABC effects can be partitioned into three-way interaction contrasts. When expressed as a mean difference contrast, a three-way interaction contrast involves two levels from each factor. That is, a $2 \times 2 \times 2$ matrix of means is analyzed, and the sum of the absolute values of the contrast weights that specify it is 8.0. Exercise 4 asks you to prove that the three-way interaction in a $2 \times 2 \times 2$ design equals the difference between all possible pairs of simple interactions, or

$$\begin{aligned} \hat{\psi}_{ABC} &= \hat{\psi}_{AB \text{ at } C1} - \hat{\psi}_{AB \text{ at } C2} = \hat{\psi}_{AC \text{ at } B1} - \hat{\psi}_{AC \text{ at } B2} \\ &= \hat{\psi}_{BC \text{ at } A1} - \hat{\psi}_{BC \text{ at } A2} \end{aligned} \tag{8.10}$$

See Keppel and Wickens (2004, Chapter 22) for examples of the specification of three-way interaction contrasts. Winer et al. (1991, pp. 333–342) described geometric interpretations of three-way interactions including interaction contrasts.

Just as in two-way designs, the derivation of effect sums of squares in factorial designs with three or more independent variables is the same regardless of whether the factors are between-subjects versus within-subjects or fixed versus random. But there may be no proper ANOVA error term for some effects, given certain combinations of fixed and random factors in complex designs. By "proper" I mean that the expected mean square of the error term estimates all sources of variance as the numerator except that of the effect being tested. There are algorithms to derive by hand expected mean squares in various factorial designs (e.g., Winer et al., 1991, pp. 369–382), and with such an algorithm it may be possible in complex designs to pool mean squares and form **quasi-*F* ratios** with proper error terms. Another method is **preliminary testing and pooling**, where parameters for higher order interactions with random variables may be dropped from the analysis based on results of statistical tests. The goal is to find a simplified model with fewer parameters that generates expected mean squares so that all effects have proper error terms. But a testing and pooling procedure based solely on statistical significance is flawed because it ignores effect size and capitalizes on chance.

Keeping track of error terms that go along with different effects in designs with both fixed and random factors is one of the complications in factorial ANOVA. It also highlights the fact that *p* values in such designs are merely estimates, and different algorithms for deriving expected mean squares can yield different *p* values for the same effect in the same sample. Considering the shortcomings of statistical tests in perhaps most studies in the behavioral sciences, however, it is best not to focus too much attention on *p* values to the neglect of other, more useful information in ANOVA source tables. This idea is elaborated next.

ANALYSIS STRATEGY

There are many effects that could be analyzed in two-way designs—main, interaction, simple, and focused comparisons for any of the aforementioned effects that are omnibus. The number of effects grows exponentially as even more factors are added to the design. One can easily get lost by estimating every possible effect in a factorial analysis. It is thus crucial to have a plan that minimizes the number of analyses while still respecting the essential research hypotheses.

Some of the worst misuses of statistical tests are seen in factorial designs where all possible effects are tested and sorted into two categories, those

found to be statistically significant and then discussed at length versus those found to be not statistically significant and then ignored. Perhaps researchers who do so believe the filter myth (see Chapter 4). These mistakes are compounded when power is ignored. This is because power can be different for various effects in a factorial design. This is so because (a) the degrees of freedom associated with the *F* statistic for different effects can be different and (b) the numbers of scores that contribute to different means vary. In a balanced 2×4 design where $n = 10$, for example, the two means for the *A* main effect are each based on 40 scores, but the four means of the *B* main effect are each based on just 20 scores. The power for the test of the *A* main effect may be different from the power of the test of the *B* main effect, and the power for the test of the *AB* effect may be different still. This is especially true in split-plot designs where some effects are between-subjects and others are within-subjects (see Table 8.5).

Too many sources emphasize statistical significance only when outlining ANOVA rituals for factorial designs, but this is a potential route to folly. It is better to estimate a priori power if statistical tests are to be used combined with effect size estimation. It also helps to realize that as the size of interaction effects becomes larger compared to those of the main effects, detailed analysis of the latter is increasingly fruitless. This is especially true for disordinal interactions, where some simple effects are reversed relative to the main effects for the same factor (see Table 8.3). A better general strategy is described next.

A factorial design affords the opportunity to estimate presumed interactions. Thus, analysis of interaction should take center stage. That is, do not take your eyes off the prize of hypotheses about interaction that motivated the specification of a factorial design. State a minimum effect size that corresponds to a substantially meaningful interaction, one that is not trivial in magnitude (Steiger, 2004). Look to meta-analytic studies in your area for guidelines about effect size. If no such studies exist, you must rely on your knowledge of prior studies, the population, and the outcome measures to make an informed estimate.

If the interaction effect is omnibus ($df > 1$) and you have a priori hypotheses about specific facets of that effect, specify and estimate the magnitudes of the corresponding interaction contrasts or interaction trends. Analyzing simple effects is an alternative when the researcher predicts the presence of interaction but not its specific form. Analyzing simple comparisons is an option for omnibus simple effects, but fishing expeditions are to be avoided if such comparisons are unplanned.

Main effects warrant little attention if the sizes of interaction effects are appreciable and they are disordinal. This advice is counter to some factorial ANOVA rituals where all "significant" effects are slavishly interpreted regard-

less of effect size or prior hypotheses. Otherwise, consider whether the magnitudes of the main effects are appreciable. If so, contrast analysis is an option for omnibus main effects, but again avoid blind reliance on significance testing if main comparisons are unplanned. Always report the full source table. In it include information about the main and interaction effects and any other effects, such as simple effects, that were analyzed. Always include effect sizes; a source table without effect sizes is incomplete.

Model Testing

Lunneborg (2000) explored the theme of treating factorial ANOVA as a model-fitting technique in the same way that regression procedures can be applied; that is, in ways not strictly exploratory. An example of this approach is when researchers compare the relative fits to the data of two different models in ANOVA, the **complete structural model** versus a **reduced structural model**. A structural model expresses a hypothetical score as the sum of population parameters that correspond to sources of variation. The complete model for a between-subjects two-way design with fixed factors is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk} \tag{8.11}$$

where Y_{ijk} is the kth score in the cell at the ith level of factor A and the jth level of factor B; μ is the population grand mean; α_i, β_j, and $\alpha\beta_{ij}$, respectively, represent the population main and interaction effects as deviations from the grand mean; and ε_{ijk} is a random error component. This model underlies the derivation of the sums of squares for the source table in the top of Table 8.5. The complete structural models that underlie the other two source tables in Table 8.5 are somewhat different because either one or both factors are within-subjects, but the idea is the same. A structural model generates predicted marginal and cell means, but these predicted means equal their observed counterparts for a complete model. That is, the observed marginal means estimate population main effects, and the observed cell means estimate the population interaction effect.

A reduced structural model does not include parameters for all effects. Parameters in the complete model are typically considered for exclusion in a sequential order beginning with the highest order interaction. If the parameters for this interaction are retained, the complete model cannot be simplified. But if the parameters that correspond to $\alpha\beta_{ij}$ are dropped, the complete model reduces to the **main effects model**

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \tag{8.12}$$

which assumes only population main effects. Two consequences arise from rejection of the complete structural model in favor of the main effects model. First, the sums of squares for the *AB* effect are pooled with those of the total within-cells variability to form a composite error term for tests of the main effects. Second, the reduced structural model generates predicted cell means that may differ from the observed cell means. It is also possible that standard errors of differences between predicted cell means may be less than those for differences between observed cell means. Accordingly, the researcher may choose to analyze the predicted cell means instead of the observed cell means. This choice also impacts effect size estimation.

In brief, there are two grounds for simplifying structural models, rational and empirical. The first is based on the researcher's domain knowledge about underlying sources of variation, and the second is based on results of statistical tests. In the latter approach, parameters for interactions that are not statistically significant become candidates for exclusion from the model. The empirical approach is controversial because it capitalizes on chance and ignores effect size.

NONORTHOGONAL DESIGNS

If all factorial designs were balanced—or at least had unequal but proportional cell sizes—there would be no need to deal with the technical problem raised next. Only two-way nonorthogonal designs are discussed, but the basic principles extend to larger nonorthogonal designs. One problem is that the factors are correlated, which means that there is no single, unambiguous way to apportion the total effects sum of squares to individual effects. A second concerns ambiguity in estimates for means that correspond to main effects. This happens because there are two different ways to compute marginal means in unbalanced designs: as arithmetic or as weighted averages of the corresponding row or column cell means. Consider the data in Table 8.6 for a nonorthogonal 2 × 2 design. The two ways to calculate the marginal mean for A_2 just mentioned are summarized respectively as

$$\frac{2.00 + 5.33}{2} = 3.67 \quad \text{versus} \quad \frac{2\,(2.00) + 6\,(5.33)}{8} = 4.50$$

The value to the right is the same as that you would find if working with the eight raw scores in the second row of the 2 × 2 matrix in Table 8.6. There is no such ambiguity in balanced designs.

TABLE 8.6
Raw Scores and Descriptive Statistics for a Nonorthogonal 2×2 Factorial Design

	B_1	B_2	Row means
A_1	2, 3, 4 3.00 (1.00)[a]	1, 3 2.00 (2.00)	2.50/2.60[b]
A_2	1, 3 2.00 (2.00)	4, 5, 5, 6, 6, 6 5.33 (.67)	3.67/4.50
Column means	2.50/2.60	3.67/4.50	3.08/3.77

[a]Cell mean (variance). [b]Arithmetic/weighted average of the corresponding cell means.

There are several methods for analyzing data from nonorthogonal designs (e.g., Maxwell & Delaney, 2004, pp. 320–343). Statisticians do not agree about optimal methods for different situations, so it is not possible to give definitive recommendations. Most of these methods attempt to correct effect sums of squares for overlap. They give the same results only in balanced designs, and estimates from different methods tend to diverge as the cell sizes become increasingly disproportional. Computer procedures for factorial ANOVA typically use by default one of the methods described next. If the default is not suitable in a particular study, the researcher must specify a better method.

An older method for nonorthogonal designs amenable to hand calculation is unweighted means analysis. Effect sums of squares are computed in this method using the equations for balanced designs, such as those in Table 8.2 for two-way designs, except that the design cell size is taken as the harmonic mean of the actual cell sizes, or

$$n_{\mathrm{h}} = \frac{ab}{\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{n_{ij}}} \tag{8.13}$$

This method estimates marginal means as arithmetic averages of the corresponding row or column cell means. It also estimates the grand mean as the arithmetic average of the cell means. A consequence of weighting all cell means equally is that overlapping variance is not attributed to any individual effect. Thus, the unweighted means method generates adjusted sums of squares that reflect unique predictive power.

A related regression-based technique called Method 1 by Overall and Spiegel (1969) estimates effect sums of squares controlling for all other effects. In a two-way design, for example, estimates for the main effect of factor *A* in this method are adjusted for both the *B* main effect and the interaction effect *AB*, and estimates for the *B* main effect are corrected for both the *A* and *AB* effects. Estimates for the interaction are always adjusted for the main effects (see, e.g., the last equation in Table 8.2). These Method 1 adjusted sums of squares may be labeled Type III in the output of statistical software programs.

The methods just described may be best for experiments designed with equal cell sizes but where there was random data loss from a few cells. This is because cells with fewer scores by chance are not weighted less heavily in either method. In nonexperimental designs, though, disproportional cell sizes may arise due to population correlations between factors. If so, it may be better for the actual cell sizes to contribute to the analysis. Two other regression-based methods do just that. They also give higher priority to one or both main effects than do the methods described earlier. Overall and Spiegel (1969) referred to these techniques as Method 2 and Method 3, and sums of squares generated by them may be labeled in computer program output as Type II for the former versus Type I for the latter.

In Method 2, main effect sums of squares are adjusted for overlap only with each other, but the interaction sum of squares is corrected for both main effects. For example, estimates for the *A* main effect are corrected only for the *B* main effect and not for the interaction effect. Corresponding corrections are made for the *B* main effect (i.e., corrected for *A* but not for *AB*). As in Method 1, estimates for the *AB* effect are adjusted for both main effects. Method 3 does not remove shared variance from the sums of squares of one main effect (e.g., *A*); it adjusts the sums of squares of the other main effect for overlap with the first (e.g., *B* adjusted for *A*) and then adjusts the interaction sum of squares for both main effects. If the researcher has no a priori hypotheses about effect priority but wishes the cell sizes to influence the results, Method 2 should be preferred over Method 3. Too many researchers neglect to state the method used to analyze data from nonorthogonal designs much less explain their choice, if one was intentionally made.

The data from the nonorthogonal 2 × 2 design in Table 8.6 were analyzed with the three regression-based approaches just described, assuming a completely between-subjects design with fixed factors. The results are summarized in Table 8.7. Observe that the sums of squares for the total effects, interaction, pooled within-cells variation, and total data set are the same across all three analyses. It is the estimates for the main effects that change depending on the method. Neither main effect has a p value less than .05 in Method 1/Type III sums of squares ($p = .094$ for both), which adjusts main effects for all other effects. Proportions of total variance explained by the main effects,

TABLE 8.7
Results of Three Different Regression Methods for the Data in Table 8.6 From a Nonorthogonal Design

Source	*SS*	*df*	*MS*	*F*	*p*	$\hat{\eta}^2$
Method 1/Type III[a]						
Total effects (*A, B, AB*)	28.97	3	9.66	9.31	.004	.756
A adjusted for *B, AB*	3.63	1	3.63	3.50	.094	.095
B adjusted for *A, AB*	3.63	1	3.63	3.50	.094	.095
AB adjusted for *A, B*	12.52	1	12.52	12.07	.007	.327
Method 2/Type II						
Total effects (*A, B, AB*)	28.97	3	9.66	9.31	.004	.756
A adjusted for *B*	5.35	1	5.35	5.15	.049	.140
B adjusted for *A*	5.35	1	5.35	5.15	.049	.140
AB adjusted for *A, B*	12.52	1	12.52	12.07	.007	.327
Method 3/Type I						
Total effects (*A, B, AB*)	28.97	3	9.66	9.31	.004	.756
A (unadjusted)	11.11	1	11.11	10.71	.010	.290
B adjusted for *A*	5.35	1	5.35	5.15	.049	.140
AB adjusted for *A, B*	12.52	1	12.52	12.07	.007	.327

Note. For all analyses, $SS_W = 9.33$; $df_W = 9$; $MS_W = 1.04$; $SS_T = 38.31$; and $df_T = 12$.
[a]Overall and Spiegel (1969) method/sum of squares type.

$\hat{\eta}^2_A = \hat{\eta}^2_B = .095$, are also the lowest in this method. The *p* values for both main effects are <.05 and have greater explanatory power in Method 2/Type II sums of squares and Method 3/Type I sums of squares, which gives them higher priority than in Method 1. Only in Method 3—which analyzes the *A*, *B*, and *AB* effects sequentially in this order—are the sums of squares and $\hat{\eta}^2$ values additive but not unique.

Which of the three sets of results in Table 8.7 is correct? From a purely statistical view, all are because there is no definitive way to estimate effect sums of squares in nonorthogonal designs. There may be a preference for one set of results given a clear rationale about effect priority. But without such a justification, there is no basis for choosing among these results.

STANDARDIZED CONTRASTS

Designs with fixed factors are assumed next. Methods for standardizing contrasts in factorial designs are not as well developed as they are for one-way designs. There is also not complete agreement across works by Glass, McGaw, and Smith (1981); Morris and DeShon (1997); Cortina and Nouri (2000);

and Olejnik and Algina (2000) that address this issue. It is therefore not possible to describe a complete method. But this discussion is consistent with a general theme of the sources just cited: how to make standardized contrasts from factorial designs comparable to those that would have occurred in nonfactorial designs. This implies that (a) estimates for effects of each independent variable in a factorial design should be comparable to effect sizes for the same factor studied in a single-factor design and (b) changing the number of factors in the design should not necessarily change the effect size estimates for any one of them.

Standardized mean differences may be preferred over measures of association if contrasts are the focus of the analysis, such as in designs where all factors have just two levels. An advantage of measures of association is that they can summarize with a single number total predictive power across the whole design, such as $\hat{\eta}^2_{A,\,B,\,AB}$ in a two-way design. There is no analogous capability with standardized contrasts. The two families of effect size statistics can also be used together; see Wayne, Riordan, and Thomas (2001) for an example.

Standardized contrasts in factorial designs have the same general form as they do in one-way designs, $\hat{\psi}/\hat{\sigma}^*$, where the denominator estimates a population standard deviation. But it is more difficult in factorial designs to figure out which standard deviation should be the standardizer. This is because what is probably the most general choice in a one-way design, the square root of MS_W, may not be the best option in a factorial design (i.e., the effect size is d_{with}). Also, there is more in the literature about standardizing main comparisons than simple comparisons in factorial designs. This is unfortunate because main comparisons may be uninteresting when there is interaction. I recommend that main and simple comparisons for the same factor have the same standardizer. This makes $\hat{\psi}/\hat{\sigma}^*$ for these two kinds of single-factor contrasts directly comparable.

Single-Factor Contrasts in Completely Between-Subjects Designs

The choice of the standardizer for a single-factor contrast, such as for a simple comparison of two levels of factor *A* at the B_1 level, is determined by (a) the distinction between the **factor of interest (targeted factor)** versus the **off-factors (peripheral factors)** and (b) whether or not the off-factors vary naturally in the population. That is, the off-factors are, respectively, **intrinsic factors** versus **extrinsic factors**. Intrinsic factors tend to be individual-difference, group, or classificatory factors that are nonexperimental, such as gender. Glass et al. (1981) referred to intrinsic factors as being of theoretical interest concerning estimation of population standard deviation. In contrast, extrinsic factors do not vary naturally, and they tend to be manipulated or experimental factors.

Suppose in a two-way design that two levels of factor *A* are compared. The factor of interest is *A*, and *B* is the off-factor. Suppose that the off-factor *B* varies naturally in the population. The square root of MS_W may not be an appropriate standardizer for contrasts between levels of *A* in this case. This is because MS_W controls for the effects of both factors, including their interaction. We can see this in the following expression for a balanced two-way design:

$$MS_W = \frac{SS_W}{df_W} = \frac{SS_T - SS_A - SS_B - SS_{AB}}{df_T - df_A - df_B - df_{AB}} \tag{8.14}$$

Because MS_W does not reflect variability due to effects of the intrinsic off-factor *B*, its square root may underestimate σ. This implies that a contrast between levels of *A* standardized against $(MS_W)^{1/2}$ may overestimate the absolute population effect size. A way to calculate an alternative standardizer that reflects the total variation on off-factor *B* is described below.

Now suppose that the off-factor *B* is extrinsic (it does not vary naturally in the population). Such factors are more likely to be manipulated or repeated measures variables than individual difference variables. For example, the theoretical population for the study of a new treatment can be viewed as follows: It is true either that every case in the population is given the treatment or that none of them are given the treatment. In either event, there is no variability because of treatment versus no treatment (Cortina & Nouri, 2000). Because extrinsic off-factors are not of theoretical interest for the sake of variance estimation, their effects should not contribute to the standardizer. In this case, the square root of MS_W from the two-way ANOVA would be a suitable denominator for standardized contrasts on factor *A* when the off-factor *B* does not vary naturally.

Described next are two methods to standardize main or simple comparisons that estimate the full range of variability on an intrinsic off-factor that varies naturally in the population. Both methods pool the variances across all levels of the factor of interest, so they also generate standardized contrasts for single-factor comparisons in factorial designs that are directly comparable with d_{with} in single-factor designs. These two methods yield the same result in balanced designs. The first is the **orthogonal sums of squares method** (Glass et al., 1981). It requires a complete source table with additive sums of squares. Assuming that *A* is the factor of interest, the following term estimates the full range of variability on the intrinsic off-factor *B*:

$$MS_{W,B,AB} = \frac{SS_W + SS_B + SS_{AB}}{df_W + df_B + df_{AB}} = \frac{SS_T - SS_A}{df_T - df_A} \tag{8.15}$$

The subscript for the mean square indicates that variability associated with the *B* and *AB* effects is pooled with error variance. Equation 8.15 also shows that $MS_{W, B, AB}$ in a two-way design has the same form as MS_W in a one-way design, where *A* is the sole factor. Indeed, the two terms just mentioned are equal in balanced two-way designs, where MS_W in a single-factor ANOVA is computed after collapsing the data across the levels of the off-factor *B*. The square root of $MS_{W, B, AB}$ is the standardizer for contrasts between levels of factor *A* in this method.

The **reduced cross-classification method** (Olejnik & Algina, 2000) does not require a complete source table with additive sums of squares. It also generates unique adjusted-variance estimates in unbalanced designs. The researcher creates with a statistical software program a reduced cross-classification of the data where the off-factor that varies naturally in the population is omitted. Next, a one-way ANOVA is conducted for the factor of interest, and the square root of the error term in this analysis is taken as the standardizer for contrasts on that factor. In balanced designs, this standardizer equals the square root of $MS_{W, B, AB}$ computed with Equation 8.15. It also equals MS_W in the one-way ANOVA for factor *A* after collapsing across the levels of factor *B*. A variation is needed when working with a secondary source that reports only cell descriptive statistics. In this case, the variance $MS_{W, B, AB}$ can be derived as follows:

$$MS_{W, B, AB} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\left[df_{ij}\,(s_{ij}^2) + n_{ij}(M_{ij} - M_{A,})^2 \right]}{N - a} \tag{8.16}$$

This equation is not as complicated as it appears. Its numerator involves the computation of a "total" sum of squares within each level of the factor of interest *A* that reflects the full range of variability on the intrinsic off-factor *B*. This is done by combining cell variances across levels of *B* and taking account of the simple effect of *B* at that level of *A*. These "total" sums of squares are added up across the levels of *A* and then divided by $N - a$, the total within-conditions degrees of freedom in the reduced cross-classification where *A* is the only factor.

The methods described for standardizing contrasts that involve one factor in the presence of an off-factor that varies naturally in the population can be extended to designs with three or more factors. For example, we can state the following general rule for the reduced cross-classification method:

> The standardizer for a single-factor comparison is the square root of MS_W from the cross-classification that includes the factor of interest and any off-factors that do not vary naturally in the population but excludes any off-factors that do.

Suppose in a three-way design that *A* is the factor of interest. Of the two off-factors, *B* varies naturally but *C* does not. According to the rule, the denominator of standardized contrasts for main or simple comparisons is the square root of the MS_W from the two-way ANOVA for the reduced cross-classification that includes factors *A* and *C* but not *B*. This standard deviation estimates the full range of variability on off-factor *B* but not on off-factor *C*. If the design were balanced, we would get the same result by taking the square root of the following variance,

$$MS_{W,\,B,\,AB\,,\,BC,\,ABC} = \frac{SS_W + SS_B + SS_{AB} + SS_{BC} + SS_{ABC}}{df_W + df_B + df_{AB} + df_{BC} + df_{ABC}} \qquad (8.17)$$

which pools the within-conditions variability in the three-way ANOVA with all effects that involve the off-factor *B*. As Cortina and Nouri (2000) noted, however, there is little statistical research that supports the general rule stated earlier for different combinations of off-factors, some intrinsic but others extrinsic, in complex designs. One hopes that such research will be forthcoming. In the meantime, you should explain in summaries of your analyses exactly how main or simple comparisons were standardized.

Let us consider an example for a balanced two-way design where factor *B* varies naturally in the population but factor *A* does not. The orthogonal sums of squares method is demonstrated using the source table at the top of Table 8.5 for the data in Table 8.4 for a completely between-subjects 2×3 design, where $n = 3$ and

$$SS_A = 18.00,\ SS_B = 48.00,\ SS_{AB} = 84.00, \text{ and } SS_W = 64.00$$

The standardizer for main or simple comparisons where *A* is the factor of interest and *B* is an intrinsic off-factor is the square root of

$$MS_{W,\,A,B,\,AB} = \frac{64.00 + 48.00 + 84.00}{12 + 2 + 2} = 12.25$$

or 3.50. As expected, the variance just computed (12.25) is greater than $MS_W = 5.33$ from the original 2×3 analysis (see Table 8.5) because the former includes effects of the off-factor *B*. Standardized contrasts for the three simple comparisons of *A* at *B* are derived as follows:

$$d_{A \text{ at } B1} = \frac{9.00 - 5.00}{3.50} = 1.14 \quad \text{and} \quad d_{A \text{ at } B2} = \frac{12.00 - 6.00}{3.50} = 1.71$$

$$d_{A \text{ at } B3} = \frac{9.00 - 13.00}{3.50} = -1.14$$

Where the standardizer, 3.50, is the square root of $MS_{W,\,B,\,AB} = 12.25$. In summary, the interaction is disordinal because at least one simple effect of *A* reverses over the levels of *B*. In particular, the mean difference between A_1 and A_2 is positive and exceeds one full standard deviation at levels B_1 and B_2, but the difference is negative—about −1.14 standard deviations—at B_3. These results precisely describe how the effect of factor *A* changes across the levels of *B* in standard deviation units. Exercise 5 asks you to apply the reduced cross-classification method (Equation 8.16) to calculate $MS_{W,\,B,\,AB} = 12.25$ based on the cell descriptive statistics in Table 8.4 for this example.

Continuing with this example, a better standardizer for main or simple comparisons on factor *B*—for which *A* is the off-factor and assuming that *A* does not varies naturally—is the square root of $MS_W = 5.33$, or 2.31, from the two-way factorial ANOVA for these data (see Table 8.5). This example shows that different sets of simple comparisons in the same factorial design may have different standardizers. The choice of which set of simple comparisons to analyze (i.e., those of *A* at *B* vs. *B* at *A*) should be based on theoretical grounds, not on whichever set would have the smaller standardizer. Other options to standardize main or simple comparisons in factorial designs are discussed by Cortina and Nouri (2000) and Morris and DeShon (1997).

Interaction Contrasts in Completely Between-Subjects Designs

Suppose that the within-cells variances in a factorial design are similar and all contrasts are standardized against the square root of MS_W. This would make sense in a study in which none of the factors vary naturally in the population (e.g., the design is experimental). The standardized two-way interaction contrast would in this case equal the difference between either pair of standardized simple comparisons, row-wise or column-wise. For example, the following relation would be observed in a 2 × 2 design where all effects are contrasts:

$$d_{\hat{\psi}_{AB}} = d_{A\text{ at }B1} - d_{A\text{ at }B2} = d_{B\text{ at }A1} - d_{B\text{ at }A2} \tag{8.18}$$

But this relation may not hold if either factor varies naturally in the population. This is because different sets of simple comparisons can have different standardizers in this case. Because interaction is a joint effect, however, there are no off-factors.

There is relatively little in the statistical literature about exactly how to standardize an interaction contrast when only some factors vary naturally. Suppose in a balanced 2 × 2 design that factor *B* varies naturally, but factor *A*

does not. Should $\hat{\psi}_{AB}$ be standardized against the square root of MS_W from the two-way ANOVA or against the square root of $MS_{W,B,AB}$? The former excludes the interaction effect. This seems desirable in a standardizer for $\hat{\psi}_{AB}$, but it also excludes variability due to the *B* main effect, which implies that σ may be underestimated. The term $MS_{W,B,AB}$ reflects variability due to the interaction, but standardizers for single-factor comparisons do not generally reflect variability because of the main effects of those factors. Olejnik and Algina (2000, pp. 251–253) described a way to choose between the variances just mentioned, but it requires designating one of the independent variables as the factor of interest. This may be an arbitrary decision for an interaction effect.

It is also possible to standardize $\hat{\psi}_{ABC}$ for a three-way interaction contrast, but it is rare to see standardized contrasts for interactions among three or more factors. If all comparisons are scaled as mean difference contrasts, the following relation would be observed in a $2 \times 2 \times 2$ design where all effects are single-*df* comparisons:

$$\begin{aligned} d_{\hat{\psi}_{ABC}} &= d_{AB \text{ at } C1} - d_{AB \text{ at } C2} = d_{AC \text{ at } B1} - d_{AC \text{ at } B2} \\ &= d_{BC \text{ at } A1} - d_{BC \text{ at } A2} \end{aligned} \tag{8.19}$$

That is, the standardized three-way interaction equals the difference between the standardized simple interactions for any two factors across the levels of the third factor. But this relation may not hold if different sets of simple interactions have different standardizers.

These uncertainties should not affect researchers who analyze simple effects instead of interaction contrasts as a way to understand a conditional effect. There should also be little problem in experimental designs where the square root of MS_W may be an appropriate standardizer for any contrast, and researchers who can specify a priori interaction contrasts also tend to work with experimental designs.

Designs With Repeated Measures Factors

Olejnik and Algina (2000) recommended standardizers in the metric of the original scores for contrasts that involve within-subjects factors. This makes standardized contrasts more directly comparable across different factorial designs. A common design is a split-plot design where unrelated samples are compared across multiple measurement occasions. Assuming homoscedasticity, the square root of MS_W is a natural choice for standardizing comparisons between groups. It also treats the repeated measures factor as an extrinsic off-factor that does not vary naturally in the population, but the same standardizer ignores cross-conditions correlations for within-subjects

factors. Cortina and Nouri (2000) described a method to standardize group comparisons after collapsing the data across levels of repeated measures factors; see also Cumming (2012, pp. 413–416) for advice about confidence intervals in split-plot designs.

The difference between the standardized mean changes for any two groups in a split-plot design is a standardized interaction contrast. For example, if the pretest-to-posttest standardized mean change is .75 for the treatment group and .10 for the control group, the standardized interaction contrast equals the difference, or .75 − .10 = .65. That is, the change for the treatment group is .65 standard deviations greater than the change for the control group.

Interval Estimation

Some of the computer tools described in previous chapters generate confidence intervals for standardized contrasts in factorial designs. For example, PSY (Bird et al., 2000; see footnote 2, Chapter 7) analyzes raw data from factorial designs with one or more between-subjects factors or one or more within-subjects factors. It standardizes all contrasts against the square root of MS_W for the whole design. The SAS/IML script by Keselman et al. (2008) analyzes trimmed means and Winsorized variances in factorial designs with fixed factors that are either between-subjects or within-subjects. It also generates bootstrapped confidence intervals based on robust standardized contrasts (see footnote 7, Chapter 5). Wilcox's (2012) WRS package for R also constructs bootstrapped confidence intervals based on robust standardized contrasts in factorial designs (see footnote 11, Chapter 2).

MEASURES OF ASSOCIATION

Outlined next are descriptive and inferential measures of association for designs with fixed or random factors.

Descriptive Measures

The effect size $\hat{\eta}^2 = SS_{effect}/SS_T$ in factorial designs with fixed factors is the proportion of total variance explained by an effect. The proportion of residual variance explained after removing all systematic effects from total variance other than that due to the effect of interest is partial $\hat{\eta}^2 = SS_{effect}/(SS_{effect} + SS_{error})$, where SS_{error} is the sum of squares for the effect error term. Some researchers report $\hat{\eta}^2$ for total effects and partial $\hat{\eta}^2$ for individual effects, such as

$$\hat{\eta}^2_{A,\,B,\,AB,}, \text{ partial } \hat{\eta}^2_A, \text{ partial } \hat{\eta}^2_B, \text{ and partial } \hat{\eta}^2_{AB}$$

in two-way designs. The rationale is that the denominators of the effect sizes just listed all have the general form, $SS_{\text{effect}} + SS_{\text{error}}$.

Pierce, Block, and Aguinis (2004) noted that too many researchers erroneously report partial $\hat{\eta}^2$ values as $\hat{\eta}^2$ for individual effects. This is potentially misleading because (a) partial $\hat{\eta}^2$ usually exceeds $\hat{\eta}^2$ for the same effect and (b) values of partial $\hat{\eta}^2$ are not generally additive even within sets of orthogonal effects. Indeed, partial $\hat{\eta}^2$ values for the individual main and interactive effects can sum to greater than 1.0, because these statistics can be based on different yet overlapping subsets of total variance.

Olejnik and Algina (2003) argued that measures of association for effects in factorial designs should reflect whether other factors are extrinsic or intrinsic. Suppose in a balanced, completely between-subjects design that *A* is a manipulated (experimental) extrinsic factor that does not vary naturally in the population, but factor *B* is gender. An appropriate effect size for factor *A* is $\hat{\eta}_A^2$ because its denominator, SS_T, reflects variability due to gender (i.e., *B*, *AB*). But a better effect size for intrinsic factor *B* is $SS_B/(SS_T - SS_A)$, where the denominator removes effects due to intrinsic factor *A* but preserves all effects of gender. This ratio is neither $\hat{\eta}_B^2$ nor partial $\hat{\eta}_B^2$. The effect size $SS_{AB}/(SS_T - SS_A)$ for the interaction has the same rationale.

Now suppose that both factors are measured, intrinsic variables. Computing effect sizes as SS_{effect}/SS_T (i.e., $\hat{\eta}^2$) for *A*, *B*, and *AB* preserves all variation due to these factors in the denominator. But if both *A* and *B* are manipulated, extrinsic factors that do not vary naturally, the effect size partial $\hat{\eta}^2$, or $SS_{\text{effect}}/(SS_{\text{effect}} + SS_W)$, removes from the denominator variation due to main or interactive effects of these factors.

Olejnik and Algina (2003) defined **generalized estimated eta-squared** for balanced factorial designs that takes account of whether factors are manipulated (extrinsic) or measured (intrinsic) and also whether they are between- or within-subjects. It controls for the presence of covariates in the analysis. Its general form is

$$\text{generalized } \hat{\eta}^2_{\text{effect}} = \frac{SS_{\text{effect}}}{(m \times SS_{\text{effect}}) + \sum SS_{\text{meas}} + \sum SS_{\text{sub, cov}}} \tag{8.20}$$

where $m = 1$ if the effect of interest concerns a manipulated factor but is zero otherwise; $\sum SS_{\text{meas}}$ is the total of the sums of squares for effects of all measured factors; and $\sum SS_{\text{S, cov}}$ is the total of the sums of squares for all effects that concern covariates (if any) or subjects. The latter includes sums of squares for the subjects effect in a correlated design or for error terms based on within-cells variation, the total of which is SS_W (see Table 8.5). If the effect is for a measured factor, SS_{effect} is already included in the expression $\sum SS_{\text{meas}}$ in

Equation 8.19; thus, $m = 0$ for such effects. But for effects of manipulated factors, setting $m = 1$ in the equation includes SS_{effect} in the denominator of generalized $\hat{\eta}^2$.

Suppose that *A* is a manipulated factor and *B* is a measured factor in the completely between-subjects analysis at the top of Table 8.5. In this analysis, where

$$SS_A = 18.00, SS_B = 48.00, SS_{AB} = 84.00, SS_W = 64.00, \text{and } SS_T = 214.00$$

$\hat{\eta}^2_{AB} = .393$ and partial $\hat{\eta}^2_{AB} = .568$, but neither effect size controls for the status of the factors as manipulated versus measured. Because factor *B* is measured, $m = 0$ in Equation 8.20 and

$$\sum SS_{meas} = SS_B + SS_{AB} = 132.00 \quad \text{and} \quad \sum SS_{S,cov} = SS_W = 64.00$$

The denominator of generalized $\hat{\eta}^2_{AB}$ is thus 132.00 + 64.00, or 196.00, which also equals $SS_T - SS_A$, or 214.00 – 18.00. The whole expression is

$$\text{generalized } \hat{\eta}^2_{AB} = \frac{84.00}{196.00} = .429$$

Thus, the interaction explains about 42.9% of the residual variable controlling only for the main effect of extrinsic factor *A*, which does not vary naturally in the population.

Inferential Measures

The effect sizes described next assume balanced designs. The inferential measures of association $\hat{\omega}^2$ or partial $\hat{\omega}^2$ for effects of fixed factors and $\hat{\rho}_I$ or partial $\hat{\rho}_I$ for effects of random factors are estimated as ratios of variance components that depend on the design (see Chapter 7). I use the symbol $\hat{\rho}_I$ only if all factors are random. An equation for directly computing $\hat{\omega}^2$ that is good for any effect in a completely between-subjects factorial design is

$$\hat{\omega}^2_{effect} = \frac{df_{effect}(MS_{effect} - MS_W)}{SS_T + MS_W} \tag{8.21}$$

An equation for partial $\hat{\omega}^2$ for any effect in the same kind of design is

$$\text{partial } \hat{\omega}^2_{effect} = \frac{df_{effect}(F_{effect} - 1)}{df_{effect}(F_{effect} - 1) + N} \tag{8.22}$$

TABLE 8.8
Equations for Variance Components Estimators in Completely Between-Subjects Two-Way Designs

Estimator	Both factors random	*A* random, *B* fixed
$\hat{\sigma}^2_A$	$\frac{1}{bn}(MS_A - MS_{AB})$	$\frac{1}{bn}(MS_A - MS_W)$
$\hat{\sigma}^2_B$	$\frac{1}{an}(MS_B - MS_{AB})$	$\frac{df_B}{abn}(MS_B - MS_{AB})$
$\hat{\sigma}^2_{AB}$	$\frac{1}{n}(MS_{AB} - MS_W)$	$\frac{1}{n}(MS_{AB} - MS_W)$

Note. In all cases, $\hat{\sigma}^2_{error} = MS_W$, and $\hat{\sigma}^2_{total}$ is the sum of $\hat{\sigma}^2_A$, $\hat{\sigma}^2_B$, $\hat{\sigma}^2_{AB}$, and $\hat{\sigma}^2_{error}$.

If some factors are random, it is actually easier to work with equations for the variance components that contribute to inferential measures of association. Listed in Table 8.8 are ANOVA-based variance components estimators for completely between-subjects two-way designs where either both factors are random or factor *A* is random but factor *B* is fixed. To calculate the desired measure of association, one just computes the appropriate estimators using the equations in Table 8.8 and then assembles them in the correct way.

Suppose that both factors are random in a completely between-subjects 3×6 design, where $n = 5$ and

$$MS_A = 48.00,\ MS_B = 40.00,\ MS_{AB} = 6.00, \text{ and } MS_W = 4.00$$

When the equations in Table 8.8 are used, the variance components estimators are

$$\hat{\sigma}^2_A = \frac{1}{6(5)}(48.00 - 6.00) = 1.400 \quad \text{and} \quad \hat{\sigma}^2_B = \frac{1}{3(5)}(40.00 - 6.00) = 2.267$$

$$\hat{\sigma}^2_{AB} = \frac{1}{5}(6.00 - 4.00) = .400 \quad \text{and} \quad \hat{\sigma}^2_{error} = 4.000$$

$$\hat{\sigma}^2_{total} = 1.400 + 2.267 + .400 + 4.00 = 8.067$$

for the interaction

$$\hat{\rho}_{I,AB} = \frac{.40}{8.067} = .050 \quad \text{and} \quad \text{partial } \hat{\rho}_{I,AB} = \frac{.40}{.40 + 4.00} = .091$$

In words, the *AB* effect explains about 5.0% of the total variance and about 9.1% of the variance controlling for the main effects. Exercise 6 asks you to calculate $\hat{\omega}^2_{AB}$ and partial $\hat{\omega}^2_{AB}$ for the same data but assuming that factor *B* is fixed.

Space limitations preclude listing equations for variance components estimators in larger factorial designs with dependent samples, but Vaughan and Corballis (1969) is a good source. Computer procedures for factorial ANOVA are gradually getting better at reporting measures of association in complex designs. Variance component estimation with maximum likelihood methods is an alternative, but large samples are needed. Olejnik and Algina (2003) described **generalized estimated omega-squared** for balanced designs with fixed factors. It has the same form as generalized $\hat{\eta}^2$ except that generalized $\hat{\omega}^2$ is based on variance component estimators, not on sums of squares. Both control for covariates and whether the factors are extrinsic versus intrinsic or between-subjects versus within-subjects. Generalized $\hat{\omega}^2$ also controls for positive bias in generalized $\hat{\eta}^2$.

Interval Estimation

Smithson's (2003) scripts for SPSS, SAS/STAT, and R calculate noncentral confidence intervals for η^2 based on the total effects and for partial η^2 based on individual main or interaction effects in completely between-subjects factorial designs (see footnotes 3–4, Chapter 5). Fidler and Thompson (2001) gave SPSS scripts for calculating noncentral confidence intervals for ω^2 or partial ω^2 in completely between-subjects factorial designs; see also W. H. Finch and French (2012), who compared different methods of interval estimation for ω^2 in two-way designs with independent samples. Sahai, Khurshid, Ojeda, and Velasco (2009) discussed interval estimation for population variance components in balanced designs with two random factors.

EXTENSIONS TO MULTIVARIATE ANALYSES

There are multivariate versions of *d* statistics and measures of association for designs with two or more continuous outcomes. For example, a Mahalanobis distance is a multivariate *d* statistic, and it estimates the difference between two group centroids (the sets of all univariate means) in standard deviation units controlling for intercorrelation. Multivariate measures of association also control for correlated outcomes; see Grissom and Kim (2011, Chapter 12) and Olejnik and Algina (2000) for more information.

RESEARCH EXAMPLES

Two examples of effect size estimation in actual factorial designs are described next.

Differential Effectiveness of Aftercare Programs for Substance Abuse

You can download the raw data for this example in SPSS format from the web page for this book. T. G. Brown, Seraganian, Tremblay, and Annis (2002) randomly assigned 87 men and 42 women who had just been discharged from residential treatment centers for substance abuse to one of two different 10-week aftercare programs, structured relapse prevention (SRP) and 12-step facilitation (TSF). The former stressed rehearsal of skills to avoid relapse, and the latter emphasized traditional methods of Alcoholics Anonymous. Reported in the top part of Table 8.9 for this 2×2 randomized blocks design are descriptive statistics for a measure of the severity of alcohol-related problems administered 6 months later where higher scores indicate more problems. The interaction is disordinal and is illustrated in Figure 8.2: Women who completed the SRP program have relatively worse outcomes than women who completed the TSF program, but men had similar outcomes regardless of aftercare program type.

Presented in the bottom part of Table 8.9 are the source table and values of standardized contrasts for single-factor effects, including the simple effects of aftercare program for each gender. The sums of squares are Type I, and the rationale for their selection in this nonorthogonal design is as follows: Men have more problems with alcohol than women, so the gender main effect (G) was not adjusted for other effects. It was less certain whether the aftercare program (P) would make any difference, so its effect was adjusted for gender. The GLM procedure of SPSS controlled through its graphical user interface does not offer an option for calculating sums of squares for user-defined simple effects, but there is an alternative.

In brief, it is possible to control SPSS by writing text-based syntax that specifies the data and analysis options. One uses the syntax editor in SPSS to write and edit the commands, and the resulting syntax file is saved with the extension .sps (for SPSS syntax). The syntax is executed by highlighting (selecting) it with the mouse cursor and then clicking on the "run" icon, which resembles the icon for "play" in a media player application. Knowing something about SPSS syntax gives the user access to capabilities that are not available through the graphical user interface of the program. For this example, the SPSS syntax listed next requests sums of squares for

TABLE 8.9
Descriptive Statistics, Analysis of Variance Results, and Effect Sizes for Severity of Alcohol-Related Problems by Gender and Aftercare Program Type

		Aftercare program		
	n	TSF	SRP	Row means
Women	42	10.54 (15.62)[a] 23[b]	27.91 (21.50) 19	18.40
Men	87	17.90 (20.12) 48	16.95 (21.55) 39	17.47
Column means		15.52	20.54	17.77

Source	*SS*	*df*	*MS*	*F*	*d*
Total effects	3,181.33	3	1,060.44	2.63[c]	—
Gender	24.37	1	24.37	<1.00	.05[g]
Program	804.28	1	804.28	2.00[d]	−.25[h]
Gender × program	2,352.69	1	2,352.69	5.84[e]	—
Simple effects of program					
Program at women	3,137.53	1	3,137.53	7.79[f]	−.85[h]
Program at men	19.43	1	19.43	<1.00	.05[h]
Within-cells (error)	50,367.22	125	402.94		
Total	53,548.55	128			

Note. These data are from T. Brown (personal communication, January 23, 2012) and are used with permission. TSF = 12-step facilitation; SRP = structured relapse prevention. [a]Cell mean (standard deviation). [b]Cell size. [c]$p = .053$. [d]$p = .160$. [e]$p = .017$. [f]$p = .006$. [g]Standardizer is $s_w = 20.07$. [h]Standardizer is $s_{w,\,G,\,GP} = 20.38$.

simple effects of aftercare program at gender and a graphical display of the interaction:

```
glm alcohol by gender program/method = sstype(1)/
  emmeans = tables (gender * program) compare(program)/
  plot = profile (gender * program).
```

Gender is intrinsic, but the aftercare program factor is assumed to be extrinsic. This is because there are far more traditional, 12-step aftercare programs than there are programs based on principles of behavior therapy. Thus, the appropriate standardizer for the gender main effect is the square root of $MS_W = 402.94$, or 20.07, which does not reflect variation due to program

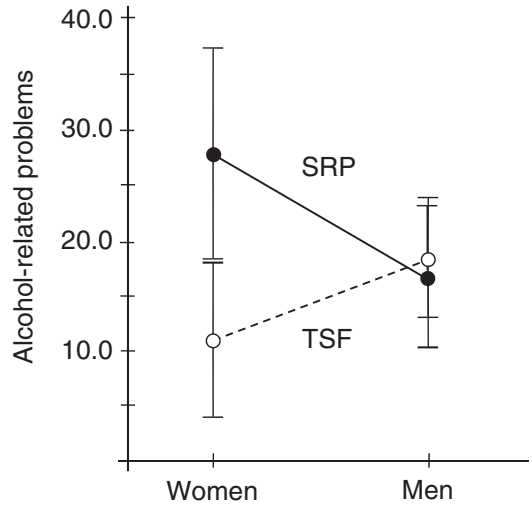

Figure 8.2. Cell means and 95% confidence intervals for μ for the data in Table 8.9. SRP = structured relapse prevention; TSF = 12-step facilitation.

type. Given the marginal means in Table 8.9, women reported more alcohol-related problems at follow-up than men did by about .05 standard deviations regardless of program type.

Because gender varies naturally, the standardizer for main or simple effects of program type is the square root of

$$MS_{W,G,GP} = \frac{SS_W + SS_G + SS_{GP}}{df_W + df_G + df_{GP}} = \frac{52{,}744.28}{127} = 415.31$$

or 20.38. Based on the marginal means in Table 8.9, the average difference between the two aftercare programs is –.25 standard deviations in favor of the TSF program. But this result is uninformative due to the presence of disordinal interaction. Standardized contrasts for the simple effects of program type are

$$d_{P \text{ at women}} = \frac{10.54 - 27.91}{20.38} = -.85 \quad \text{and} \quad d_{P \text{ at men}} = \frac{17.90 - 16.95}{20.38} = .05$$

These results say that women in the SRP program reported more alcohol-related problems than women in the TSF program did by about 85% of a standard deviation. The magnitude of the corresponding difference for men was only 5% in standard deviation units, but men did somewhat better in the TSF program than in the SRP program. The difference between the two standardized simple effects is a standardized interaction contrast, or

$$d_{P \text{ at women}} - d_{P \text{ at men}} = -.85 - .05 = -.90$$

That is, gender difference in the effect of aftercare program type is almost a full standard deviation in magnitude.

Earwitness Testimony and Moderation of the Face Overshadowing Effect

The face overshadowing effect (FOE) happens when identification of the once-heard voice of a stranger in conditions that resemble a police lineup is worse if the speaker's face is seen at the time of exposure. Cook and Wilding (2001) evaluated whether the FOE is affected by hearing the voice more than once or by explicit instructions to attend to the voice instead of the face. In total, 216 young adults were randomly assigned to one of eight conditions in this balanced $2 \times 2 \times 2$ experimental design where the fixed factors are face (present or absent), voice repetition (once or three times), and instruction (intentional, specifically told to focus on the voice; incidental, no specific instructions given). All participants heard two different voices, one a man's and the other a woman's, say two different sentences. One week later the participants were asked to pick each voice out of separate gender voice lineups. The outcome variable was the number of correct identifications. Before analyzing these data, I multiplied the scores by the constant 10.00 in order to avoid very small sums of squares. This change does not affect values of *F, p,* or the effect sizes reported next.

Listed in the top part of Table 8.10 are cell descriptive statistics, and the source table is reported in the bottom part. For the total effects, $\hat{\eta}^2 = .117$, 95% CI [.026, .175], so the main and interactive effects together explain about 11.7% of the total observed variance in correct identifications. The repetition main effect is the best individual predictor, partial $\hat{\eta}^2 = .090$, 95% CI [.030, .169]. As expected, there are more correct identifications when the voice is heard three times than when it is heard just once. The main effect of the face–no face factor was the second best predictor, partial $\hat{\eta}^2 = .021$, 95% CI [.004, .073], and the marginal mean is indeed higher when the face is not present (10.85) than when the face is present (8.90).

Observed proportions of residual variance explained by the remaining effects are close to zero except for the interaction between the face (*F*) and repetition (*R*) factor, partial $\hat{\eta}^2 = .011$, 95% CI [.004, .054]. Means on the outcome variable for this two-way interaction averaged over the instruction factor are

	Repeat 1×	Repeat 3×
No face	9.45	12.25
Face	6.10	11.70

TABLE 8.10
Descriptive Statistics, Analysis of Variance Results, and Effect Sizes for Accuracy of Voice Recognition by Instruction, Repetition, and Presence Versus Absence of the Speaker's Face

	Instruction	
Condition	Incidental	Intentional
Voice once	9.60 (6.50)[a]	9.30 (6.80)
Voice three times	12.60 (7.60)	11.90 (6.20)
Voice once + face	5.90 (6.40)	6.30 (6.90)
Voice three times + face	11.50 (7.20)	11.90 (6.80)

Source	*SS*	*df*	*MS*	*F*	Partial $\hat{\eta}^2$
Total effects	1,275.90	7	182.27	3.93[b]	.117 [.026, .175][e]
Instruction (*I*)	.14	1	.14	<1.00	<.001
Face (*F*)	205.34	1	205.34	4.42[c]	.021 [.004, .073]
Repetition (*R*)	952.56	1	952.56	20.52[b]	.090 [.030, .169]
$I \times F$	10.94	1	10.94	<1.00	<.001
$I \times R$	.54	1	.54	<1.00	<.001
$F \times R$	105.84	1	105.84	2.28[d]	.011 [.004, .054]
$I \times F \times R$	.54	1	.54	<1.00	<.001
Within-cells (error)	9,654.73	208	46.42		
Total	10,930.63	215			

Note. Cell descriptive statistics are from "Earwitness Testimony: Effects of Exposure and Attention on the Face Overshadowing Effect," by S. Cook and J. Wilding, 2001, *British Journal of Psychology, 92,* p. 621. Copyright 2001 by John Wiley and Sons. Reprinted with permission. Partial $\hat{\eta}^2$ for the total effects = $\hat{\eta}^2$ for those same effects.
[a]Cell mean (standard deviation); $n = 27$ for all cells. [b]$p < .001$. [c]$p = .037$. [d]$p = .133$.
[e]Noncentral 95% confidence interval reported in brackets for effect sizes > .001.

We can see in this matrix that the size of the FOE is greater when the voice is heard just once instead of three times. The unstandardized interaction contrast based on these cell means is

$$\hat{\psi}_{FR} = 9.45 - 12.25 - 6.10 + 11.70 = 2.80$$

Assuming that none of factors vary naturally, standardizing this contrast against the square root of $MS_W = 46.42$ for the whole design gives us

$$d_{\hat{\psi}_{FR}} = \frac{2.80}{\sqrt{46.42}} = .41$$

Thus, the magnitude of the FOE is .41 standard deviations larger given one repetition of the voice than it is given three repetitions. Because intentional versus incidental instruction does not appreciably moderate the two-way interaction just analyzed, Cook and Wilding (2001) attributed the FOE to an involuntary preference for processing face information that is not overcome on hearing an unfamiliar voice just once.

CONCLUSION

Estimation of the magnitudes and precisions of interaction effects should be the focus of the analysis in factorial designs. Methods to calculate standardized mean differences for contrasts in such designs are not as well developed as those for one-way designs. Standardizers for single-factor contrasts should reflect variability as a result of intrinsic off-factors that vary naturally in the population, but variability due to extrinsic off-factors that do not vary naturally should be excluded. Measures of association may be preferred in designs with three or more factors or where some factors are random. They can also evaluate the predictive power of several effects analyzed together. Characteristics of designs with fixed factors can affect values of $\hat{\eta}^2$ and $\hat{\omega}^2$, but there are generalized forms of both effect sizes that control for covariates and whether factors are extrinsic versus intrinsic or between-subjects versus within-subjects. The intraclass correlation $\hat{\rho}_I$ can be calculated for effects of random factors.

LEARN MORE

Montgomery, Peters, and Little (2003) give suggestions for reporting the results of factorial analyses. Olejnik and Algina (2000) review effect sizes for factorial designs, and Pierce et al. (2004) caution about the failure to distinguish between $\hat{\eta}^2$ and partial $\hat{\eta}^2$.

Montgomery, A. A., Peters, T. J., & Little, P. (2003). Design, analysis and presentation of factorial randomised controlled trials. *BMC Medical Research Methodology, 3,* Article 26. doi:10.1186/1471-2288-3-26

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25,* 241–286. doi:10.1006/ceps.2000.1040

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement, 64,* 916–924. doi:10.1177/0013164404264848

EXERCISES

1. Calculate sums of squares for the main and interaction effects and for the simple effects of drug at gender for the data in the left side of Table 8.3 for $n = 10$ and $MS_W = 125.00$.
2. Show that the interaction contrasts in **(I)** and **(II)** are orthogonal in a balanced design.
3. Show that the sum of squares for the omnibus interaction in the completely between-subjects analysis at the top of Table 8.5 can be uniquely broken down in the sums of squares for the interaction contrasts specified in **(I)** and **(II)**.
4. Cell means for a balanced $2 \times 2 \times 2$ design are presented as follows. Show that Equation 8.10 holds for these data:

	C_1		C_2	
	B_1	B_2	B_1	B_2
A_1	15.00	14.00	17.00	10.00
A_2	10.00	8.00	18.00	10.00

5. For the completely between-subjects analysis in the top part of Table 8.5, show for this balanced design that the reduced cross-classification method generates $MS_{W, B, AB} = 12.25$ based on the cell descriptive statistics in Table 8.4 using Equation 8.16.
6. Given $MS_A = 48.00$, $MS_B = 40.00$, $MS_{AB} = 6.00$, $MS_W = 4.00$, $a = 3$, $b = 6$, and $n = 5$, $\hat{\rho}_I = .050$ and partial $\hat{\rho}_I = .091$ for the interaction effect assuming that both factors are random. Recalculate proportions of explained variance assuming that factor B is fixed.

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank

9

REPLICATION AND META-ANALYSIS

> For life is not a tournament. Its race is not always to the swift nor its battle to the strong. What counts is enduring to the end.
>
> —Gilbert Meilaender (2011, p. 20)

Replication is a foundational scientific activity but one neglected in the behavioral sciences. This is a paradox: Most behavioral researchers along with their colleagues in the natural sciences would probably endorse replication as a gold standard. Replication is common in the natural sciences, but it is hard to find studies in our own literature conducted specifically as replications. There are also obstacles in the behavioral sciences in the form of disincentives and outright biases against replication research. These cultural factors discourage genuine appreciation for replication. If the behavioral sciences are ever to mature out of their extended adolescence, this neglect of replication must end. Considered next are basic kinds of replication, attitudes and policies that work against replication, and the role of meta-analysis. A key point is that meta-analysis is not a substitute for systematic replication.

DOI: 10.1037/14136-009
Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, Second Edition, by R. B. Kline

CONCEPTS ABOUT REPLICATION

Basic definitions and concepts about replication are introduced in this section.

Theoretical and Empirical Cumulativeness

The historian and philosopher Thomas S. Kuhn (1996) described science as alternating between two basic states. One is the steady state of **normal science**, characterized by a high level of paradigm development. A **paradigm** is a shared set of theoretical structures, methods, and definitions that supports the essential activity of normal science, **puzzle solving**, the posing and working out of problems. If a paradigm's empirical and theoretical structures build on one another in a way that permits results of current research to extend earlier work, it provides **theoretical cumulativeness** (Hedges, 1987). The second state involves crises that arise when certain persistent problems, or **anomalies**, cannot be solved under the current paradigm. These crises lead to challenges by scholars who may be younger or who have backgrounds in different fields than those who defend the current paradigm. A scientific revolution occurs when the old paradigm is replaced by a new one. The assumptions of the new paradigm may be so different that the subsequent course of the discipline is radically altered. It is normal science that concerns us, in particular its cumulative nature through replication and synthesis of results.

That the behavioral or "soft" sciences lack a true scientific paradigm is a matter of debate. Note that our use of a more-or-less common set of statistical techniques does not by itself constitute a paradigm. The use of common tools is only a small part of a paradigm. The rest involves shared assumptions and methods that together identify the main problems of interest and how to go about solving them. There is little agreement in the behavioral sciences about just what the main problems are and exactly how to study them. This basic disagreement reflects our preparadigmatic (i.e., prescientific) state.

Another requirement for a cumulative science is **empirical cumulativeness**, or the extent of agreement of replicated results and whether such results fall into patterns that make sense (Hedges, 1987). Perhaps behavioral research results are simply less replicable than those in the natural sciences. This may be especially true in studies with human participants. In animal studies, the subjects can literally be transported to and from the experimental situation at the behest of the researcher, but those who conduct human studies know all too well that they are "studying complex organisms that do have moods, can be generally uncooperative, and are known to evidence behavioral inconsistencies over time" (Easley, Madden, & Dunn, 2000, p. 84). Human research participants miss scheduled appointments, try to guess the

purpose of the study when they do show up, react to the knowledge that they are being measured, and sometimes fail to complete all requested forms or withdraw from the study altogether.

When physical scientists study things like neutrons and protons and observe how neutrons and protons react in each other's presence, they do not have to qualify their results by saying "generally," "for most neutrons," or "only for neutrons with good nutrition during proton gestation." Social scientists study people, who by their nature are idiosyncratically individual. The uniqueness of every person is what makes people so interesting, but it is also what makes generalizing about people so daunting a prospect. In this sense the very subject matter of behavioral research may be more complex than many phenomena studied in the natural sciences (Lykken, 1991).

There are also strong familial or social context effects for many aspects of human behavior. This is another way to describe interaction, which concerns associations that change over situations or cultures. Context effects can also be era dependent. For example, social conditions concerning the status of women changed dramatically over the last few decades, and certain effects of disparate treatment of women versus men are different today than in the past. Probably as a result of improved access to educational resources by women, gender differences in math skills may have narrowed over the 1960s–1980s to the point where little substantive difference may now exist (Lindberg, Hyde, Petersen, & Linn, 2010). There is also evidence that international variation in gender differences at the highest levels of mathematics achievement is related to inequality in the labor market and differences in the social status of women and men (Penner, 2008). Behavioral researchers seem to underestimate the impact of both sampling error and context effects on their results (Shadish et al., 2001).

Hedges (1987) evaluated whether empirical cumulativeness may be inherently lower for behavioral data than for natural science data. He estimated the consistency of results in physics research about the mass and lifetime of stable particles, such as neutrons or protons, with the consistency of results in the "hard" area of gender differences in cognitive abilities and the "soft" area of effects of educational programs on achievement. Surprisingly, he found similar degrees of consistency in the physics and behavioral research areas just mentioned as measured by a standard index of between-studies variability in meta-analysis (the *Q* statistic, described later). These findings suggested that physical science data may not be inherently more empirically cumulative than behavioral science data, at least in the domains studied by Hedges (1987).

Types of Replication

There is no single nomenclature to classify replication studies (e.g., Easley et al., 2000), but there is enough consensus to outline at least the broad types

described next. B. Thompson (1997) distinguished between internal and external replication. **Internal replication** includes statistical resampling and cross-validation by the original researcher(s). Resampling includes bootstrapping and related computer-based methods, such as the jackknife technique, that randomly combine the cases in an original data set in different ways to estimate the effect of idiosyncrasies in the sample on the results (e.g., Figure 2.5). Such procedures are not replication in the usual scientific sense. The total sample in **cross-validation** is randomly divided into a **derivation sample** and a **cross-validation sample**, and the same analyses are conducted in each one. **External replication** is conducted by people other than the original researchers, and it involves new samples collected at different times or places.

There are two broad contexts for external replication. The first concerns different kinds of replications of experimental studies. One is **exact replication**, also known as **direct replication**, **literal replication**, or **precise replication**, where all major aspects of an original study—its sampling methods, design, and outcome measures—are closely copied. True exact replications exist more in theory than in practice because it is difficult to perfectly duplicate a study, especially when human factors among participants and researchers inevitably vary over time, settings, and samples. Other sources of variation include differences in equipment or procedures across laboratories. Another type is **operational replication**—also referred to as **partial replication** or **improvisational replication**—where just the sampling and methods of an original study are duplicated. Operational replication tests whether a result can be found by a researcher who follows just the basic "recipe" in the Method section of an original study. The outcome of operational replication is potentially more informative than that of literal replication, because robust effects should stand out against variations in procedures, settings, or samples.

In **balanced replication**, operational replications are used as control conditions. Other conditions may represent the manipulation of additional substantive variables to test new hypotheses. For example, a drug condition from an original study could be replicated in a new study. Additional conditions in the latter may feature the administration of the same drug at different dosages, other kinds of drugs, or a different type of treatment. The logic of balanced replication is similar to that of strong inference, which features designing studies to rule out competing explanations, and to that of **dismantling research**. The aim of the latter is to study elements of treatments with multiple components in smaller combinations to find the ones responsible for treatment efficacy.

A researcher who conducts a **construct replication** or **conceptual replication** avoids close imitation of the specific methods of an original study. An ideal construct replication would be carried out by telling a skilled researcher

little more than the original empirical result. This researcher would then specify the design, measures, and data analysis methods deemed appropriate to test whether a finding has generality beyond the particular situation studied in an original work. This provides an even more demanding test of the robustness of some finding. But it is possible that the nature of the phenomenon could actually change depending on how it is measured, the particular sample studied, or the specific experimental method used. Without a systematic cataloging of how construct replications differ from each other, it may be difficult to associate study characteristics with observed changes in the effect (if any). Meta-analysis can partially fill this role, as discussed below.

A second context for replication concerns psychometrics, which seems to have a stronger tradition of replication. This may be in part a result of professional standards that outline benchmarks for establishing score validity, such as *Standards for Educational and Psychological Testing,* developed jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999). The demonstration of construct validity requires more than one line of evidence, which includes the use of multiple methods among other variations in assessment procedures. There is also an appreciation of the need to cross-validate tests that generate scores based on mathematically weighted combinations of predictor variables. These weights—usually regression coefficients—are susceptible to capitalization on chance. It is thus necessary to determine whether their values are observed in other samples.

REPLICATION IN THE BEHAVIORAL SCIENCES

There is evidence that only small proportions—in some cases < 1%—of all published studies in the behavioral sciences are specifically described as replications (e.g., Easley et al., 2000; Kmetz, 2002). Some possible reasons are listed next:

1. *Misinterpretation of statistical significance.* Many widespread false beliefs about the meaning of statistical significance undoubtedly discourage replication. Among the obvious suspects are the replicability, odds-against-chance, inverse probability, and valid research hypothesis fallacies. The combined effect of cognitive distortions about *p* values could lead researchers to be so overconfident about their results that replication is seen as unnecessary.
2. *Editorial preference for novelty.* It is easy to see the clear preference among journal editors and reviewers for work characterized as *original;* that is, providing new theoretical, methodological,

or substantive contributions to the field. From this perspective, replication studies may be seen as derivative rehashes of old ideas and the researchers who conduct them as uncreative, scholarly dullards who can imitate but not innovate. About 95% of the editors of behavioral science research journals surveyed by Neuliep and Crandall (1990) said that replication studies were not explicitly encouraged for submission in editorial policy. Most journal reviewers also prefer original studies over replications (Neuliep & Crandall, 1993). If researchers correctly perceive that the odds are slim for getting replication studies published, it is no wonder that they would shy away from conducting them.

3. *Other disincentives for replication.* Most graduate programs require that dissertation research should make contributions to knowledge resulting from original and independent research. These requirements do not explicitly rule out replication, but doctoral students may be dissuaded from conducting such studies due to the novelty requirement. Faculty members are also aware they may be evaluated less positively if their research portfolios are weighted toward replication.

All of these factors combine to create a kind of cultural bias against replication in the behavioral sciences. These pressures also favor works in which new theory is generated over those that develop or refine theory. This explains why our research literature is awash with unsubstantiated claims based on one-shot studies about a myriad of "new" theories, most of which have the staying power of castles in the sand. But this cogent recommendation would send a powerful message to authors of empirical studies: "The publishing of a paper that has relied on results from a single study (and, thus, has not been replicated) should be unacceptable to the discipline because of the inherent variability in human subjects" (Easley et al., 2000, p. 89).

K. Hunt (1975), S. Schmidt (2009), and others have argued that most replication in the behavioral sciences occurs covertly in the form of **follow-up studies**, which combine direct replication (or at least construct replication) with new procedures, measures, or hypotheses in the same investigation. Such studies may be described by their authors as "extensions" of previous works with new elements but not as "replications," probably to avoid the stigma of replication with journal editors and reviewers. The problem with this informal approach to replication is that it is not explicit and therefore is unsystematic. Authors of follow-up studies may not document all the ways in which their investigation differed from those of previous studies

in the same area. Suppose that results in a follow-up study based loosely on a prior study disagree with those in the original work. Now, what does this outcome mean? One possibility is that the original finding is not robust over minor variation in participants, methods, or settings. If those variations are major, though, the results may not be directly comparable over the original and new studies.

The issue just described is the **apples and oranges problem**, well known in the meta-analytic literature, and it refers to doubt concerning whether it is reasonable to directly compare results from different studies. Although there are ways in meta-analysis to address this problem, they are not magic, especially when the meta-analyst must infer the factors that account for variation in results over studies not conducted as explicit replications. This is why meta-analysis could never cure the replication deficit in the behavioral sciences. But meta-analysis is superior to old-fashioned, narrative literature reviews, especially ones based on the **box-score (vote counting) method**, where tallies of the numbers and directions of null hypothesis rejections over a set of studies determined the conclusion. That effect sizes are synthesized in most meta-analyses also reminds us of the importance of this aspect of describing results.

Perhaps replication would be more highly valued if confidence intervals were reported more often. Then readers of empirical articles would be able to see the low precision with which many studies are conducted. Widths of confidence intervals for behavioral data are often, to quote Cohen (1994, p. 999), "so embarrassingly large!" (see also Cumming, 2012). Wide confidence intervals indicate that a study contains only limited information, a fact that is concealed when only results of statistical tests are reported (F. L. Schmidt & Hunter, 1997).

META-ANALYSIS

One cannot deny that meta-analysis is an important and widely used technique for research synthesis. Since the publication of the first modern meta-analysis—the classic Smith and Glass (1977) study of psychotherapy outcomes measured with standardized mean differences—thousands of meta-analytic articles have been published in psychology, psychiatry, education, and medicine, among other disciplines. There are also introductions to meta-analysis in areas such as behavioral medicine (Nestoriuc, Kriston, & Rief, 2010) and criminal justice (Pratt, 2010), plus many books that introduce meta-analysis to wider audiences (e.g., Card, 2012). Next, I emphasize aspects of meta-analysis that highlight how effect sizes are synthesized and limitations of the technique.

Predictors

If a set of studies is made up of exact replications, there may be little quantitative analysis to do other than estimate the central tendency and variability of the results. The former could be seen as a better estimate of the population parameter than the result in any one study, and the latter could be used to identify individual results that are outliers. Because exact replications are inherently similar, outliers may be more a result of chance than systematic differences among studies. This is less certain for operational or construct replications and even less so for follow-up studies. For the latter, observed variability in results may reflect actual changes in the effect due to differences in samples, measures, or designs over studies.

Because sets of related investigations in the behavioral sciences are generally made up of follow-up studies, the explanation of observed variability in their results is a common goal in meta-analysis. That is, the meta-analyst tries to identify and measure characteristics of follow-up studies that give rise to variability among the results. These characteristics include attributes of samples (e.g., mean age, gender), settings in which cases are tested (e.g., inpatient vs. outpatient), and the type of treatment administered (e.g., duration, dosage). Other factors concern properties of the outcome measures (e.g., self-report vs. observational), quality of the research design, source of funding (e.g., private vs. public), professional backgrounds of the authors, or date of publication. The last reflects the potential impact of temporal factors such as changing societal attitudes. These characteristics can be classified as low versus high inference. A **low-inference characteristic** is one that is readily apparent in the text or tables of a primary study, such as the measurement method. In contrast, a **high-inference characteristic** requires a judgment. The quality of the research design is an example of a high-inference characteristic because it must be inferred from the information reported in the study.

Study factors are conceptualized as meta-analytic predictors, and study outcome measured with the same standardized effect size is typically the criterion. Each predictor is actually a moderator variable, which implies interaction. This is because the criterion, study effect size, usually represents the association between the independent and dependent variables. If observed variation in effect sizes across a set of studies is explained by a meta-analytic predictor, the relation between the independent and dependent variables changes across the levels of that predictor. For the same reason, the terms **moderator variable analysis** and **meta-regression** describe the process of estimating whether study characteristics explain variability in results. The latter term is especially appropriate because study factors can covary, such as when different variations of a treatment tend to be administered to patients with

acute versus chronic forms of a disorder. If meta-analytic predictors covary, it is necessary to control for overlapping explained proportions of variability in effect sizes.

It is also possible for meta-analytic predictors to interact, which means that they have a joint influence on observed effect sizes. Interaction also implies that to understand variability in results, one must consider the predictors together. This is a subtle point, one that requires some elaboration: Each individual predictor in meta-analysis is a moderator variable. But the relation of one meta-analytic predictor to study outcome may depend on another predictor. For example, the effect of treatment type on observed effect sizes may depend on whether cases with mild versus severe forms of an illness were studied.

A different kind of phenomenon is mediation, or indirect effects among study factors. Suppose that one factor is degree of early exposure to a toxic agent and another is illness chronicity. The exposure factor may affect study outcome both directly and indirectly through its influence on chronicity. Indirect effects can be estimated in meta-analysis by applying techniques from structural equation modeling to covariance matrices of study factors and effect sizes pooled over related studies. The use of both techniques together is called **mediational meta-analysis** or **model-driven meta-analysis**. Estimation of mediation requires specific a priori hypotheses about patterns of direct or indirect effects of study factors on effect sizes.

Steps

The basic steps in meta-analysis are similar to those in a primary study. They may be iterative in both because it is often necessary to return to an earlier step for refinement when problems are discovered at later stages. They are listed next:

1. Formulate the research question.
2. Collect the data (primary studies).
3. Evaluate the quality of the data (i.e., study design, procedures, and measurement).
4. Identify and measure the predictors (study factors) and criterion (effect sizes).
5. Analyze the data (synthesize effect sizes).
6. Describe, interpret, and report the results.

Because the steps just listed are described in many published introductions to meta-analysis, I will elaborate on just a few critical issues. It is just as important in meta-analysis as when conducting a primary study to clearly specify the hypotheses and operational definitions of constructs.

These specifications in meta-analysis should also help to distinguish between relevant and irrelevant studies. A meta-analysis obviously requires that research about a topic exists, which raises the question of how many studies are necessary. A researcher can use meta-analytic methods to synthesize as few as two studies, but more are typically needed. Although there is no absolute minimum, it seems to me that at least 20 different primary studies would be required before a meta-analysis is feasible. This assumes that the studies are relatively homogeneous and that only a small number of moderator variables are analyzed. The failure to find sufficient numbers of studies indicates a knowledge gap.

Data collection is characterized by computer searches in multiple sources including published works, such as articles, books, or reports from public agencies, and unpublished studies. The latter include conference presentations, papers submitted for publication but rejected, student theses, and technical reports from private agencies. There are computer databases for some unpublished kinds of studies, such as doctoral dissertations, but other kinds of unpublished works are not always stored in accessible databases or even available at all through the Internet. This makes them harder to find.

A related issue is the **file drawer problem**, which is that some studies may be conducted but never reported, and results from unreported studies could differ on average from results that are reported. There are ways to estimate in meta-analysis what is known as the **fail-safe N**, which is the number of additional studies where the average effect size is zero that would be needed to increase the *p* value in a meta-analysis for the test of the mean observed effect size to > .05 (i.e., the nil hypothesis is not rejected). These additional studies are assumed to be file drawer studies or to be otherwise not found in the literature search of a meta-analysis. If the estimated number of such studies is so large that it is unlikely that so many studies (e.g., 2,000) with a mean nil effect size could exist, more confidence in the results may be warranted. But estimates of fail-safe *N* are just that despite what is implied by their name.

Studies from each source are subject to different types of biases. For example, bias for statistical significance implies that published studies have more H_0 rejections and larger effect sizes than do unpublished studies (e.g., Table 2.1). There are techniques in meta-analysis for estimating the extent of publication bias (e.g., Gilbody, Song, Eastwood, & Sutton, 2000). If such bias is indicated, a meta-analysis based mainly on published sources may be inappropriate. Results from unpublished studies may be prone to distortion because of design or analysis problems that otherwise may have been detected in peer review, but coding of study source as a meta-analytic predictor permits direct evaluation of its effect on study outcome.

For two reasons, it is crucial to assess the high-inference characteristic of research quality for each found primary study. The first is to eliminate from further consideration studies so flawed that their results are untrustworthy. This helps to avoid the **garbage in, garbage out problem**, where results from bad studies are synthesized along with those from sound studies. The other reason concerns the remaining (nonexcluded) studies, which may be divided into those that are well designed versus those with significant limitations. Results synthesized from the former group may be given greater weight in the analysis than those from the latter group. There are some standard systems for coding quality of primary studies (e.g., Conn & Rantz, 2003). Nowadays it should be standard practice for meta-analysts to describe how they evaluated the research designs in found studies and specify the criteria used to retain or reject these studies from further analysis. Relatively high proportions of found studies in meta-analyses are often discarded due to poor rated quality, a sad comment on the status of a research literature.

The computation of standardized effect sizes based on descriptive or test statistics reported in a set of primary studies is the main way to convert all findings to a common metric. But if very different types of outcome measures are used across the studies, their results may not be directly comparable even if the same kind of standardized effect size is computed for each one. This is the apples and oranges problem concerning effect sizes, which are the data points in meta-analysis. Suppose that gender differences in aggression are estimated over a series of studies. There is more than one type of aggressive behavior (e.g., verbal, physical) and more than one way to measure it (e.g., self-report, observational). An average *d* statistic that compares men and women and is computed across a diverse set of aggression measures may not be very meaningful. A way to deal with this problem is to code the content or measurement method of the outcome variable and represent this information in the analysis as one or more study factors, but doing so requires that the meta-analyst knows to make this distinction.

It is common in meta-analysis to weight the standardized effect sizes by a factor that represents sample size and error variance. This gives greater weight to results based on larger samples, which are less subject to sampling error. It is also possible to weight effect sizes by other characteristics, such as score reliability (see Chapter 5). Hunter and Schmidt (2004) described an extensive set of corrections for attenuation in effect sizes for problems such as artificial dichotomization in continuous outcome variables and range restriction, but primary studies do not always report sufficient information for one to apply these corrections.

There is also the problem of **correlated effect sizes**, or nonindependence of study results. It seldom happens that each individual result comes from an independent study where a single hypothesis is tested. In some studies,

the same research participants may be tested with multiple outcome measures. If these measures are intercorrelated, effect sizes across these measures are not independent. Likewise, effect sizes for the comparison of variations of a treatment against a common control group are probably not independent. Fortunately, statistical techniques that handle correlated effect sizes are available.

Synthesizing Effect Sizes

Summarized next are the basic iterative phases of effect size synthesis in meta-analysis:

1. Decide whether to combine results across studies and what to combine.
2. Estimate a common (average) effect size.
3. Estimate the heterogeneity in effect sizes across studies, and attempt to explain it—that is, find an appropriate statistical model for the data.
4. Assess the potential for bias.

The first step is often the computation of a weighted average effect size. If it can be assumed that the observed effect sizes estimate a single population effect size—that is, a fixed effects model—their average takes the form

$$M_{ES} = \frac{\sum_{i=1}^{k} w_i ES_i}{\sum_{i=1}^{k} w_i} \tag{9.1}$$

where ES_i is the effect size (e.g., d) for the ith result in a set of k effect sizes and w_i is the weight for that result. A weight for each effect size that minimizes the variance of M_{ES} is

$$w_i = \frac{1}{s^2_{ES_i}} \tag{9.2}$$

where the denominator is the conditional variance (squared standard error) of an effect size, or the **within-studies variance**. The equation for the conditional variance depends on the particular effect size (e.g., Equation 5.20 for d_{pool} in two-sample designs), but it generally varies inversely with sample size but directly with the extent of within-groups variability. Thus, results based on larger samples and more homogeneous groups in comparative studies are given greater weight.

The conditional variance of the weighted average effect size M_{ES} is determined by the total number of effect sizes and their weights:

$$s^2_{M_{ES}} = \frac{1}{\sum_{i=1}^{k} w_i} \tag{9.3}$$

The square root of Equation 9.3 is the standard error of the average weighted effect size. The general form of a 100 $(1 - \alpha)$% confidence interval for the population effect size μ_{ES} is

$$M_{ES} \pm s_{M_{ES}} (z_{2\text{-tail}, \alpha}) \tag{9.4}$$

If a confidence interval for μ_{ES} includes zero and $z_{2\text{-tail}, \alpha} = 1.96$, the nil hypothesis that the population effect size is zero cannot be rejected at the .05 level. This is an example of a statistical test in meta-analysis. The power of this test will be low if the number of study effect sizes is relatively small, but even trivial average effect sizes will be statistically significant given sufficiently many primary studies. These tests also assume that the found studies were randomly sampled from the population of all studies, but this is not how primary studies wind up being included in most meta-analyses (i.e., this is another instance of the design–analysis gap). Thus, statistical tests in meta-analysis are subject to the same basic limitations as in primary studies.

Weighting of effect sizes as just described assumes a fixed effects model, or a **conditional model**. It assumes that (a) there is one population of studies with a single true effect size and (b) study effect size departs from true effect size due to within-studies variance only. Thus, effect sizes in conditional models are weighted solely by functions of their conditional variances (Equation 9.2). Other variation in observed effect sizes is viewed as systematic and as a result of identifiable differences due to meta-analytic predictors (study factors). Generalizations in a fixed effects model are limited to studies such as those actually found.

An alternative model for a meta-analysis is a random effects model, also called an **unconditional model**. There is no single population of studies or a constant population effect size presumed to underlie all studies in a random effects model. It assumes instead that (a) there is a distribution of population effect sizes (i.e., there is a different true effect size for each study) and (b) there are two sources of error variance. One is within-studies variation, which in an unconditional model is conceptualized as the difference between an observed effect size and the population effect size estimated by that particular study, just as in a fixed effects model. The second source is **between-studies variance**, which concerns the distribution of all population effect sizes around the population

grand mean effect size. It is commonly assumed that the distribution of population effect sizes is normal, which simplifies the estimation of error variance in a random effects model (Cumming, 2012). A random effects model assumes that the between-studies variance is completely random. In contrast, a mixed effects model assumes that between-studies variation may be a result both of systematic factors that can be identified, such as study factors, and of random sources that cannot. In both random and mixed models, the estimation of two sources of error variance instead of just one, as in a fixed effects model, may improve prediction of observed effect sizes. It is also consistent with generalization of results to studies not identical to the set of found studies.

If the between-studies variance is about zero, there is essentially a single population effect size, which suggests a fixed effects model. This implies that within-studies variance alone is sufficient to explain variation in observed effect sizes and that those effect sizes are therefore homogeneous. But greater variation of population effect sizes says just the opposite, that the observed effect sizes are expected to be heterogeneous because they do not estimate a common parameter. In a random effects model, a weighted mean effect size estimates the grand mean of all population effect sizes. Because there is an additional source of presumed error variance, widths of confidence intervals around weighted mean effect sizes in random effects models are typically wider than the corresponding confidence intervals, assuming a fixed effects model.

An older practice in meta-analysis was to assume a fixed effects model and then estimate the variability of the observed effect sizes. If this variability were too large, a fixed effects model would be rejected in favor of a random effects model (or a mixed effects model). There is a statistic known as Q that measures the degree of heterogeneity within a set of observed effect sizes. It is the total weighted sum of the squares for between-studies variation in effect sizes, and it is calculated as

$$Q = \sum_{i=1}^{k} w_i \left(ES_i - M_{ES}\right)^2 = \sum_{i=1}^{k} w_i ES_i^2 - \frac{\left(\sum_{i=1}^{k} w_i ES_i\right)^2}{\sum_{i=1}^{k} w_i} \tag{9.5}$$

where the mean effect size M_{ES} and the weights w_i for each of the k effect sizes are computed assuming a fixed effects model (Equations 9.1–9.2). The expression in the right side of Equation 9.5 is a computational version more amenable to hand calculation.

Under the null hypothesis of a single population effect size, the Q statistic is distributed as a central chi-square with $k - 1$ degrees of freedom. The latter is the expected value in a central chi-square distribution. If the null hypothesis about a fixed effects model is false, more and more sample χ^2

statistics will exceed the expected value. Suppose that $Q = 10.00$ for a set of 15 effect sizes. Here, the value of Q (10.00) is actually below the expected value (14.00), assuming a fixed effects model. Also, the critical value for χ^2 (14) at the .05 level is 23.68, so the null hypothesis that the 15 results reflect a common population effect size is clearly not rejected at $p < .05$. But the homogeneity hypothesis would be rejected for the same number of effect sizes if $Q = 25.00$ because χ^2 (14) = 25.00, $p = .035$. The test statistic Q is subject to all the same limitations as any other significance test, including low power for small numbers of effect sizes.

Assuming that $Q > df$, the quantity $Q - df$ estimates the noncentrality parameter of the chi-square distribution, or in this case the degree to which the null hypothesis of homogeneity is false. In particular, it reflects the extra variation between studies beyond that expected in a fixed effects model (Cumming, 2012). If a fixed effects model is rejected, between-studies variation is assumed to be an additional source of error variance. One way to estimate this extra variance is to calculate the T^2 statistic, which is

$$T^2 = \frac{Q - df}{C} \tag{9.6}$$

where C is a scaling factor that controls for the fact that Q is a sum of squares, not a variance. In Equation 9.6, it is the division of $Q - df$ by C that estimates the between-studies variance in the same metric as the within-studies variance (Equation 9.2). A computational formula for C is

$$C = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \tag{9.7}$$

The parameter estimated by T^2 is τ^2 (tau-squared), the variance of population effect sizes around the population grand mean effect size in a random effects model.

A more modern approach is to routinely assume random effects models (e.g., Cumming, 2012; F. L. Schmidt, 2010). The rationale for this strategy is threefold: First, incorrect specification of a fixed effects model when the true model is random implies that confidence intervals based on weighted mean effect sizes will be too narrow, which overstates the precision of the results. Second, low power of the test for homogeneity based on the Q statistic could lead to the false retention of a fixed effects model. Third, if there is low between-studies variance, results assuming a fixed effects model versus a random effects model tend to be quite similar.

In fixed effects models, the weight for each effect size reflects just the within-studies variance, s^2_{ES}. But in a random effects model, both within-studies and between-studies variance contribute to the weight for each effect size. One way to estimate weights in random effects models is to add the quantity T^2, which estimates between-studies variance (Equation 9.6), to the within-studies variance of each individual effect size. Because weights are the inverse of error variance, they are computed in a random effects model as

$$w_i^* = \frac{1}{s^2_{ES_i} + T^2} \tag{9.8}$$

where the asterisk designates a random effects model. As the value of T^2 increases—that is, there is greater estimated between-studies variation—the weights for a set of effect sizes become more similar. This implies that effect sizes based on smaller versus larger sample sizes are more similarly weighted in random effects models than in fixed effects models. Cumming (2012) explained it this way: The effect size from a particular study estimates a unique parameter in a random effects model. Because this effect is the only estimate of its corresponding parameter, it cannot be ignored even if the sample size is relatively small.

Computation of the average weighted effect size and its standard error in a random effects model is based on the weights defined by Equation 9.8. The corresponding formulas are

$$M^*_{ES} = \frac{\sum_{i=1}^{k} w_i^* ES_i}{\sum_{i=1}^{k} w_i^*} \tag{9.9}$$

$$s^2_{M^*_{ES}} = \frac{1}{\sum_{i=1}^{k} w_i^*} \tag{9.10}$$

The general form of a 100 $(1 - \alpha)$% confidence interval for the population grand mean effect size μ^*_{ES} has the following general form:

$$M^*_{ES} \pm s_{M^*_{ES}} (z_{2\text{-tail}, \alpha}) \tag{9.11}$$

Confidence intervals for μ^*_{ES} assuming a random effects model are generally wider than those for μ_{ES} assuming a fixed effects model for the same data and level of α. Thus, the choice between the two models in meta-analysis affects the relative contribution of individual effect sizes and the estimation of both the weighted average effect size and its precision.

TABLE 9.1
Study Effect Sizes and Weights for a Fixed Effects Model Versus a Random Effects Model in a Meta-Analysis

					Fixed effects model			Random effects model		
Study	d_{pool}	n_1	n_2	s_d^2	w	w (%)	wd	w^*	w^* (%)	w^*d
1	.50	22	22	.0938	10.667	6.9	5.334	4.285	10.8	2.143
2	.50	12	24	.1285	7.784	5.1	3.892	3.730	9.4	1.865
3	1.20	40	40	.0590	16.949	11.0	20.339	5.034	12.7	6.041
4	.80	80	80	.0270	37.037	24.0	29.630	6.001	15.2	4.801
5	1.30	50	45	.0511	19.563	12.7	25.432	5.242	13.2	6.815
6	1.20	14	85	.0905	11.054	7.2	13.265	4.346	11.0	5.215
7	1.00	55	120	.0294	34.046	22.1	34.046	5.917	14.9	5.917
8	2.00	30	100	.0587	17.031	11.0	34.061	5.041	12.7	10.082
Total					154.131		165.999	39.596		42.878

Note. $\Sigma w_i^2 = 3{,}787.470$; $\Sigma w_i d_i^2 = 203.872$; $\Sigma w_i^{*2} = 200.416$; $\Sigma w_i^* d_i^2 = 54.292$.

Listed in the left side of Table 9.1 are the results of eight hypothetical studies each based on a two-sample design. The observed effect sizes are d_{pool} (Equations 5.3–5.4). I used ESCI (Cumming, 2012; see footnote 4, Chapter 2) to calculate the within-studies variances and weights for a fixed effects model that are reported in the table.[1] Also reported in Table 9.1 are percentages that indicate the relative contribution of each result to the weighted average. These percentages are derived as the ratio of the weight for each study over the total of all the weights, which is 154.130 for a fixed effects model. For example, the weight for study 1 in Table 9.1 is 10.667, so the relative contribution of this effect size is 10.667/154.131 = .069, or about 6.9%.

Given these results from Table 9.1

$$\sum w_i = 154.131 \text{ and } \sum w_i d_i = 165.999$$

the average weighted effect size and its estimated standard error are computed as

$$M_d = \frac{165.999}{154.131} = 1.077 \quad \text{and} \quad s_{M_d} = \sqrt{\frac{1}{154.131}} = 0.0805$$

Based on these results, the 95% confidence interval for μ_{ES} is

$$1.077 \pm .0805(1.96)$$

[1]The ESCI program calculates the error variance of d_{pool} using the square of Equation 5.20 except that the overall sample size N replaces the expression $df = N - 2$ in this equation.

which defines the interval [.92, 1.23] at two-decimal accuracy. Thus, the population effect size could be as low as .92 standard deviations or as high as 1.23 standard deviations, with 95% confidence and assuming a fixed effect model for the eight results in Table 9.1.

Given $\Sigma w_i d_i^2 = 203.872$ from Table 9.1, the value of Q is

$$Q = 203.872 - \frac{165.999^2}{154.131} = 25.091$$

With a total of $k = 8$ studies, the degree of freedom are 7, and the p value for χ^2 (7) = 25.091 is .001. If using a conventional significance test, we would reject the hypothesis of homogeneity that there is a common population effect size at the .05 level.

Now assuming a random effects model for the data in Table 9.1, we estimate the between-studies variance as

$$\sum w_i^2 = 3{,}787.470$$

$$Q - df = 25.091 - 7 = 18.091$$

$$C = 154.131 - \frac{3{,}787.470}{154.131} = 129.558$$

$$T^2 = \frac{18.091}{129.558} = .140$$

The last result, .140, is the estimated between-studies variance expressed in the metric of the within-studies variances for these data.

You should verify with Equation 9.8 that the weights in Table 9.1 for the random effects model are computed for each effect size as the inverse of the sum of the within-conditions variances and $T^2 = .140$. Also note in the table that, as expected, the relative contributions of the individual effect sizes in the random effects models are more generally equal than those in the fixed effects model. This new set of weights for the random effects model also implies new values for the average weighted effect size, its standard error, and the corresponding 95% confidence interval for the population grand mean effect size compared with the fixed effects model. These values for the random effects model, given the data in Table 9.1, are computed as follows:

$$\sum w_i^* = 39.596 \quad \text{and} \quad \sum w_i^* d_i = 42.878$$

$$\sum w_i^{*2} = 200.416 \quad \text{and} \quad \sum w_i^* d_i^2 = 54.292$$

$$M_d^* = \frac{42.878}{39.596} = 1.083 \quad \text{and} \quad s_{M_d}^* = \sqrt{\frac{1}{39.596}} = .1589$$

$$95\%\ \text{CI for}\ \mu_{ES}^*,\ 1.083 \pm .1589\,(1.96),\ \text{or}\ [.77,\ 1.39]$$

As expected, the width of the 95% confidence interval for the random effects model, or [.77, 1.39], is wider than that for the fixed effects model, or [.92, 1.23].

The results just described for the data in Table 9.1 are summarized with the forest plots in Figure 9.1. The noncentral 95% confidence intervals for δ

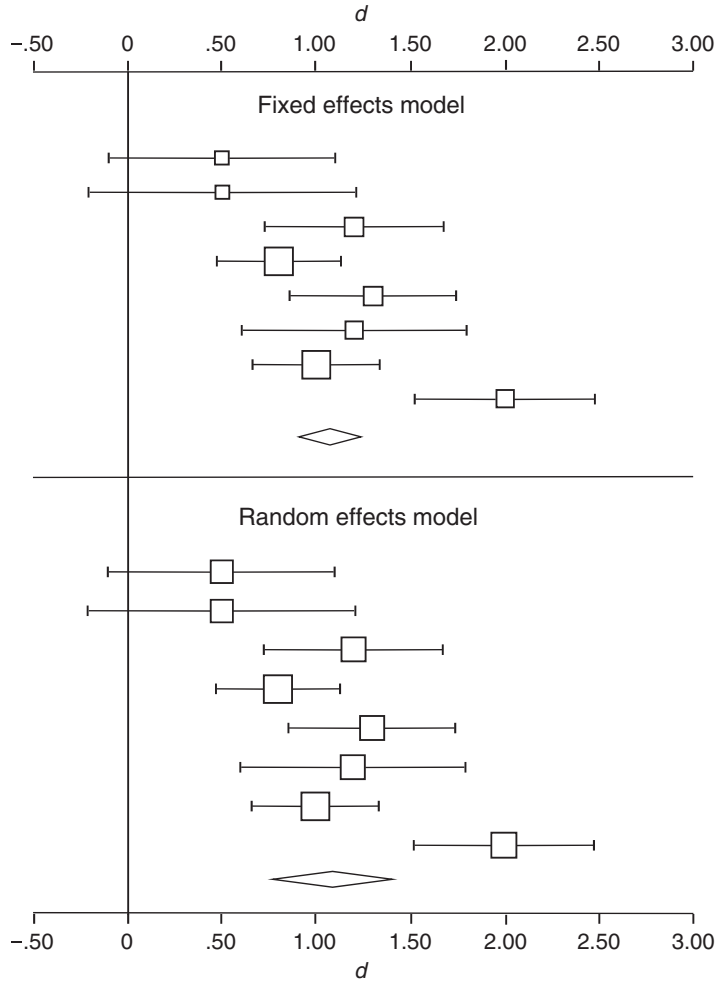

Figure 9.1. Forest plot for a fixed effects model and a random effects model for the data in Table 9.1.

based on each of the eight primary studies are shown oriented horizontally in the figure in the same order as they are listed in the table. Each point estimate of δ is represented in the figure with rectangles depicted in relative sizes that reflect the weight for each study in Table 9.1. Represented as diamonds in the figure are the 95% confidence intervals based on the weighted average effect size from all eight studies. As expected, widths of the confidence intervals based on the aggregated results are narrower than those from any individual study. Comparing the forest plots in the upper and lower parts of Figure 9.1, we can see that the relative contributions of each effect size are generally more similar in the random effects model than in the fixed effects model.

Statistical Techniques

After selecting a model for error variance, the meta-analyst typically chooses a statistical technique for analyzing weighted effect sizes. Techniques include analogs of ANOVA and multiple regression. For example, it is possible to disaggregate studies by the levels of two crossed, categorical study factors (e.g., gender and illness severity) and perform a two-way ANOVA on the weighted effect sizes. This analysis would estimate both main and interactive effects of the predictors. Regression analysis in meta-analyses can include either categorical or continuous factors, such as average patient age in each study, in the same equation. Regression methods also allow individual predictors or blocks of predictors to be entered into or removed from the equation in an order specified by the researcher. Borenstein et al. (2009) described other techniques for analyzing weighted effect sizes.

As is probably obvious by now, many decisions made while analyzing effect sizes can influence the results of a meta-analysis. Conducting a **sensitivity analysis** is one way to address this issue. In meta-analysis, this means that effect sizes are reanalyzed under different assumptions, and the results are compared with the original findings. If the two sets of results are similar, the original meta-analytic findings may be robust with regard to the manipulated assumptions. Suppose that the criteria for study inclusion are modified in a reasonable way, and a somewhat different subset of all found studies is retained. If the meta-analysis is repeated with the new subset and the results are not appreciably different from those under the original criteria, the overall findings are robust concerning the inclusion criteria.

Threats to the Validity of a Meta-Analysis

Those who have promoted (e.g., M. Hunt, 1997) or dismissed (e.g., Eysenck, 1995) meta-analysis would probably agree that it is no less subject to many of the problems that can beset the primary studies on which

it is based. Meta-analysis is also susceptible to additional limitations specific to the technique. Some examples were mentioned, including the apples and oranges problem and the garbage in, garbage out problem. Other possible threats are outlined next; see also Borenstein et al. (2009, Chapter 43) for more information.

Although it is useful to know average effect sizes in some research areas, effect size by itself says little about substantive significance (see Chapter 5, this volume). It is also true that explaining a relatively high proportion of observed variance in outcomes with a set of study factors does not imply that these variables are actually the ones involved in the underlying process. It is possible that an alternative set of study factors may explain just as much of the variance or that some of the measured predictors are confounded with other, unmeasured factors that are actually more important. Meta-analysis is not a substitute for primary studies. Despite their limitations, primary studies are the basic engine of science. Indeed, a single brilliant empirical or theoretical work could be worth more than hundreds of mediocre studies synthesized in a meta-analysis. There is also concern about the practice of guarding against experimenter bias by having research assistants code the primary studies. The worry is about a crowding out of wisdom that may occur if what is arguably the most thought-intensive part of a meta-analysis—the careful reading of the individual studies—is left to others.

It is probably best to see meta-analysis as a way to better understand the status of a research area than as an end in itself or some magical substitute for critical thought. Its emphasis on effect sizes and the explicit description of study retrieval methods and assumptions is an improvement over narrative literature reviews. It also has the potential to address hypotheses not directly tested in primary studies. If the results of a meta-analysis help researchers conduct better primary studies, little more could be expected. But meta-analysis does not solve the replication crisis in the behavioral sciences. It is merely a stopgap until we change our mentality and behavior so that explicit replication is both expected and rewarded. The availability of meta-analysis should not prevent the behavioral sciences from growing up in this regard.

Meta-Analysis and Statistics Reform

Meta-analysis can bring clarity to a research problem previously considered only through the lens of significance testing. For example, Lytton and Romney (1991) conducted one of the first meta-analyses in the area of differential socialization, which refers to the encouragement of certain traits or behaviors more for children of one gender than of another. It is also possible that fathers may make greater differences between sons and daughters than do mothers or vice versa. Some narrative reviews published in the 1970s

and 1980s concluded that "significant" differential socialization effects were found in about half the studies and called for more research to resolve the ambiguity. But no amount of additional research would have ever resolved the ambiguity if power were about .50, which is consistent with the pattern just described.

Lytton and Romney (1991) synthesized standardized mean differences from studies where mothers and fathers reported about their emphasis on various traits or behaviors in the upbringing of their sons versus daughters. They further partitioned the effect sizes in the retained set of about 160 studies from North America by eight different socialization areas, such as warmth, discipline, achievement, and gender-typed activities. Weighted average *d* statistics for boys versus girls were generally close to zero for both mothers and fathers in most areas. For example, both mothers and fathers emphasized achievement more with their sons than with their daughters, but the mean effect size in this area was $d = .05$ for mothers and $d = .11$ for fathers. Both mothers and fathers encouraged reasoning more strongly among sons than daughters, but the average effect size for both parents was only about $d = .01$. Lytton and Romney (1991) found evidence for stronger differential socialization in just one area, gender-typed activities. For mothers, the average effect size was $d = .34$, and for fathers it was $d = .49$. That is, both mothers and fathers emphasized gender-typed activities more strongly among sons than daughters, but fathers tended to differentiate more strongly than mothers between boys and girls in this area. See Cumming (2012) for descriptions of other occasions when results of meta-analytic studies have brought clarity to research areas long dominated by significance testing.

CONCLUSION

There is little evidence for direct or even approximate replication in the behavioral science literature. This fact contradicts what is expected from the most basic presumption of science, but the reality in the behavioral sciences is that there are few incentives for students or researchers to conduct explicit replications. Instead, replication is more often carried out implicitly in the form of follow-up studies that extend prior investigations by adding new elements. The problem is that differences between original and follow-up studies may not be systematically cataloged, which makes it difficult to interpret the meaning of an apparent failure to find consistent results over studies. Meta-analysis generally estimates average weighted effect sizes from primary studies and evaluates whether various study factors, such as characteristics of participants or treatments, explain variation in effect sizes. Crucial questions about the validity of meta-analysis concern the selection of studies, assessment

of their quality and measurement of their characteristics, how lack of independence in effect sizes is handled, and the underlying statistical model assumed. A good meta-analysis should summarize the status of a literature and suggest new directions. It is not a substitute for explicit replication, but the behavioral sciences are not yet mature enough to do without meta-analysis.

LEARN MORE

Card (2012) gives a clear introduction to meta-analysis, and Borenstein et al. (2009) is for researchers who intend to conduct meta-analyses. S. Schmidt (2009) analyzes replication in the behavioral sciences.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley. doi:10.1002/9780470743386

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York, NY: Guilford Press.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90–100. doi: 10.1037/a0015108

This page intentionally left blank

10

BAYESIAN ESTIMATION AND BEST PRACTICES SUMMARY

> Friends do not let friends compute *p* values.
>
> —John K. Kruschke (2010, p. 294)

This chapter introduces the Bayesian approach to hypothesis testing and interval estimation. Potential advantages include the capabilities to estimate probabilities of hypotheses, directly compare the likelihoods of competing hypotheses, and estimate inferential confidence intervals, all under explicit assumptions. It also provides a systematic framework for revising the plausibility of hypotheses as new data are collected. Entire books are devoted to Bayesian statistics, so it is not possible to give a complete account here. The more modest goal is to make you aware of an inference model that increasing numbers of behavioral science researchers see as a viable alternative to significance testing. Presented at the end of this chapter are best practice recommendations based on all topics considered to this point.

DOI: 10.1037/14136-010
Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, Second Edition, by R. B. Kline

CONTEXTS FOR BAYESIAN ESTIMATION

We considered many limitations of significance testing. One is that *p* values do not estimate the probability that H_0 is true, given the data. They tell us instead the conditional probability of the data or results even more extreme under H_0, designated as p (Data+$|H_0$). They are subject to many decisions, or researcher degrees of freedom, not all of which are always explicit. The default form of H_0 is a nil hypothesis, but such hypotheses are usually implausible. Percentages associated with standard confidence intervals, such as 95%, cannot generally be interpreted as the chance that the interval contains the corresponding parameter.

Bayesian estimation is not constrained by these limitations, in part because it is not based on a frequentist model of probability. Also, a parameter in Bayesian methods is generally conceptualized as a random variable with its own distribution that summarizes the current state of knowledge about that parameter. The distribution's expected value is the single best guess about the true value of the parameter, and its variability reflects the amount of uncertainty. In contrast, a parameter in significance testing is viewed as a constant that should be estimated with sample statistics. This difference explains why symbols for parameters are printed in italic font to emphasize that they are variables in Bayesian estimation, such as μ for a random population mean in Bayesian statistics instead of μ for a constant parameter in classical statistics.

Since the late 1950s, Bayesian methods have been used in disciplines such as economics, medicine, engineering, and computer science (e.g., Pourret, Naïm, & Marcot, 2008). Your computer may be connected to a network protected by a dynamic Bayesian system that continually updates estimated probabilities of hacking (intentional malevolent access) for each user, given recent activities on the network. If this probability exceeds some threshold (e.g., > .75), that user or account may be locked out of the network (e.g., Christina, 2010). Introduced to psychology in the 1960s by authors such as Edwards, Lindman, and Savage (1963), Bayesian statistics never really caught on among behavioral scientists over the next 40 years.

Recently, there has been a surge of interest in Bayesian methods among behavioral scientists. An example is the controversy about Bem (2011), who reported evidence for psi, the ability to anticipate events before they happen, based on statistically significant results in eight of nine experiments with a mean effect size of about $d = .25$. The same results analyzed from a Bayesian perspective, however, are not as convincing (e.g., Kruschke, 2011; Wetzels et al., 2011). Another example of increasing attention is the special edition about Bayesian methods in *Trends in Cognitive Science* (e.g., Chater, Tenenbaum, & Yuille, 2006). As I edited this chapter, the *Journal*

of Management issued a call for papers for a special issue titled "Bayesian Probability and Statistics in Management Research: A New Horizon," to be published in 2014.

Once some fundamentals are mastered, Bayesian hypothesis testing is closer to intuitive scientific reasoning than significance testing. For instance, the following principles are all supported in Bayesian analysis:

1. Not all hypotheses are equally plausible before there is evidence, and implausible hypotheses require stronger evidence to be supported. This is a basic tenet of science that extraordinary claims require extraordinary proof. I mentioned that *p* values under implausible nil hypotheses are generally too low, which exaggerates the relative rarity of the data. In contrast, Bayesian methods take explicit account of hypothesis plausibility.
2. Not all researchers will see the same hypothesis as equally plausible before data collection. This is another basic fact of science, if not human nature. The effect of assuming different degrees of plausibility for the same hypothesis can also be explicitly estimated in a Bayesian analysis. Doing so is a kind of sensitivity analysis that makes plain the effects of differing initial assumptions.
3. The impact of initial differences in the perceived plausibility of a hypothesis tends to become less important as results accumulate. So open-minded scientists with different initial beliefs are generally driven toward the same conclusion as new data are collected. The real long-term effect of initial differences in belief is that skeptics will require more data to reach the same level of belief as that held by those more enthusiastic about a theory.
4. Data that are not precise will have less sway on the subsequent plausibility of a hypothesis than data that are more precise. This is a principle of meta-analysis, too.

A longtime objection is that Bayesian methods are associated with a subjectivist view of probability. Recall that a subjectivist view does not distinguish between repeatable and unrepeatable (unique) events, and probabilities are considered as degrees of personal belief that may vary from person to person. There is a perception among those unfamiliar with Bayesian statistics that prior probabilities of hypotheses are wholly subjective guesses just plucked out of thin air, perhaps to suit some whim or prejudice.

These perceptions of Bayesian statistics are false. If nothing is known about some hypothesis, the researcher has little choice other than to guess

about plausibility. But it is rare for researchers to have absolutely no previous information on which to base estimates of prior probabilities. There are heuristic methods for eliciting consistent prior probabilities from content experts that try to avoid common difficulties that arise when people reason with probabilities. One is the **conjunction fallacy**, which occurs when a higher probability is estimated for two joint events than for the individual events. Some of these methods include the posing of questions in a frequency format instead of a probability format, which may help to avoid inconsistent or illogical reasoning with probabilities. Estimates of prior probabilities in Bayesian analyses are explicit and thus open to debate. It is also possible to estimate conditional probabilities of hypotheses under a range of estimates about their prior probabilities.

BAYES'S THEOREM

The starting point is **Bayes's theorem**, which is from a posthumous publication (1763) of a letter by Rev. Thomas Bayes in the *Philosophical Transactions of the Royal Society*. It is based on the mathematical fact that the joint probability of two events, D and H, is the product of the probability of the first event and the conditional probability of the second event given the first, or

$$p(D \wedge H) = p(D)p(H|D) = p(H)p(D|H) \tag{10.1}$$

where the logical connective $\wedge$ designates the conjunctive *and*.

Now let us assume that D in Equation 10.1 stands for a particular result in a primary study and does not include all more extreme results. Next we designate this particular result as Data in order to distinguish it from Data+, which in significance testing includes all more extreme results under H_0. Let us also take the symbol H in Equation 10.1 to mean hypothesis but not necessarily a point hypothesis such as H_0: $\mu_1 - \mu_2 = 0$ in significance testing. A hypothesis in Bayesian estimation about a continuous parameter can be either a point hypothesis or a range hypothesis. In significance testing, H_1 is almost always a range hypothesis (e.g., H_1: $\mu_1 - \mu_2 > 0$), but H_0 is usually a point hypothesis. The larger issue is that the specification of hypotheses is more flexible in Bayesian inference than in classical significance testing.

With these definitions in mind, solving Equation 10.1 for the conditional probability $p\,(H|\text{Data})$ gives us the basic form of Bayes's theorem:

$$p(H|\text{Data}) = \frac{p(H)p(\text{Data}|H)}{p(\text{Data})} \tag{10.2}$$

In words, p (H|Data) is the **posterior probability** of the hypothesis, and it estimates the probability of that hypothesis in view of a particular result. It is a function of two **prior (marginal, unconditional) probabilities**, p (H) and p (Data), and a conditional probability called the **likelihood**, or p (Data|H). The latter is the probability of a result under the hypothesis, and it is analogous to p (Data+|H_0) in significance testing, given the differences between the terms Data and Data+ and between the terms H and H_0 just explained.

The term p (H) in Equation 10.2 is the probability of the hypothesis before the data are collected, and p (Data) is the probability of the data irrespective of the truth of any hypothesis. Bayes's theorem thus takes an initial belief about the hypothesis, p (H), and combines it with information from the sample to generate an updated belief, p (H|Data). It also shows us that to correctly translate the conditional probability of the data to the conditional probability of the hypothesis, we need also to estimate the prior probabilities of both the data and the hypothesis.

If extant theory makes no relevant prediction or there are no empirical studies, the specification p (H) = .50 is consistent with this lack of information. The specification p (H) < .50 is more consistent with a skeptic's view, but stating that p (H) > .50 may be warranted when there is already evidence that favors the hypothesis. A lower initial assignment of p (H) means that more evidence will be required to eventually raise the estimate of p (H|Data) to a level that more clearly supports the hypothesis. But the specification of prior probabilities of hypotheses in Bayesian estimation must be explicit, and the consequences of different assumptions about p (H) can be evaluated in a sensitivity analysis.

This example by Dixon and O'Reilly (1999) illustrates a simple application of Bayes's theorem. Suppose we want to estimate the probability that it will snow sometime during the day, given a below-freezing temperature in the morning. The chance of snow on any particular day of the year is only 10%, so p (H) = .10. The chance of a below-freezing morning temperature on any particular day is 20%, so p (Data) = .20. Of all days it snowed, the chance of a below-freezing temperature in the morning is 80%, so p (Data|H) = .80. When Equation 10.2 is used, the posterior probability is

$$p(H|\text{Data}) = \frac{.10(.80)}{.20} = .40$$

which says that there is a 40% chance that it will snow on days when it is cold in the morning.

BAYESIAN TESTING FOR POINT HYPOTHESES

Next we assume k mutually exclusive and exhaustive point hypotheses about a parameter. The sum of their prior probabilities is 1.0, or

$$\sum_{i=1}^{k} p(H_i) = 1.0 \tag{10.3}$$

The prior probability of the data in Equation 10.2 for Bayes's theorem can now be expressed as

$$p(\text{Data}) = \sum_{i=1}^{k} p(H_i)\, p(\text{Data}|H_i) \tag{10.4}$$

which is the sum of the products of the prior probabilities for each of the k discrete hypotheses and the likelihood of the data under it.

Suppose that the distribution on a continuous variable in a population is normal, the variance is known—that is, it is a constant, not a variable, and we assume $\sigma^2 = 144.00$—but the mean is not known (i.e., it is a random variable, μ). There are two competing hypotheses, or

$$H_1: \mu = 100.00 \quad \text{and} \quad H_2: \mu = 110.00$$

Assuming no previous information, the two hypotheses are judged to be equally likely, or

$$p(H_1) = p(H_2) = .50$$

The assignment of equal prior probabilities to all competing hypotheses when there are no grounds to favor any one of them follows the principle of indifference (see Chapter 2). Related descriptive terms include **agnostic priors** and **uninformative priors**. In contrast, **informative priors** reflect greater confidence in one hypothesis than the other. For example, specification of the prior probabilities

$$p(H_1) = .40 \quad \text{and} \quad p(H_2) = .60$$

would reflect greater confidence in H_2 than H_1. The ratio $p\,(H_2)/p(H_1)$, or .60/.40 = 1.67, is the **prior odds** (i.e., 3:2) that H_2 is correct. But if $p(H_1) = p(H_2)$, as in this example, the prior odds that either hypothesis is correct are 1.0 (i.e., 1:1, no difference).

A sample of 16 cases is collected, and the observed mean is $M_1 = 106.00$. Because the population variance is known, the standard error of the mean is $(144.00/16)^{1/2} = 3.00$. Under the normality assumption, the conditional probability of the sample mean under each hypothesis—the likelihoods—can be found with the standard normal density function

$$\text{ndf}(z) = \frac{e^{-z/2}}{\sqrt{2\ \text{pi}}} \tag{10.5}$$

where z is a normal deviate, e is the natural base (about 2.7183), and pi is approximately 3.1416. The function takes a z score and returns its probability, which is the height of the normal curve with a mean of zero and a standard deviation of 1.0 at that point.[1] The z score equivalents of the sample mean under each hypothesis are

$$z_{H_1} = \frac{106.00 - 100.00}{3.00} = 2.000 \quad \text{and} \quad z_{H_2} = \frac{106.00 - 110.00}{3.00} = -1.333$$

and the likelihoods of $M_1 = 106.00$ under each hypothesis are

$$p(\text{Data}_1 | H_1) = \frac{\text{ndf}(2.000)}{2} = \frac{.0540}{2} = 0.0270$$

$$p(\text{Data}_1 | H_2) = \frac{\text{ndf}(-1.333)}{2} = \frac{.1640}{2} = 0.0820$$

The results of the ndf function are divided by two because there are two hypotheses. These results say that the probabilities of the data under H_1 and H_2 are, respectively, .0270 and .0820. The prior probability of the data ($M_1 = 106.00$) is

$$p(\text{Data}_1) = .50(.0270) + .50(.0820) = .0545$$

and applying Bayes's theorem tells us that the posterior probabilities for each hypothesis are

$$p(H_1 | \text{Data}_1) = \frac{.50(.0270)}{.0545} = .2477$$

$$p(H_2 | \text{Data}_1) = \frac{.50(.0820)}{.0545} = .7523$$

[1] In Microsoft Excel, the function NORMDIST(z, 0, 1, True) returns the likelihood of z.

In summary, our revised estimate of the probability of H_2: $\mu = 110.00$ (about .75) is higher than that for H_1: $\mu = 100.00$ (about .25), given $M_1 = 106.00$.

Posterior Odds and the Bayes Factor

The **posterior odds** are the ratio of the conditional probabilities of two competing hypotheses for the same data. For the example

$$\text{Posterior odds}_1 = \frac{p(H_2|\text{Data}_1)}{p(H_1|\text{Data}_1)} = \frac{.7523}{.2477} = 3.04$$

which says that the odds are about 3:1 in favor of H_2 that the population mean is 110.00 over H_1 that this mean is 100.00 after observing $M_1 = 106.00$. Which of the two hypotheses is represented in the numerator is arbitrary. For this example, the ratio .2477/.7523, or .329, is the posterior odds for H_1 relative to H_2 (i.e., about 1:3 against H_1).

With Equation 10.2 used for Bayes's theorem, it can be demonstrated that posterior odds can be expressed as the product of the prior odds and the likelihood ratio, or the **Bayes factor** (BF). That is,

$$\text{Posterior odds} = \text{Prior odds} \times \text{BF} \tag{10.5}$$

where the prior odds are $p(H_2)/p(H_1)$ and the Bayes factor is

$$\text{BF} = \frac{p(\text{Data}|H_2)}{p(\text{Data}|H_1)} \tag{10.6}$$

which summarizes the relative likelihood of the same data under the two hypotheses. (Compare Equations 6.8 and 10.5.) The Bayes factor also summarizes the results of the study that allow the update of the odds of the two hypotheses from what they were before collection of the data (prior odds) to what they should be given the data. If the prior odds do not favor one hypothesis over the other (i.e., it is 1.0), the value of BF directly equals that of the posterior odds. For the example where the prior odds are 1.0, the value of the Bayes factor is

$$\text{BF}_1 = \frac{p(\text{Data}_1|H_2)}{p(\text{Data}_1|H_1)} = \frac{.0820}{.0270} = 3.04$$

which equals the posterior odds for this example calculated earlier (3.04) as the ratio of the likelihood of the two hypotheses, given $M_1 = 106.00$.

The Bayes factor is a continuous measure of the likelihood of the data under two competing hypotheses. If the value of BF is close to 1.0, the results of the study failed to differentiate between the hypotheses. Otherwise, values of BF that exceed about 3.00 (or are less than .33) are generally taken as indicating support for one hypothesis over another, but this rule of thumb should not be rigidly applied (i.e., do not dichotomize the Bayes factor). This heuristic is useful for comparing *p* values in significance testing to BF values computed for the same data. For example, Jeffreys (1961) suggested that $p < .05$ from statistical tests would generally correspond to about $BF > 3.00$ (or $BF < .33$) for typical sample sizes.

Wetzels et al. (2011) provided a more empirical basis for relating *p* and BF in the same samples. They compared results from a total of 855 *t* tests reported in 252 articles published in the 2007 volumes of two different experimental psychology research journals. They found that about 70% of *p* values between .01 and .05 were associated with a value of the Bayes factor that indicated only anecdotal support for the alternative hypothesis (e.g., $BF < 3.00$ assuming the numerator corresponds to H_1 of the *t* test). Although *p* values and BF values were strongly correlated, Wetzels et al. (2011) suggested that the Bayesian approach was generally more conservative than the standard *t* test over this collection of studies.

Sample size affects *p* values and BF values in different ways, too. If H_0 is true, *p* does not converge to any particular value as more data are collected. There is also the problem that a researcher is practically guaranteed to get statistically significant results by simply collecting more data (the stopping rule issue; see Chapter 3). In contrast, values of BF are driven toward zero as more data are collected, if H_0 is true. But when H_0 is false, *p* values tend to decrease and BF values tend to increase as more cases are added (Dienes, 2011). Wetzels et al. (2011) reminded us that the Bayes factor is not synonymous with effect size and that estimating effect sizes complements a Bayesian analysis, too. It is also possible in Bayesian methods to directly analyze effect sizes.

Updating Posterior Odds

The posterior odds can be updated as additional data are collected by iteratively applying Bayes's theorem. This is why Edwards et al. (1963) said that key principles of Bayesian estimation are "that probability is orderly opinion, and that inference from data is nothing other than the revision of such opinion in the light of relevant new information" (p. 194). Recall the previous example where the posterior odds are 3.04 in favor of H_2 that $\mu = 110.00$ against H_1 that $\mu = 100.00$, given $M_1 = 106.00$, $N = 16$. Suppose in a second sample of 16 cases it is found that $M_2 = 107.50$. The posterior odds from the previous analysis ($BF_1 = 3.04$) become the prior odds in the

new analysis. The updated posterior odds after observing the new result are calculated as the product of the Bayes factor from the original analysis and the new Bayes factor, or

$$\text{Posterior odds}_2 = \text{BF}_1 \times \text{BF}_2 \tag{10.7}$$

The value of the normal deviate for $M_2 = 107.50$ is 2.500, given $\sigma_M = 3.00$ and assuming $\mu = 100.00$ under H_1, so the likelihood of the second mean under this hypothesis is

$$p(\text{Data}_2 | H_1) = \frac{\text{ndf}(2.500)}{2} = \frac{.0175}{2} = .0088$$

Under H_2, which assumes $\mu = 110.00$, the normal deviate for the second mean is −.833, so the likelihood under this hypothesis is

$$p(\text{Data}_2 | H_2) = \frac{\text{ndf}(-.833)}{2} = \frac{.2820}{2} = .1410$$

The value of the new Bayes factor is

$$\text{BF}_2 = \frac{p(\text{Data}_2 | H_2)}{p(\text{Data}_2 | H_1)} = \frac{.1410}{.0088} = 16.02$$

so the likelihood of $M_2 = 107.50$ is about 16 times greater under H_2 than H_1. This is also the factor by which the posterior odds from the first analysis will be updated given the second result. The new posterior odds are

$$\text{Posterior odds}_2 = 3.04 \times 16.02 = 48.71$$

which now favor H_2 that $\mu = 110.00$ over H_1 that $\mu = 100.00$ even more strongly than the original posterior odds when only the first result (3.04) was available.

BAYESIAN TESTING FOR RANGE HYPOTHESES

It is rare that hypothesis testing is as narrow as described in the previous example. Researchers do not typically know the population variance, nor do they generally evaluate competing point hypotheses about the value of an unknown population parameter. Although the default null hypothesis in sig-

nificance testing is a point hypothesis, the alternative hypothesis is usually a range hypothesis. If the alternative hypothesis in Bayesian estimation concerns a continuous random parameter, such as the directional range hypothesis H_1: $\mu_1 - \mu_2 > 0$ tested against the point hypothesis H_0: $\mu_1 - \mu_2 = 0$, the researcher must specify the prior distribution of that parameter under each hypothesis. A prior distribution for a range hypothesis is usually described with a probability density function that defines the likelihood of any value contained within the distribution. It is the mathematical operation of integration on this function that gives the probability for a range of values, and the integral over the whole range is 1.0.

The trick in Bayesian estimation is to specify an appropriate prior distribution for a range hypothesis. If this specification is grossly wrong, subsequent estimates of the conditional probabilities of the data under range hypotheses may also be incorrect. The same thing goes for the posterior distribution, which is basically an updated version of the prior distribution after observing the data. Because the whole framework of Bayesian estimation is iterative, the posterior distribution at the conclusion of one study can be specified as the prior distribution in a subsequent study about the same random parameter and so on.

Presented in Figure 10.1 are examples of prior distributions for hypotheses about a random population parameter. Figures 10.1(a)–10.1(e) represent prior distributions for the difference between two random population means, $\mu_1 - \mu_2$, of the type discussed by Dienes (2011), Kruschke (2011), and Rouder et al. (2009) for Bayesian versions of the *t* test. Depicted in Figure 10.1(a) is the probability distribution for the point null hypothesis H_0: $\mu_1 - \mu_2 = 0$. This distribution has only one value (zero), and its likelihood is assumed to be 1.0. The prior distribution in Figure 10.1(b) includes a good-enough belt around the point hypothesis $\mu_1 - \mu_2 = 0$. The majority of the values contained within this distribution are considered as practically equivalent and uninteresting departures from zero. These deviations are assumed to be normally distributed. In Bayesian equivalence testing, the prior distribution for the null hypothesis could be specified in a similar manner.

Represented in Figure 10.1(c) is the continuous uniform (rectangular) prior distribution for the directional alternative hypothesis

$$H_1 : 0 < (\mu_1 - \mu_2) < 5.0$$

which predicts $\mu_1 > \mu_2$ but also limits the upper bound of the expected population mean difference to 5.0. The distribution in Figure 10.1(c) also represents every result within the range 0–5.0 as equally plausible. Specification of the lower and upper bounds of a rectangular distribution is sometimes justified by the scale on which means are calculated. If scores on that scale range from 0 to 5, the difference between two means cannot exceed 5.0.

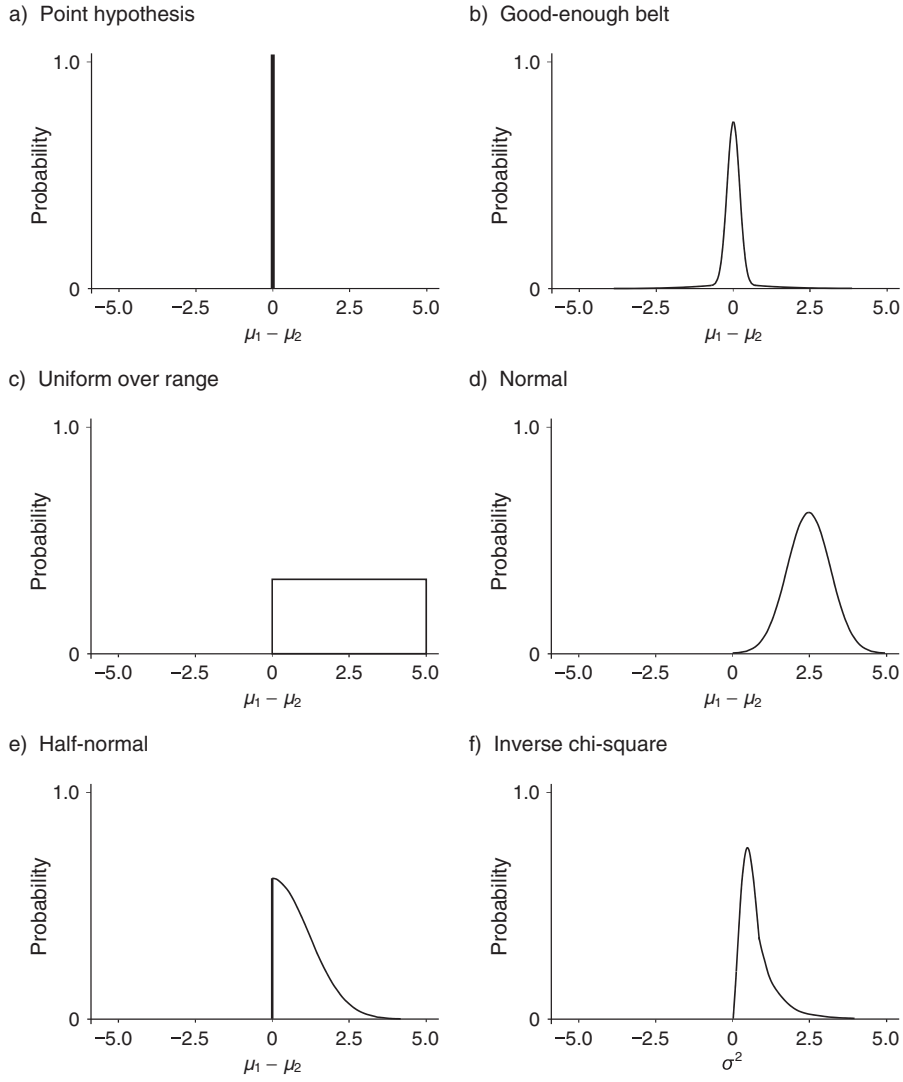

Figure 10.1. Examples of prior distributions for hypotheses about differences between random population means (a–e) and a random population variance (f).

It may be unreasonable for some research problems to assume that all possible outcomes are equally probable. The prior distribution depicted in Figure 10.1(d) is for a directional alternative hypothesis where the most plausible estimate of $\mu_1 - \mu_2$ is 2.5, but where values lower or higher than 2.5 are represented as progressively unlikely. The drop-off in likelihoods moving away from the center of the distribution in Figure 10.1(d) is assumed to follow a normal curve. The prior distribution in Figure 10.1(e) is also for

a directional alternative hypothesis. It is a half-normal distribution with a mode of zero that explicitly assumes that smaller effects are more likely than increasingly larger effects.

If we assume a common population variance but its precise value is unknown, the random parameter σ^2 has its own prior distribution. Depending on the analysis, it may also be necessary to specify a probability density function for σ^2. An example is presented in Figure 10.1(f), which depicts an **inverse chi-square distribution** with a single degree of freedom. The expected value for σ^2 is 1.0, and the prior distribution in the figure represents the prediction that likelihood falls off sharply for very small and very large values of σ^2. There are also times when it makes sense to expect that effect sizes follow similar distributions (Rouder et al., 2009). Other candidates for prior distributions of random variances include inverse chi-square distributions where $df \geq 2$ and **inverse-gamma distributions**, which have parameters for shape and scale. Selection of a suitable prior distribution for a random variance should also be guided by theory and empirical results.

There are many other theoretical probability density functions, such as the binomial distribution for proportions and the multivariate normal distribution for joint random variables such as covariances, and a Bayesian analysis is easier if a known distribution can be selected to model the prior distributions. The same family of known probability distributions—also called **conjugate distributions**—may be used in the analysis to specify both the prior distribution and the posterior distribution. If so, the prior distribution is referred to as the **conjugate prior** when estimating the likelihood of the data under each hypothesis.

Selection of an appropriate prior distribution is a question of statistical model fitting. The choice can affect the results, but consequences of specifying different distributions can be evaluated in a sensitivity analysis. An alternative is to specify a noninformative prior, which assumes no specific knowledge. A noninformative prior for a continuous random parameter is just a flat distribution with infinite variance, ∞. The limit of the ratio $1/\infty$ is infinitely small, so the precision of the knowledge about a parameter described by a flat distribution (noninformative prior) is also infinitely small (i.e., practically zero). In general, it takes more evidence to eventually support a particular hypothesis with an uninformed prior than with an informed prior that is more approximately correct.

Dienes (2011) described a Bayesian version of the *t* test for either a single sample or two samples (independent or dependent) that estimates the Bayes factor for comparing a range alternative hypothesis against a point nil hypothesis. The test assumes normality and homoscedasticity in two-sample tests, and sample variance is assumed to estimate the random population variance. The latter means that no specification for the prior distribution of

the population variance is required. There is also a freely available online calculator that computes Bayes factor values.[2] To use the calculator, the researcher must enter the mean (or mean difference) and its standard error (i.e., Equation 2.6, 2.12, or 2.20). The researcher must also specify whether H_1 is one- or two-tailed and the form of the prior distribution under H_1. There are three choices:

1. A uniform distribution for a definite range over which all values are represented as equally plausible. The lower and upper bounds must be specified; see Figure 10.1(c).
2. A normal distribution where the mean equals the predicted value under H_1 and where values lower or higher are represented as progressively less likely. The researcher must also specify the standard deviation in this normal prior distribution. If the mean of the distribution is not zero, Dienes (2011) suggested, a reasonable specification would be .50 times the value of the distribution's mean, such as .50 × 2.50, or 1.25 for the normal curve in Figure 10.1(d). Otherwise, a value that equals .50 times the range of plausible values is a reasonable specification for the standard deviation. For example, if the lower and upper bounds for this plausible range are, respectively, 0 and 5.0, then one half of the range is 2.5, which is the specification for the standard deviation.
3. A half-normal distribution where the mode is centered on zero and increasingly larger values are represented as progressively less plausible; see Figure 10.1(e). A reasonable specification for the standard deviation of this distribution is .50 times the range of plausible values for the parameter under H_1.

Suppose the results of the standard t test in a balanced design with two independent samples where $M_1 - M_2 = 2.00$, $s_{M_1-M_2} = .90$, and $n = 30$ are

$$t(58) = 2.22,\ p = .015 \text{ for } H_0: \mu_1 - \mu_2 = 0 \text{ against } H_1: \mu_1 - \mu_2 > 0$$

Based on these results, the conventional nil hypothesis is rejected at the .05 level. In a Bayesian version of this test, plausible values for $\mu_1 - \mu_2$ range from zero to 10.0. Listed next are values of Bayes factor computed for these data with Dienes's (2008) online calculator:

1. For a uniform prior distribution over the range 0–10.0, BF = 2.63, which does not strongly support H_1 over H_0.

[2]http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/bayes_factor.swf

2. For a normal prior distribution where the mean is 5.0 and the standard deviation is 2.5, BF = 2.12, which also does not clearly support H_1.
3. But for a half-normal distribution centered at zero and assuming a standard deviation of 2.5, BF = 5.92, which indicates more reasonable support for H_1.

Thus, the outcome of a Bayesian *t* test for the same data depends on assumptions about prior distributions for $\mu_1 - \mu_2$. In contrast, the standard *t* test is not sensitive to hypotheses about the distributional form of the effect under study. Rouder et al. (2009) described Bayesian versions of the *t* test and also Bayesian tests for regression analyses and binomial data, such as the proportion of successful learning trials over all trials. An online calculator that computes BF values for these tests is available.[3] See Kruschke (2011) and Morey and Rouder (2011) for more information about testing nil hypotheses in a Bayesian framework. Masson (2011) described Bayesian hypothesis testing in designs where the technique of ANOVA is used.

BAYESIAN CREDIBLE INTERVALS

Methods for **Bayesian parameter estimation** do not directly compare competing hypotheses. Instead, they are analogous to interval estimation in more standard analyses except that **credible intervals** in Bayesian estimation—also called **highest density regions** or **Bayesian confidence intervals**—establish relative probabilities for a range of candidate values of a random parameter. These intervals are represented in posterior probability distributions that are updated as new data are collected. The variances of posterior distributions generally decrease as more and more new results are synthesized along with the old, just as in meta-analysis. But unlike those for traditional confidence intervals, percentages associated with credible intervals, such as 95%, are interpreted as the probability that the true value of the random parameter is between the lower and upper bounds of the interval.

With one exception, traditional confidence intervals are not to be interpreted this way (see Chapter 2). The exception occurs when the prior distribution is flat (uninformative), which implies that the parameters of the posterior distribution are estimated solely with the sample data. In this case, the Bayesian confidence interval is asymptotically identical in large samples to the traditional confidence interval for the unknown parameter,

[3]http://pcl.missouri.edu/bayesfactor

given the same distributional assumptions. But when the prior distribution is informative, the parameters of the posterior distribution are basically a weighted combination of those from the prior distribution and those estimated in the sample. The weights reflect the precision of each source of information. In this case the Bayesian confidence interval is also generally different from the traditional confidence interval for the same result.

Suppose the mean and variance in a normal prior distribution for a random population mean are, respectively, μ_0 and σ_0^2. The precision of this distribution is $prc_0 = 1/\sigma_0^2$. The shape of the posterior distribution will be normal, too, if (a) the distribution of scores in the population is normal and (b) the sample size is not small, such as $N > 50$, in which case at least approximate normality may hold (Howard, Maxwell, & Fleming, 2000). The latter also permits reasonable estimation of the population variance with the sample variance. The observed mean and error variance in a sample are, respectively, M_1 and $s_{M_1}^2$, and the precision of the sample mean is $prc_1 = 1/s_{M_1}^2$. Given the assumptions stated earlier, the mean in the posterior distribution, μ_1, is the weighted combination of the mean in the prior distribution and the observed mean:

$$\mu_1 = \left(\frac{prc_0}{prc_0 + prc_1}\right)\mu_0 + \left(\frac{prc_1}{prc_0 + prc_1}\right)M_1 \tag{10.8}$$

The variance of the posterior distribution, σ_1^2, is estimated as

$$\sigma_1^2 = \frac{1}{prc_0 + prc_1} \tag{10.9}$$

Note in Equation 10.8 that the relative contribution of new knowledge, the observed mean M_1, depends on its precision, prc_1, and the precision of all prior knowledge taken together, prc_0.

An example demonstrates the iterative estimation of the posterior distribution for a random population mean as new data are collected. The distributional characteristics stated earlier are assumed. Suppose that the researcher has no basis to make a prior prediction about the value of μ, so a flat prior distribution with infinite variance is specified as the prior distribution. A sample of 100 cases is selected, and the results are

$$M_1 = 106.00,\ s_1 = 25.00, \text{ and } s_{M_1} = 2.50$$

The traditional 95% confidence interval for the population mean computed with $z_{2\text{-tail, }.05} = 1.96$ instead of $t_{2\text{-tail, }.05}\,(99) = 1.98$ is

$$106.00 \pm 2.50(1.96), \text{ or } [101.10, 110.90]$$

The precision of the observed mean is the reciprocal of the error variance, or $prc_1 = 1/2.50^2$, which is .16. But because the precision of the prior distribution is $prc_0 = 1/\infty$, or essentially zero, the mean and standard deviation of the posterior distribution given the data, respectively,

$$\mu_1 = 106.00 \text{ and } \sigma_1 = 2.50$$

equal the observed mean and standard error, respectively. The Bayesian 95% credible interval for the random population mean μ calculated in the posterior distribution is

$$106.00 \pm 2.50(1.96), \text{ or } [101.10, 110.90]$$

which defines exactly the same interval as the traditional 95% confidence interval calculated earlier. We can say, based on the data, that the probability is .95 that the interval [101.10, 110.90] includes the true value of μ. But after something is known about the parameter (i.e., there are data), traditional confidence intervals are no longer interpreted this way.

All of the information just described is summarized in the first row of Table 10.1. The remaining rows in the table give the characteristics of the prior and posterior distributions and results in three subsequent samples, each based on 100 cases. For each new result, the posterior distribution from the previous study is taken as the prior distribution for that result. For example, the posterior distribution, given just the results of the first sample, with the characteristics

$$\mu_1 = 106.00, \sigma_1 = 2.50, \text{ and } prc_1 = 1/2.50^2 = .16$$

becomes the prior distribution for the results in the second sample, which are

$$M_2 = 107.50, s_{M_2} = 3.00, prc_2 = 1/3.00^2 = .11$$

TABLE 10.1
Means and Standard Deviations of Prior Distributions and Posterior Distributions Given Data From Four Different Studies

	Prior distribution		Data		Posterior distribution		
Study	*M*	σ	*M*	s_M	μ	σ	95% CI
1	—	∞	106.00	2.50	106.00	2.50	[101.10, 110.90]
2	106.00	2.50	107.50	3.00	106.61	1.92	[102.85, 110.37]
3	106.61	1.92	112.00	2.80	108.33	1.58	[105.32, 111.43]
4	108.33	1.58	109.00	2.50	108.52	1.34	[105.89, 109.86]

Note. The sample size for all studies is $N = 100$. The prior distribution for Study 1 is a flat prior distribution with infinite variance and where no prediction is made about the population mean. CI = Bayesian credible interval.

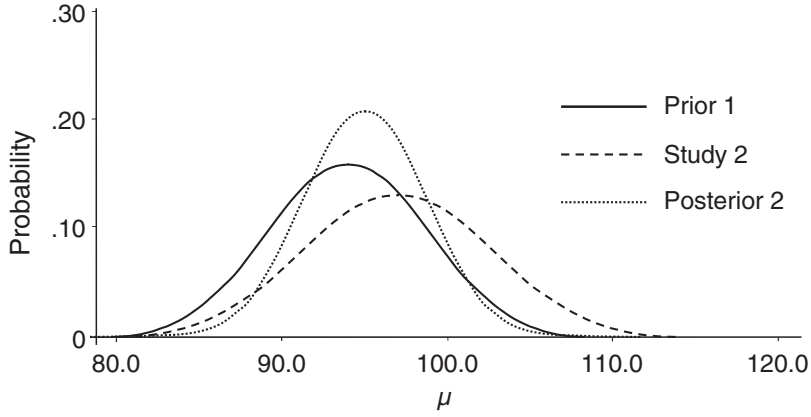

Figure 10.2. Plots of the prior distribution before collecting the second sample, the distribution in the second study, and the posterior distribution for the data in Table 10.1. Prior 1 ($\mu_1 = 106.00$, $\sigma_1 = 2.50$), Study 2 ($M_1 = 107.50$, $s_{M_1} = 3.00$), Posterior 2 ($\mu_2 = 106.61$, $\sigma_2 = 1.92$).

The mean and standard deviation in the posterior distribution, given the results in the first and second samples, are

$$\mu_2 = \left(\frac{.16}{.16+.11}\right)106.00 + \left(\frac{.11}{.16+.11}\right)107.50 = 106.61$$

$$\sigma_2 = \sqrt{\frac{1}{.16+.11}} = 1.92$$

That is, our best single guess for the true population mean has shifted slightly from 106.00 to 106.61 after the second result, and the standard deviation in the posterior distribution is reduced from 2.50 before collecting the second sample to 1.92 after observing the second sample. Our new Bayesian 95% credible interval is [102.85, 110.37], which is slightly narrower than the previous 95% credible interval, [101.10, 110.90].

I used an online plotter by Dienes (2008) to display the prior, sample, and posterior distributions shown in Figure 10.2 for the results just described.[4] This graphic shows the change from the prior to posterior distributions after observing the results in the second sample. The last two rows in Table 10.1 show changes in the prior and posterior distributions as results from two additional samples are synthesized. Note in the table that the widths of the

[4]http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/bayes_normalposterior.swf

posterior distributions get gradually narrower, which indicates decreasing uncertainty with more information.

Conventional meta-analysis and Bayesian analysis are both methods for research synthesis, and it is worthwhile to briefly summarize their relative strengths. Both methods accumulate evidence about a parameter of interest and generate confidence intervals for that parameter. Both methods also allow sensitivity analysis of the consequences of making different kinds of decisions that may affect the results. Because meta-analysis is based on traditional statistical methods, it tests basically the same kinds of hypotheses that are evaluated in primary studies with traditional statistical tests. This limits the kinds of questions that can be addressed in meta-analysis. For example, a standard meta-analysis cannot answer the question, What is the probability that treatment has an effect? It could be determined whether zero is included in the confidence interval based on the average effect size across a set of studies, but this would not address the question just posed. In contrast, there is no special problem in dealing with this kind of question in Bayesian statistics. A Bayesian approach takes into account both previous knowledge and the inherent plausibility of the hypothesis, but meta-analysis is concerned only with the former. It is possible to combine meta-analytical and Bayesian methods in the same analysis (see Howard et al., 2000).

EVALUATION

Bayesian methods are flexible and can evaluate the kinds of questions that researchers would really like answered. An obstacle to their wider use in the behavioral sciences was that many older reference works for Bayesian statistics were quite technical. They often required familiarity with integral notation for probability distributions and estimation techniques for the parameters of different kinds of probability distributions. Such presentations are not accessible for applied researchers without strong quantitative backgrounds. But this situation is changing, and there are now some books that introduce Bayesian methods to a wider audience in the behavioral sciences (e.g., Dienes, 2008).

A second obstacle was the relative paucity of Bayesian software tools for behavioral scientists, but things have improved in this area, too. A freely available software tool for Bayesian analysis is WinBUGS (Bayesian Inference Using Gibbs Sampling; Lunn, Thomas, Best, & Spiegelhalter, 2000) for personal computers.[5] There is an open-source version of WinBUGS that

[5]http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml

runs under the LINUX/UNIX, Microsoft Windows, and Apple Macintosh operating systems.[6] The OpenBugs computer tool will eventually supplant WinBUGS. Wetzels, Raaijmakers, Jakab, and Wagenmakers (2009) described the WinBUGS implementation of a Bayesian *t* test.

Bayesian methods are no more magical than any other set of statistical techniques. One drawback is that there is no direct way in Bayesian estimation to control Type I or Type II errors regarding the dichotomous decision to reject or retain some hypothesis. Researchers can do so in traditional significance testing, but too often they ignore power (the complement of the probability of a Type II error) or specify an arbitrary level of Type I error (e.g., $\alpha = .05$), so this capability is usually wasted. Specification of prior probabilities or prior distributions in Bayesian statistics affects estimates of their posterior counterparts. If these specifications are grossly wrong, the results could be meaningless (e.g., Hurlbert & Lombardi, 2009). But assumptions in Bayesian analyses should be explicitly stated and thus open to scrutiny. Bowers and Davis (2012) criticized the application of Bayesian methods in neuroscience. They noted in particular that Bayesian methods offer little improvement over more standard statistical techniques, but they also noted problems with use of the former, such as the specification of prior probabilities or utility functions in ways that are basically arbitrary. As with more standard statistical methods, Bayesian techniques are not immune to misuse. Overall, behavioral researchers comfortable with structural equation modeling or other statistical modeling techniques should be able to manage the basics of Bayesian estimation. But do give careful thought to the representation of your hypotheses in this approach, which is critical in any kind of analysis.

BEST PRACTICE RECOMMENDATIONS

Summarized next are suggestions for best practices in reporting results from empirical studies; see also Cumming (2012), Ellis (2010), and Ziliak and McCloskey (2008).

1. Do not express research hypotheses solely in terms of statistical significance. For example, to say something like "It is expected that the effect of learning incentive on performance will be *significant*" is to make a hollow prediction, one that betrays confusion of statistical significance with scientific relevance. Instead, state hypotheses in terms of expected directions and magnitudes for effects of interest.

[6]http://www.openbugs.info/w/

2. To make predictions about effect size, you need to have a sense of typical effect sizes observed in previous studies. If there is no meta-analysis, you may be able to compute effect sizes based on descriptive or test statistics reported in primary studies. Doing so is harder for newer research topics for which there are few studies, but estimating effect sizes over even a small number of previous works is still worthwhile.
3. If a relevant meta-analysis is based on a fixed effects model, look skeptically at confidence intervals for average weighted effect sizes. The widths of these intervals may be too narrow if the true model is really a random effects model (see Figure 9.1).
4. Given your outcome measures, estimate minimum effect sizes needed before the results would be considered substantively significant. For example, establish a good-enough belt around zero that marks the boundaries of unappreciable effect sizes. Any result beyond this belt would be considered as potentially of scientific interest.
5. It is harder to relate effect sizes to substantive significance when metrics of outcome variables are arbitrary instead of meaningful. This explains in part why the illusion that statistical significance indicates scientific relevance is so appealing when the metrics of outcome variables have no real-world referents. In treatment outcome studies, where the ultimate goal is to evaluate clinical significance, this problem should motivate researchers to consider alternative outcomes or look deeper into the literature to find meaningful correlates of scores expressed in arbitrary metrics.
6. Select a standardized effect size for results measured in arbitrary metrics that would be most familiar in your research area (e.g., *d*-type vs. *r*-type effect sizes). But report unstandardized effect sizes for outcomes scales in meaningful metrics.
7. In treatment outcome studies, estimate effect size at both the group and case levels. Results of the latter should describe the degree to which treated versus control cases are distinct.
8. Select measures and procedures in ways to reduce measurement error or other sources of irrelevant variance. Do not assume that measures or procedures have acceptable psychometric properties just because they were used in previous studies.
9. If scores on predictor or outcome variables come from psychological tests, estimate and report reliability coefficients in your own sample. If the results for a set of scores are unsatisfactory,

such as $r_{XX} < .50$, then skip analyzing those scores (Little, Lindenberger, & Nesselroade, 1999, described some exceptions to this rule).

10. If it is not possible to estimate score reliabilities in your own sample, then (a) report values of these coefficients given in other studies but (b) explicitly compare characteristics of samples from other studies with those of your own (i.e., justify reliability induction).
11. Build replication into the study plan. One way is to collect sufficient cases until there are both a derivation sample and a cross-replication sample (i.e., conduct internal replication).
12. Follow an analysis plan that respects both theory and results of previous empirical studies but also minimizes the total number of analyses. Avoid "snooping" using statistical tests to find potentially interesting results.
13. Do not hide HARKing—hypothesizing after the results are known—from your readers. That is, do not invent a rationale for the study after conducting preliminary analyses. It is better to explicitly state that the original hypotheses were not supported. Some of the most interesting findings in science have come from studies where expected results were not found. Such "failures" can lead to new discoveries.
14. Select a minimally sufficient statistical technique, or the simplest one that addresses the hypotheses. It is also critical that you understand the output of this technique. Otherwise, you are not ready to use that technique in a meaningful way for your intended audience.
15. If results of significance testing are to be reported, then (a) estimate a prior power for expected population effect sizes and (b) specify α intelligently (e.g., Equation 3.3), not arbitrarily. Be prepared to explain why the assumption of random—or at least representative—sampling in significance testing is not grossly incorrect in your study. Also justify why testing a nil hypothesize is warranted.
16. Report effect sizes for all effects of substantive interest, not just for those that are statistically significant.
17. Do not use the word *significant* in ambiguous ways. For example, never use the word significant without the qualifier *statistically* when describing statistical test outcomes. There is no requirement to use the word significant for results where $p < \alpha$. One context is when p values are reported but not dichotomized relative to some arbitrarily specified level of α.

18. Describe data integrity in the very first paragraph of the Results section. Give information about complications, such as missing data, and steps taken to deal with these problems. Reassure readers that assumptions of statistical techniques, such as requirements for normality or homoscedasticity, are not untenable. If assumptions are violated, describe any intervention, such as transformations, taken to remedy the trouble (Wilkinson & the TFSI, 1999).
19. Do not refer to the T-shirt effect sizes of small, medium, or large, especially if there is no basis in your research area for distinguishing between smaller versus larger effects.
20. Whenever possible, report confidence intervals for effect sizes of substantive interest. Treat the lower and upper bounds of these intervals as estimating the range of effects that are equivalent to your point estimates within the limits of sampling error and assuming that all other sources of error are nil.
21. Evaluate the substantive significance of your observed effect sizes. Doing so may require that you find a way to communicate with stakeholders about how to differentiate between trivial and meaningful results. Without doing so, it is difficult to communicate meaningfully with practitioners, clinicians, managers, or other audiences without strong research backgrounds.
22. Report sufficient summary statistics so that others can, without access to your raw data file, reproduce at least your main analyses. For example, report correlations, standard deviations, and means for all variables in regression analyses, and list cell means, standard deviations, and sizes for each dependent variable in ANOVA. Also report the correlation matrix in each group for within-subjects factors. Make these summaries available online if there is not enough space in a published report to provide this information. Zientek and Thompson (2009) described how following this practice can improve research reports.
23. Even better, make your raw data available online. There are online repositories for data sets in some research areas, such as for clinical drug trials. Otherwise, put your data files on your own web page. Doing so makes a strong statement about openness and transparency. (This recommendation assumes that confidentiality obligations are respected.)
24. Never forget that the point of data analysis is not to report statistically significant results or effect sizes that seem relatively

large. These outcomes are incidental to the real purpose of empirical science, which is to test good ideas about outcomes of potential theoretical, practical, or substantive significance.

CONCLUSION

In an ideal world, students in the behavioral sciences would be taught the basics of Bayesian inference either along with or in lieu of traditional significance testing. This is not to say that Bayesian estimation is without potential limitations. There is no such thing as a perfect inference model that works equally well in all situations. But a Bayesian approach comes closer to the goal of directly evaluating hypotheses in light of the data. Bayesian estimation also requires that all assumptions are explicitly stated in the form of specifications about prior probabilities or distributions. This aspect of Bayesian statistics directly acknowledges the role of *reasoned judgment* in science, which is hidden behind a veneer of objectivity in significance testing.

The main point of this chapter—and that of the whole book—is that there are alternatives to the unthinking overreliance on significance testing that has handicapped the behavioral sciences for so long. And if you have gained new perspectives on your research in the course of reading this book, then I have attained my goals for writing it. In a science fiction story from the 1950s by Alfred Bester (1979), the dark wizard protagonist poses a question to a talented but immature young artist: "It's late. Time to make up your mind. Which will it be? The reality of dreams or the dream of reality?" (p. 221). The artist eventually chooses the hard road of reality and thus opens new prospects. May our own choices about how to analyze data and describe the results be so brave. All the best.

LEARN MORE

Articles by Dienes (2011), Kruschke (2010), and Wetzels et al. (2011) clearly describe applications of Bayesian methods in the behavioral sciences.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6,* 274–290. doi: 10.1177/1745691611406920

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14,* 293–300. doi:10.1016/j.tics.2010.05.001

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 291–298. doi: 10.1177/1745691611406925

REFERENCES

Abelson, R. P. (1997a). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117–141). Mahwah, NJ: Erlbaum.

Abelson, R. P. (1997b). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8,* 12–15. doi:10.1111/j.1467-9280.1997.tb00536.x

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley. doi:10.1002/0470114754

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods, 13,* 515–539. doi:10.1177/1094428109333339

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., Roediger, H. L., III, Scarr, S., . . . Sherman, S. J. (1990). Measurement in psychology: A survey of PhD programs in North America. *American Psychologist, 45,* 721–734. doi:10.1037/0003-066X.45.6.721

Algina, J., & Keselman, H. J. (2003). Approximate confidence intervals for effect sizes. *Educational and Psychological Measurement, 63,* 537–553. doi:10.1177/0013164403256358

Algina, J., Keselman, H. J., & Penfield, R. D. (2005a). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence intervals in the two independent groups case. *Psychological Methods, 10,* 317–328. doi:10.1037/1082-989X.10.3.317

Algina, J., Keselman, H. J., & Penfield, R. (2005b). Effect sizes and their intervals: The two-level repeated measures case. *Educational and Psychological Measurement, 65,* 241–258. doi:10.1177/0013164404268675

Algina, J., Keselman, H. J., & Penfield, R. D. (2006). Confidence intervals for an effect size when variances are not equal. *Journal of Modern Applied Statistical Methods, 5,* 2–13. doi:10.1177/0013164406288161

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Andersen, M. B. (2007). But what do the numbers really tell us? Arbitrary metrics and effect size reporting in sport psychology research. *Journal of Sport & Exercise Psychology, 29,* 664–672. Retrieved from http://journals.humankinetics.com/jsep

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management, 64,* 912–923. doi:10.2307/3803199

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting, 23,* 321–327. doi:10.1016/j.ijforecast.2007.03.004

Aron, A., & Aron, E. N. (2002). *Statistics for the behavioral and social sciences* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Austin, P. C., Mamdani, M. M., Juurlink, D. N., & Hux, J. E. (2006). Testing multiple statistical hypotheses resulted in spurious associations: A study of astrological signs and health. *Journal of Clinical Epidemiology, 59,* 964–969. doi:10.1016/j.jclinepi.2006.01.012

Baguley, T. (2004). An introduction to sphericity. Retrieved from http://homepages.gold.ac.uk/aphome/spheric.html

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100,* 601–617. doi:10.1348/000712608X377117

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437. doi:10.1037/h0020412

Bayes, T. (1763). A letter to John Canton. *Philosophical Transactions of the Royal Society of London, 53,* 293–295.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emory, G. (1979). *Cognitive therapy of depression.* New York, NY: Guilford Press.

Belasco, J., & Stayer, R. (1993). *Flight of the buffalo: Soaring to excellence, learning to let employees lead.* New York, NY: Warner.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10,* 389–396. doi:10.1037/1082-989X.10.4.389

Bellinger, D. C. (2007). Interpretation of small effect sizes in occupational and environmental neurotoxicology: Individual versus population risk. *Neurotoxicology, 28,* 245–251. doi:10.1016/j.neuro.2006.05.009

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

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37,* 325–335. doi:10.1080/01621459.1942.10501760

Bester, A. (1979). 5,271,009. In M. H. Greenberg & J. Olander (Eds.), *Science fiction of the fifties* (pp. 187–221). New York, NY: Avon Books. (Original work published 1954)

Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement, 62,* 197–226. doi:10.1177/0013164402062002001

Bird, K. D., Hadzi-Pavlovic, D., & Isaac, A. (2000). PSY [Computer program]. Retrieved from http://www.psy.unsw.edu.au/research/resources/psyprogram.html

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61,* 27–41. doi:10.1037/0003-066X.61.1.27

Bonett, D. G., & Price, R. M. (2002). Statistical inference for a linear function of medians: Confidence interval, hypothesis testing, and sample size requirements. *Psychological Methods, 7,* 370–383. doi:10.1037/1082-989X.7.3.370

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221–235). New York, NY: Russell Sage Foundation.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2005). Comprehensive Meta-Analysis (Version 2) [Computer software]. Englewood, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley. doi:10.1002/9780470743386

Boring, E. G. (1919). Mathematical vs. scientific importance. *Psychological Bulletin, 16,* 335–338. doi:10.1037/h0074554

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138,* 389–414. doi:10.1037/a0026450

Breaugh, J. A. (2003). Effect size estimation: Factors to consider and mistakes to avoid. *Journal of Management, 29,* 79–97. doi:10.1177/014920630302900106

Brock, F. (2003). The "power" of international business research. *Journal of International Business Studies, 34,* 90–99. doi:10.1057/palgrave.jibs.8400006

Brown, J. S., Bradley, C. S., Subak, L. L., Richter, H. E., Kraus, S. R., & Brubaker, L., . . . Grady, D. (2006). The sensitivity and specificity of a simple test to distinguish between urge and stress urinary incontinence. *Annals of Internal Medicine, 144,* 715–723. Retrieved from http://www.annals.org/

Brown, T. G., Seraganian, P., Tremblay, J., & Annis, H. (2002). Matching substance abuse aftercare treatments to client characteristics. *Addictive Behaviors, 27,* 585–604. doi:10.1016/S0306-4603(01)00195-2

Browne, M. W., & Du Toit, S. H. C. (1991). Models for learning data. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 47–68). Washington, DC: American Psychological Association.

Bruce, C. R., Anderson, M. E., Fraser, S. F., Stepko, N. K., Klein, R., Hopkins, W. G., & Hawley, J. A. (2000). Enhancement of 2000-m rowing performance after caffeine ingestion. *Medicine and Science in Sports and Exercise, 32,* 1958–1963. doi:10.1097/00005768-200011000-00021

Canadian Task Force on Preventive Health Care. (2011). Recommendations on screening for breast cancer in average-risk women aged 40–74 years. *Canadian Medical Association Journal, 183,* 1991–2001. doi: 10.1503/cmaj.110334

Capraro, R. M., & Capraro, M. (2002). Treatments of effect sizes and statistical significance in textbooks. *Educational and Psychological Measurement, 62,* 771–782. doi:10.1177/001316402236877

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York, NY: Guilford Press.

Cartwright, D. (1973). Determinants of scientific progress: The case of research on the risky shift. *American Psychologist, 28,* 222–231. doi:10.1037/h0034445

Carver, R. P. (1978). The case against significance testing. *Harvard Educational Review, 48,* 378–399. Retrieved from http://www.hepg.org/main/her/Index.html

Casscells, W., Schoenberger, A., & Graboys, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299,* 999–1001. doi:10.1056/NEJM197811022991808

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10,* 287–291. doi:10.1016/j.tics.2006.05.007

Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers* (2nd ed.). Hoboken, NJ: Wiley.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine, 19,* 3127–3131. doi:10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M

Christina, R. (2010). *Extreme risk management: Revolutionary approaches to evaluating and measuring risk*. New York, NY: McGraw Hill.

Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153. doi:10.1037/h0045186

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70,* 426–443. doi:10.1037/h0026714

Cohen, J. (1969). *Statistical power analyses for the behavioral sciences*. New York, NY: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003. doi:10.1037/0003-066X.49.12.997

Colliver, J. A., & Markwell, S. J. (2006). ANCOVA, selection bias, statistical equating, and effect size: Recommendations for publication. *Teaching and Learning in Medicine, 18,* 284–286. doi:10.1207/s15328015tlm1804_1

Conn, V. S., & Rantz, M. J. (2003). Research methods: Managing primary study quality in meta-analyses. *Research in Nursing & Health, 26,* 322–333. doi:10.1002/nur.10092

Cook, S., & Wilding, J. (2001). Earwitness testimony: Effects of exposure and attention on the face overshadowing effect. *British Journal of Psychology, 92,* 617–629. doi:10.1348/000712601162374

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs*. Thousand Oaks, CA: Sage.

Crawford, J. R., Garthwaite, P. H., & Betkowska, K. (2009). Bayes' theorem and diagnostic tests in neuropsychology: Interval estimates for post-test probabilities. *Clinical Neuropsychologist, 23,* 624–644. doi:10.1080/13854040802524229

Culpepper, S. A., & Aguinis, H. (2011). Using analysis of covariance (ANCOVA) with fallible covariates. *Psychological Methods, 16,* 166–178. doi:10.1037/a0023355

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3,* 286–300. doi:10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York, NY: Routledge.

Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. *Journal of Cell Biology, 177,* 7–11. doi:10.1083/jcb.200611141

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61,* 532–574. doi:10.1177/00131640121971374

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods, 11,* 217–227. doi:10.1037/1082-989X.11.3.217

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3,* 299–311. doi:10.1207/s15328031us0304_5

Dawes, R. M. (2001). *Everyday irrationality: How pseudo-scientists, lunatics, and the rest of us systematically fail to think rationally.* Cambridge, MA: Westview Press.

Deeks, J. J., & Altman, D. G. (2004). Diagnostic tests 4: Likelihood ratios. *BMJ, 329,* 168–169. doi:10.1136/bmj.329.7458.168

den Hollander, B., Schouw, M., Groot, P., Huisman, H., Caan, M., Barkhof, F., & Reneman, L. (2012). Preliminary evidence of hippocampal damage in chronic users of ecstasy. *Journal of Neurology, Neurosurgery & Psychiatry, 83,* 83–85. doi:10.1136/jnnp.2010.228387

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference.* New York, NY: Palgrave Macmillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6,* 274–290. doi:10.1177/1745691611406920

Dixon, P., & O'Reilly, T. (1999). Scientific versus statistical inference. *Canadian Journal of Experimental Psychology, 53,* 133–149. doi:10.1037/h0087305

Dodd, D. H., & Schultz, R. F. (1973). Computational procedures for estimating magnitude of effect for some analysis of variance designs. *Psychological Bulletin, 79,* 391–395. doi:10.1037/h0034347

Drobatz, K. J. (2009). Measures of accuracy and performance of diagnostic tests. *Journal of Veterinary Cardiology, 11*(Suppl. 1), S33–S40. doi:10.1016/j.jvc.2009.03.004

Dunleavy, E. M., Barr, C. D., Glenn, D. M., & Miller, K. R. (2006). Effect size reporting in applied psychology: How are we doing? *The Industrial-Organizational Psychologist, 43*(4), 29–37. Retrieved from http://www.siop.org/tip/tip.aspx

Easley, R. W., Madden, C. S., & Dunn, M. G. (2000). Conducting marketing science: The role of replication in the research process. *Journal of Business Research, 48,* 83–92. doi:10.1016/S0148-2963(98)00079-4

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242. doi:10.1037/h0044139

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7,* 1–26. doi:10.1214/aos/1176344552

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results.* New York, NY: Cambridge University Press. doi:10.1017/CBO9780511761676

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63,* 591–601. doi:10.1037/0003-066X.63.7.591

Eysenck, H. J. (1995). Meta-analysis squared—Does it make sense? *American Psychologist, 50,* 110–111. doi:10.1037/0003-066X.50.2.110

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5,* 75–98. doi:10.1177/0959354395051004

Faul, F., Erdfelder, E., Lang, A.-G., & 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. doi:10.3758/BF03193146

Ferguson, C. J. (2009). Is psychology research really as good as medical research? Effect size comparisons between psychology and medicine. *Review of General Psychology, 13,* 130–136. doi:10.1037/a0015103

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems. *Journal of Consumer Research, 23,* 89–105. doi:10.1086/209469

Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R., & Thomason, N. (2006). Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conservation Biology, 20,* 1539–1544. doi:10.1111/j.1523-1739.2006.00525.x

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., . . . Schmitt, R. (2005). Toward improved statistical reporting in the *Journal of Consulting and Clinical Psychology*. *Journal of Consulting and Clinical Psychology, 73,* 136–143. doi:10.1037/0022-006X.73.1.136

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science, 15,* 119–126. doi:10.1111/j.0963-7214.2004.01502008.x

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement, 61,* 575–604. doi:10.1177/0013164401614003

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology:* Little evidence of reform. *Educational and Psychological Measurement, 61,* 181–210. doi:10.1177/00131640121971167

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., . . . Goodman, O. (2004). Reform of statistical inference in psychology: The case of *Memory & Cognition*. *Behavior Research Methods, Instruments, & Computers, 36,* 312–324. doi:10.3758/BF03195577

Finch, W. H., & French, B. F. (2012). A comparison of methods for estimating confidence intervals for omega-squared effect size. *Educational and Psychological Measurement, 72,* 68–77. doi:10.1177/0013164411406533

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver & Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh, Scotland: Oliver & Boyd.

Fleiss, J. L., & Berlin, J. A. (2009). Effect sizes for dichotomous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 237–253). New York, NY: Russell Sage Foundation.

Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C., & Fawcett, J. (2010). Antidepressant drug effects and depression severity: A patient-level meta-analysis. *Journal of the American Medical Association, 303,* 47–53. doi:10.1001/jama.2009.1943

Freiman, J. A., Chalmers, T., Smith, H., & Kuebler, R. R. (1978). The importance of beta, the Type II error and sample design in the design and interpretation of the randomized control trial: Survey of 71 negative trials. *New England Journal of Medicine, 299,* 690–694. doi:10.1056/NEJM197809282991304

Friederich, J., Buday, E., & Kerr, D. (2000). Statistical training in psychology: A national survey and commentary on undergraduate programs. *Teaching of Psychology, 27,* 248–257. doi:10.1207/S15328023TOP2704_02

Friedman, G. (2009). *The next 100 years: A forecast for the 21st century*. New York, NY: Doubleday.

Geary, R. C. (1947). Testing for normality. *Biometrika, 34,* 209–242. doi:10.1093/biomet/34.3-4.209

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *American Statistician, 60,* 328–331. doi:10.1198/000313006X152649

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol. 1. Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33,* 587–606. doi:10.1016/j.socec.2004.09.033

Gigerenzer, G., & Murray, D. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Gilbody, S. M., Song, F., Eastwood, A. J., & Sutton, A. (2000). The causes, consequences and detection of publication bias in psychiatry. *Acta Psychiatrica Scandinavica, 102,* 241–249. doi:10.1034/j.1600-0447.2000.102004241.x

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Newbury Park, CA: Sage.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed analysis of variance and covariance. *Review of Educational Research, 42,* 237–288. doi:10.2307/1169991

Gleick, J. (1987). *Chaos: Making a new science.* New York, NY: Viking Penguin.

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *Journal of Experimental Education, 71,* 83–92. doi:10.1080/00220970209602058

Gorard, S. (2006). Towards a judgment-based statistical analysis. *British Journal of Sociology of Education, 27,* 67–80. doi:10.1080/01425690500376663

Gouzoulis-Mayfrank, E., Daumann, J., Tuchtenhagen, F., Pelz, S., Becker, S., Kunert, H.-J., . . . Sass, H. (2000). Impaired cognitive performance in drug free users of recreational ecstasy (MDMA). *Journal of Neurology, Neurosurgery & Psychiatry, 68,* 719–725. doi:10.1136/jnnp.68.6.719

Gray, P. O. (2002). *Psychology* (4th ed.). New York, NY: Worth.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology, 33,* 175–183. doi:10.1111/j.1469-8986.1996.tb02121.x

Grimes, D. A., & Schulz, K. F. (2002). Uses and abuses of screening tests. *Lancet, 359,* 881–884. doi:10.1016/S0140-6736(02)07948-5

Grissom, R. J., & Kim, J. J. (2011). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York, NY: Routledge.

Guthery, F. S., Lusk, J. J., & Peterson, M. J. (2001). The fall of the null hypothesis: Liabilities and opportunities. *Journal of Wildlife Management, 65,* 379–384. doi:10.2307/3803089

Halkin, A., Reichman, J., Schwaber, M., Paltiel, O., & Brezis, M. (1998). Likelihood ratios: Getting diagnostic testing into perspective. *Quarterly Journal of Medicine, 91,* 247–258. doi:10.1093/qjmed/91.4.247

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online, 7*(1), 1–17. Retrieved from http://www.dgps.de/fachgruppen/methoden/mpr-online/

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Harris, R. J. (1997). Reforming significance testing via three-valued logic. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 145–174). Mahwah, NJ: Erlbaum.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? *American Psychologist, 42,* 443–455. doi:10.1037/0003-066X.42.5.443

Herbert, R. (2011). Confidence Interval Calculator [Computer software]. Available from http://www.pedro.org.au/english/downloads/confidence-interval-calculator/

Hoekstra, R., Finch, S., Kiers, H., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and misuse of *p* values. *Psychonomic Bulletin & Review, 13,* 1033–1037. doi:10.3758/BF03213921

Hoffer, E. (1973). *Reflections on the human condition.* New York, NY: Harper & Row.

Holroyd-Leduc, J. M., & Straus, S. E. (2004). Management of urinary incontinence in women: Scientific review. *Journal of the American Medical Association, 291,* 986–995. doi:10.1001/jama.291.8.986

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods, 5,* 315–332. doi:10.1037/1082-989X.5.3.315

Hsu, J. C. (1996). *Multiple comparisons: Theory and methods.* New York, NY: Chapman & Hall.

Hubbard, R., & Armstrong, J. S. (2006). Why we don't really know what "statistical significance" means: A major educational failure. *Journal of Marketing Education, 28,* 114–120. doi:10.1177/0273475306288399

Hubbard, R., Bayarri, M. J., Berk, K. N., & Carlton, M. A. (2003). Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing. *American Statistician, 57,* 171–178. doi:10.1198/0003130031856

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology—and its future prospects. *Educational and Psychological Measurement, 60,* 661–681. doi:10.1177/00131640021970808

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman–Pearson views in textbooks. *Journal of Experimental Education, 61,* 317–333. Retrieved from http://www.tandf.co.uk/journals/titles/00220973.asp

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62,* 227–240. doi:10.1177/0013164402062002002

Huberty, C. J., & Lowman, L. L. (2000). Group overlap as the basis for effect size. *Educational and Psychological Measurement, 60,* 543–563. doi:10.1177/00131640021970718

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis* (2nd ed.). Hoboken, NJ: Wiley.

Hunt, K. (1975). Do we really need more replications? *Psychological Reports, 36,* 587–593. doi:10.2466/pr0.1975.36.2.587

Hunt, M. (1997). *How science takes stock*. New York, NY: Russell Sage Foundation.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman–Pearson decision theory framework and rise of the neoFisherian. *Annales Zoologici Fennici, 46*, 311–349. Retrieved from http://www.sekj.org/AnnZool.html

Hyde, J. S. (2001). Reporting effect sizes: The role of editors, textbook authors, and publication manuals. *Educational and Psychological Measurement, 61*, 225–228. doi:10.1177/0013164401612005

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*, 581–592. doi:10.1037/0003-066X.60.6.581

International Committee of Medical Journal Editors. (2010). Uniform requirements for manuscripts submitted to biomedical journals: Writing and editing for biomedical publication. Retrieved from http://www.icmje.org/urm_full.pdf

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. doi:10.1371/journal.pmed.0020124

Iverson, G. J., & Lee, M. D. (2009). p_{rep} misestimates the probability of replication. *Psychonomic Bulletin & Review, 16*, 424–429. doi:10.3758/PBR.16.2.424

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika, 38*, 324–329. doi:10.1093/biomet/38.3-4.324

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, England: Oxford University Press.

Johnson, D. H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management, 63*, 763–772. doi:10.2307/3802789

Johnson, M. K., & Liebert, R. M. (1977). *Statistics: Tool of the behavioral sciences*. Englewood Cliffs, NJ: Prentice Hall.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude–treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74*, 657–690. doi:10.1037/0021-9010.74.4.657

Kazdin, A. (2006). Arbitrary metrics: Implications for identifying evidence-based treatments. *American Psychologist, 61*, 42–49. doi: 10.1037/0003-066X.61.1.42

Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20*(8). Retrieved from http://www.jstatsoft.org/v20/i08

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*, 137–152. doi:10.1037/a0028086

Kennedy, M. L., Willis, W. G., & Faust, D. (1997). The base-rate fallacy in school psychology. *Journal of Psychoeducational Assessment, 15,* 292–307. doi:10.1177/073428299701500401

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology, 54,* 1–20. doi:10.1348/000711001159357

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13,* 110–129. doi:10.1037/1082-989X.13.2.110

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., . . . Levin, J. R. (1998). Statistical practices of education researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68,* 350–368. doi:10.3102/00346543068003350

Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. M., & Wilcox, R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical and Statistical Psychology, 53,* 175–191. doi:10.1348/000711000159286

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16,* 345–353. doi:10.1111/j.0956-7976.2005.01538.x

Killeen, P. R. (2006). Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review, 13,* 549–562. doi:10.3758/BF03193962

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9,* 137–163. doi:10.1093/oxfordjournals.pan.a004868

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759. doi:10.1177/0013164496056005002

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61,* 213–218. doi:10.1177/00131640121971185

Kirk, R. E. (2012). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage.

Kline, R. B. (2009). *Becoming a behavioral science researcher: A guide to producing research that matters.* New York, NY: Guilford Press.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Kmetz, J. L. (2002). *The skeptic's handbook: Consumer guidelines and a critical assessment of business and management research.* doi:10.2139/ssrn.334180

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14,* 293–300. doi:10.1016/j.tics.2010.05.001

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 299–312. doi:10.1177/1745691611406925

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago, IL: University of Chicago Press.

Kupfersmid, J., & Fiala, M. (1991). A survey of attitudes and behaviors of authors who publish in psychology and education journals. *American Psychologist, 46,* 249–250. doi:10.1037/0003-066X.46.3.249

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory & Psychology, 22,* 67–90. doi:10.1177/0959354311429854

Lenth, R. V. (2006–2009). Java applets for power and sample size. Retrieved from http://www.stat.uiowa.edu/~rlenth/Power

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136,* 1123–1135. doi: 10.1037/a0021276

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods, 4,* 192–211. doi:10.1037/1082-989X.4.2.192

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumptions violations revisited: A quantitative review of alternatives to the one-way analysis of variance *F* test. *Review of Educational Research, 66,* 579–619. doi:10.3102/00346543066004579

Loftus, G. R. (1993). Editorial comment. *Memory & Cognition, 21,* 1–3. doi:10.3758/BF03211158

Longford, N. T. (2005). Editorial: Model selection and efficiency: Is "which model . . . ?" the right question? *Journal of the Royal Statistical Society: Series A, 168,* 469–472. doi:10.1111/j.1467-985X.2005.00366.x

Lunn, D. J., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine, 28,* 3049–3082. doi:10.1002/sim.3680

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10,* 325–337. doi:10.1023/A:1008929526011

Lunneborg, C. (2000). *Modeling experimental and observational data.* Belmont, CA: Duxbury Press.

Lunneborg, C. E. (2001). Random assignment of available cases: Bootstrap standard errors and confidence intervals. *Psychological Methods, 6,* 402–412. doi:10.1037/1082-989X.6.4.402

Lykken, D. T. (1991). What's wrong with psychology, anyway? In D. Cicchetti & W. Grove (Eds.), *Thinking clearly about psychology* (Vol. 1, pp. 3–39). Minneapolis: University of Minnesota Press.

Lytton, H., & Romney, D. M. (1991). Parents' differential socialization of boys and girls: A meta-analysis. *Psychological Bulletin, 109,* 267–296. doi:10.1037/0033-2909.109.2.267

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences.* Stanford, CA: Stanford University Press.

Maserejian, N. N., Lutfey, K. E., & McKinlay, J. B. (2009). Do physicians attend to base rates? Prevalence data and statistical discrimination in the diagnosis of coronary heart disease. *Health Services Research, 44,* 1933–1949. doi:10.1111/j.1475-6773.2009.01022.x

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43,* 679–690. doi:10.3758/s13428-010-0049-5

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163. doi:10.1037/1082-989X.9.2.147

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

McBride, G. B. (1999). Equivalence testing can enhance environmental science and management. *Australian & New Zealand Journal of Statistics, 41,* 19–29. doi:10.1111/1467-842X.00058

McCloskey, D. N., & Ziliak, S. T. (2009). The unreasonable ineffectiveness of Fisherian "tests" in biology, and especially in medicine. *Biological Theory, 4,* 44–53. doi:10.1162/biot.2009.4.1.44

McGrath, R. E. (2011). *Quantitative models in psychology.* Washington, DC: American Psychological Association. doi:10.1037/12316-000

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods, 11,* 386–401. doi:10.1037/1082-989X.11.4.386

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111,* 361–365. doi:10.1037/0033-2909.111.2.361

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction.* New York, NY: Guilford Press.

McLean, J., & Kaufman, A. S. (Eds.). (1998). Statistical significance testing [Special issue]. *Research in the Schools, 5*(2). Retrieved from http://www.msera.org/rits.htm

McWhaw, K., & Abrami, P. C. (2001). Student goal orientation and interest: Effects on students' use of self-regulated learning strategies. *Contemporary Educational Psychology, 26,* 311–329. doi:10.1006/ceps.2000.1054

Meehl, P. E. (1990). Why summaries on research on psychological theories are often uninterpretable. *Psychological Reports, 66* (Monograph Suppl. 1-V66), 195–244. doi:10.2466/PR0.66.1.195-244

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216. doi:10.1037/h0048070

Meilaender, G. (2011). Playing the long season. *First Things, 214,* 19–20. Retrieved from http://www.firstthings.com/

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166. doi:10.1037/0033-2909.105.1.156

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110,* 40–48. doi:10.1037/0021-843X.110.1.40

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review, 16,* 617–640. doi:10.3758/PBR.16.4.617

Montgomery, A. A., Peters, T. J., & Little, P. (2003). Design, analysis and presentation of factorial randomised controlled trials. *BMC Medical Research Methodology, 3,* Article 26. doi:10.1186/1471-2288-3-26

Moons, K. G. M., van Es, G.-A., Deckers, J. W., Habbema, J. D. F., & Grobbee, D. E. (1997). Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: A clinical example. *Epidemiology, 8,* 12–17. doi:10.1097/00001648-199701000-00002

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16,* 406–419. doi:10.1037/a0024377

Morris, S. B., & DeShon, R. P. (1997). Correcting effect sizes computed with factorial analyses of variance for use in meta-analysis. *Psychological Methods, 2,* 192–199. doi:10.1037/1082-989X.2.2.192

Mossman, D., & Berger, J. O. (2001). Intervals for posttest probabilities: A comparison of 5 methods. *Medical Decision Making, 21,* 498–507. doi:10.1177/0272989X0102100608

Myers, J. L., Well, A. D., & Lorch, R. F., Jr. (2010). *Research design and statistical analysis* (3rd ed.). New York, NY: Routledge Academic.

Neal, D. E., Donovan, J. L., Martin, R. M., & Hamdy, F. C. (2009). Screening for prostate cancer remains controversial. *Lancet, 374,* 1482–1483. doi:10.1016/S0140-6736(09)61085-0

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41,* 1299–1301. doi:10.1037/0003-066X.41.11.1299

Nestoriuc, Y., Kriston, L., & Rief, W. (2010). Meta-analysis as the core of evidence-based behavioral medicine: Tools and pitfalls of a statistical approach. *Current Opinion in Psychiatry, 23,* 145–150. doi:10.1097/YCO.0b013e328336666b

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality, 5,* 85–90. Retrieved from http://www.rickcrandall.com/services/jsbp/#posts

Neuliep, J. W., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior and Personality, 8,* 21–29. Retrieved from http://www.rickcrandall.com/services/jsbp/#posts

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine, 17,* 857–872. doi:10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A, 231,* 289–337. doi:10.1098/rsta.1933.0009

Nickerson, R. S. (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

Oakes, M. (1986). *Statistical inference.* New York, NY: Wiley.

O'Keefe, D. J. (2007). Post hoc power, observed power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of power analyses. *Communication Methods and Measures, 1,* 291–299. doi:10.1080/19312450701641375

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25,* 241–286. doi:10.1006/ceps.2000.1040

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8,* 434–447. doi: 10.1037/1082-989X.8.4.434

Oliveri, M., & Calvo, G. (2003). Increased visual cortical excitability in ecstasy users: A transcranial magnetic stimulation study. *Journal of Neurology, Neurosurgery & Psychiatry, 74,* 1136–1138. doi:10.1136/jnnp.74.8.1136

Onwuegbuzie, A. J. (2002). Common analytical and interpretational errors in educational research: An analysis of the 1998 volume of the *British Journal of Educational Psychology. Educational Research Quarterly, 26,* 11–22.

Onwuegbuzie, A. J., & Levin, J. R. (2003). Without supporting statistical evidence, where would reported measures of substantive importance lead? To no good effect. *Journal of Modern Applied Statistical Methods, 2,* 133–151.

Osborne, J. W. (2008). Sweating the small stuff in educational psychology: How effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educational Psychology, 28,* 151–160. doi:10.1080/01443410701491718

Overall, J. E., & Spiegel, D. K. (1969). Concerning least-squares analysis of experimental data. *Psychological Bulletin, 72,* 311–322. doi:10.1037/h0028109

Park, R. L. (2003). The seven warning signs of voodoo science. *Think, 1,* 33–42. doi:10.1017/S1477175600000427

Penfield, R. D., Algina, J., & Keselman, H. J. (2004a). ES Bootstrap: Correlated Groups [Computer software]. Available from http://plaza.ufl.edu/algina/index.programs.html

Penfield, R. D., Algina, J., & Keselman, H. J. (2004b). ES Bootstrap: Independent Groups [Computer software]. Available from http://plaza.ufl.edu/algina/index.programs.html

Penfield, R. D., Algina, J., & Keselman, H. J. (2006). ES Bootstrap 2 [Computer software]. Available from http://plaza.ufl.edu/algina/index.programs.html

Penner, A. M. (2008). Gender differences in extreme mathematical achievement: An international perspective on biological and social factors. *American Journal of Sociology, 114*(Suppl. 1), S138–S170. doi:10.1086/589252

Perkins, N. J., & Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints using two criteria based on the receiver operating characteristic curve. (2006). *American Journal of Epidemiology, 163,* 670–675. doi:10.1093/aje/kwj063

Perlis, A. J. (1982). Epigrams on programming. *ACM SIGPLAN Notices, 17*(9), 7–13. doi:10.1145/947955.1083808

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement, 64,* 916–924. doi:10.1177/0013164404264848

Platt, J. R. (1964, October 16). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science, 146,* 347–353. doi:10.1126/science.146.3642.347

Poitevineau, J., & Lecoutre, B. (2001). The .05 cliff effect may be overstated. *Psychonomic Bulletin & Review, 8,* 847–850. doi:10.3758/BF03196227

Pollard, P. (1993). How significant is "significance"? In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol. 1. Methodological issues* (pp. 449–460). Hillsdale, NJ: Erlbaum.

Pourret, O., Naïm, P., & Marcot, B. (Eds.). (2008). *Bayesian networks: A practical guide to applications.* New York, NY: Wiley.

Pratt, T. C. (2010). Meta-analysis in criminal justice and criminology: What it is, when it's useful, and what to watch out for. *Journal of Criminal Justice Education, 21,* 152–168. doi:10.1080/10511251003693678

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160–164. doi:10.1037/0033-2909.112.1.160

Provalis Research. (1994–2004). SimStat (Version 2.5.8) [Computer software]. Montréal, Québec, Canada: Author.

Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 259–284). Mahwah, NJ: Erlbaum.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21–26. doi:10.3102/0013189X026005021

Robinson, D. H., & Wainer, H. (2002). On the past and future of null hypothesis significance testing. *Journal of Wildlife Management, 66,* 263–271. doi:10.2307/3803158

Rodgers, J. L. (2009). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research, 34,* 441–456. doi:10.1207/S15327906MBR3404_2

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65,* 1–12. doi:10.1037/a0018326

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research*. New York, NY: Cambridge University Press.

Rothman, K. J. (1998). Writing for *Epidemiology*. *Epidemiology, 9,* 333–337. doi:10.1097/00001648-199805000-00019

Rouder, J. N., Speckman, P. L., Sun, D., & Morey, R. D. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. doi:10.3758/PBR.16.2.225

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416–428. doi:10.1037/h0042040

Rutherford, A. (2011). *ANOVA and ANCOVA: A GLM approach*. Hoboken, NJ: Wiley.

Rutledge, T., & Loh, C. (2004). Effect sizes and statistical testing in the determination of clinical significance in behavioral medicine research. *Annals of Behavioral Medicine, 27,* 138–145. doi:10.1207/s15324796abm2702_9

Sagan, C. (1996). *The demon-haunted world: Science as a candle in the dark*. New York, NY: Random House.

Sahai, H., Khurshid, A., Ojeda, M. M., & Velasco, F. (2009). Simultaneous confidence intervals for variance components in two-way balanced crossed classification random effects model with interaction. *Revista Investigación Operacional, 30,* 250–265. Retrieved from http://rev-inv-ope.univ-paris1.fr/

Salibián-Barrera, M., & Zamar, R. H. (2002). Bootstrapping robust estimates of regression. *Annals of Statistics, 30,* 556–582. doi:10.1214/aos/1021379865

Schervish, M. J. (1996). *p* values: What they are and what they are not. *American Statistician, 50,* 203–206. doi:10.2307/2684655

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1,* 115–129. doi:10.1037/1082-989X.1.2.115

Schmidt, F. L. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science, 5,* 233–242. doi:10.1177/1745691610369339

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: Erlbaum.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13,* 90–100. doi:10.1037/a0015108

Schuster, C., & von Eye, A. (2001). The relationship of ANOVA models with random effects and repeated measurement designs. *Journal of Adolescent Research, 16,* 205–220. doi:10.1177/0743558401162006

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York, NY: Wiley.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316. doi:10.1037/0033-2909.105.2.309

Seggar, L. B., Lambert, M. J., & Hansen, N. B. (2002). Assessing clinical significance: Application to the Beck Depression Inventory. *Behavior Therapy, 33,* 253–269. doi:10.1016/S0005-7894(02)80028-4

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. New York, NY: Houghton Mifflin.

Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science, 8,* 1–2. doi:10.1111/j.1467-9280.1997.tb00533.x

Simel, D. L., Samsa, G. P., & Matchar, D. B. (1991). Likelihood ratios with confidence: Sample size estimation for diagnostic test studies. *Journal of Clinical Epidemiology, 44,* 763–770. doi:10.1016/0895-4356(91)90128-V

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

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760. doi:10.1037/0003-066X.32.9.752

Smithson, M. (2003). *Confidence intervals*. Thousand Oaks, CA: Sage.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61,* 334–349. Retrieved from http://www.tandfonline.com/toc/vjxe20/current

Sohn, D. (2000). Significance testing and the science. *American Psychologist, 55,* 964–965. doi:10.1037/0003-066X.55.8.964

Spence, G. (1995). *How to argue and win every time: At home, at work, in court, everywhere, everyday*. New York, NY: St. Martin's Press.

Statistics.com. (2009). *Resampling Stats (Version 4)* [Computer software]. Arlington, VA: Author.

Steering Committee of the Physicians' Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing Physicians' Health Study. *New England Journal of Medicine, 318,* 262–264. doi:10.1056/NEJM198801283180431

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9,* 164–182. doi:10.1037/1082-989X.9.2.164

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger

(Eds.), *What if there were no significance tests?* (pp. 221–257). Mahwah, NJ: Erlbaum.

Sterner, W. R. (2011). What is missing in counseling research? Reporting missing data. *Journal of Counseling & Development, 89,* 56–62. doi:10.1002/j.1556-6678.2011.tb00060.x

Stevens, J. J. (1999). Interaction effects in ANOVA. Retrieved from http://pages.uoregon.edu/stevensj/interaction.pdf

Stigler, S. M. (1978). Francis Ysidro Edgeworth, statistician. *Journal of the Royal Statistical Society, Series A, 141,* 287–322. doi:10.2307/2344804

Stigler, S. M. (1986). *The history of statistics.* Cambridge, MA: Belknap.

Streiner, D. L. (1996). Maintaining standards: Differences between the standard deviation and standard error, and when to use each. *Canadian Journal of Psychiatry, 41,* 498–502. Retrieved from http://publications.cpa-apc.org/browse/sections/0

Student [W. S. Gosset]. (1927). Errors of routine analysis. *Biometrika, 19,* 151–164. doi:10.2307/2332181

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology, 102,* 989–1004. doi:10.1037/a0019507

Sutcliffe, A. (2002). *User-centred requirements engineering.* London, England: Springer-Verlag. doi:10.1007/978-1-4471-0217-5

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers.* Mahwah, NJ: Erlbaum.

Tabachnick, B. G., & Fidell, L. S. (2001). *Computer-assisted research design and analysis.* Boston, MA: Allyn & Bacon.

The Canadian Press. (2011, November 25). Controversy over new mammogram guidelines continues. Retrieved from http://www.ctv.ca/CTVNews/Health/20111125/mammogram-guidelines-111125/

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling & Development, 70,* 434–438. doi:10.1002/j.1556-6676.1992.tb01631.x

Thompson, B. (Ed.). (1993). Statistical significance testing in contemporary practice [Special issue]. *Journal of Experimental Education, 61*(4). Retrieved from http://www.tandfonline.com/loi/vjxe20

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55,* 525–534. doi:10.1177/0013164495055004001

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26–30. doi:10.3102/0013189X025002026

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26*(5), 29–32. doi:10.3102/0013189X026005029

Thompson, B. (1999). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review, 11,* 157–169. doi:10.1023/A:1022028509820

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 70,* 80–93. doi:10.1080/00220970109599499

Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues.* Thousand Oaks, CA: Sage.

Thompson, B. (2006a). *Foundations of behavioral statistics: An insight-based approach.* New York, NY: Guilford Press.

Thompson, B. (2006b). Research synthesis: Effect sizes. In J. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 583–603). Washington, DC: American Educational Research Association.

Thompson, S. K. (2012). *Sampling* (3rd ed.). Hoboken, NJ: Wiley.

Thompson, W. L. (2001). 402 citations questioning the indiscriminate use of null hypothesis significance tests in observational studies. Retrieved from htttp://warnercnr.colostate.edu/~anderson/thompson1.html

Toffler, A. (1970). *Future shock.* New York, NY: Random House.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6,* 371–386. doi:10.1037/1082-989X.6.4.371

Tryon, W. W., & Lewis, C. (2008). An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods, 13,* 272–277. doi:10.1037/a0013158

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Tukey, J. W., & McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization 1. *Sankhya, Series A, 25,* 331–352. Retrieved from http://sankhya.isical.ac.in/index.html

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110. doi:10.1037/h0031322

U.S. Census Bureau. (2010). *Door-to-door visits begin for 2010 census* [Press release]. Retrieved from http://2010.census.gov/news/releases/operations/door-to-door-visits-begin.html

Vacha-Haase, T., & Ness, C. N. (1999). Statistical significance testing as it relates to practice: Use within *Professional Psychology: Research and Practice. Professional Psychology: Research and Practice, 30,* 104–105. doi:10.1037/0735-7028.30.1.104

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization. *Measurement and Evaluation in Counseling and Development, 44,* 159–168. doi:10.1177/0748175611409845

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the *CL* common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25,* 101–132. doi:10.3102/10769986025002101

Vaughan, G. M., & Corballis, M. C. (1969). Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin, 72,* 204–213. doi:10.1037/h0027878

Viechtbauer, W. (2007). Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples design. *Journal of Educational and Behavioral Statistics, 32,* 39–60. doi:10.3102/1076998606298034

Wayne, J. H., Riordan, C. M., & Thomas, K. M. (2001). Is all sexual harassment viewed the same? Mock juror decisions in same- and cross-gender cases. *Journal of Applied Psychology, 86,* 179–187. doi:10.1037/0021-9010.86.2.179

Webb, A. J. S., Fischer, U., & Rothwell, P. M. (2011). Effects of β-blocker selectivity on blood pressure variability and stroke: A systematic review. *Neurology, 77,* 708–709. doi:10.1212/WNL.0b013e31822b007a

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29,* 350–362. doi:10.1093/biomet/29.3-4.350

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Boca Raton, FL: Chapman & Hall. doi:10.1201/EBK1439808184

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 291–298. doi:10.1177/1745691611406923

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian *t* test. *Psychonomic Bulletin & Review, 16,* 752–760. doi:10.3758/PBR.16.4.752

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology, 75,* 1182–1189. doi:10.1111/j.1365-2656.2006.01141.x

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300–314. doi:10.1037/0003-066X.53.3.300

Wilcox, R. R. (2003). *Applying contemporary statistical techniques.* New York, NY: Academic Press.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). San Diego, CA: Academic Press.

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8,* 254–274. doi:10.1037/1082-989X.8.3.254

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604. doi:10.1037/0003-066X.54.8.594

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). Boston, MA: McGraw-Hill.

Wood, M. (2005). Bootstrapped confidence intervals as an approach to statistical inference. *Organizational Research Methods, 8,* 454–470. doi:10.1177/1094428105280059

Yuen, K. K. (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika, 61,* 165–170. doi:10.1093/biomet/61.1.165

Zientek, L. R., & Thompson, B. (2009). Matrix summaries improve research reports: Secondary analyses using published literature. *Educational Researcher, 38,* 343–352. doi:10.3102/0013189X09339056

Ziliak, S., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.

INDEX

ABOUT THE AUTHOR

Rex B. Kline, **PhD**, is a professor of psychology at Concordia University in Montréal, Canada. He has a doctorate in clinical psychology. His areas of research and writing include the psychometric evaluation of cognitive abilities, cognitive and scholastic assessment of children, structural equation modeling, the training of behavioral science researchers, and usability engineering in computer science. He has published six books and nine chapters in these areas (see http://tinyurl.com/rexkline).