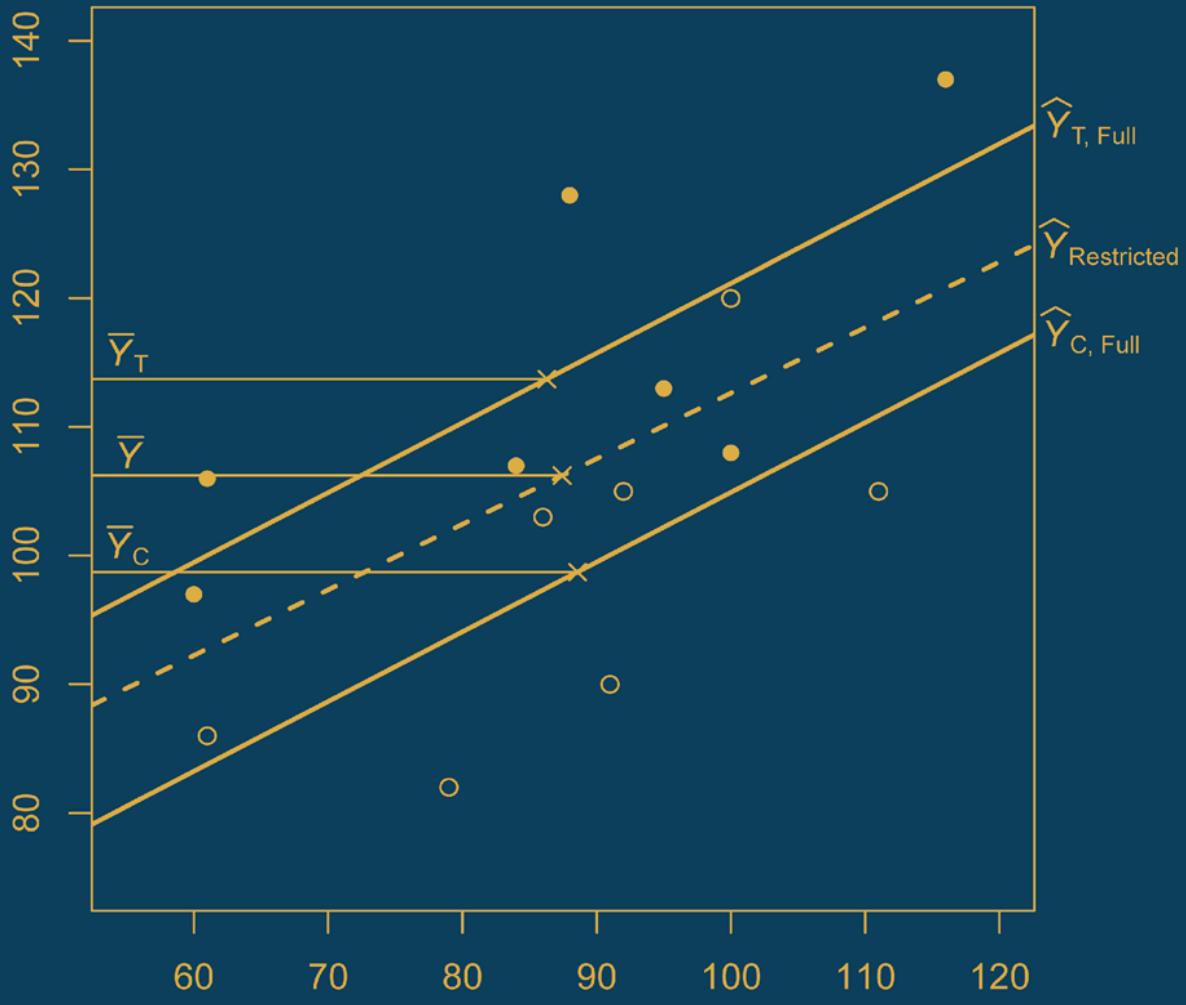
Designing Experiments and Analyzing Data

THIRD EDITION

A MODEL COMPARISON PERSPECTIVE



SCOTT E. MAXWELL, HAROLD D. DELANEY, and KEN KELLEY

ROUTLEDGE

DESIGNING EXPERIMENTS AND ANALYZING DATA

Designing Experiments and Analyzing Data: A Model Comparison Perspective (Third edition) offers an integrative conceptual framework for understanding experimental design and data analysis. Maxwell, Delaney, and Kelley first apply fundamental principles to simple experimental designs followed by an application of the same principles to more complicated designs. Their integrative conceptual framework better prepares readers to understand the logic behind a general strategy of data analysis that is appropriate for a wide variety of designs, which allows for the introduction of more complex topics that are generally omitted from other books. Numerous pedagogical features further facilitate understanding: **examples of published research** demonstrate the applicability of each chapter's content; **flowcharts** assist in choosing the most appropriate procedure; **end-of-chapter lists of important formulas** highlight key ideas and assist readers in locating the initial presentation of equations; **useful programming code and tips** are provided throughout the book and in associated resources available online; and **extensive sets of exercises** help develop a deeper understanding of the subject. **Detailed solutions** for some of the exercises and **realistic data sets** are included on the website (*DesigningExperiments.com*). The pedagogical approach used throughout the book enables readers to gain an overview of experimental design, from conceptualization of the research question to analysis of the data. The book and its **companion websites** with web apps, tutorials, and detailed code are ideal for students and researchers seeking the optimal way to design their studies and analyze the resulting data.

Scott E. Maxwell is the Fitzsimons Professor of Psychology at the University of Notre Dame. His research interests are in the areas of research methodology and applied behavioral statistics, with much of his recent work focusing on statistical power and accuracy in parameter estimation, especially in randomized designs. He has served as editor of *Psychological Methods*; received the Samuel J. Messick Award for Distinguished Scientific Contributions by the American Psychological Association's Division of Evaluation, Measurement, and Statistics; and has received multiple teaching awards.

Harold D. Delaney is Emeritus Professor of Psychology at the University of New Mexico, where he received the University's Outstanding Graduate Teacher of the Year award for his course on experimental design and analysis, and where he directed the Psychology Honors program for 30 years. His research interests in applied statistics include methods that accommodate individual differences among people. He received a Fulbright Award from the U.S. Department of State to spend an academic year lecturing in Budapest, Hungary, and continues to offer courses there.

Ken Kelley is Professor of Information Technology, Analytics, and Operations (ITAO) and the Associate Dean for Faculty and Research in the Mendoza College of Business at the University of Notre Dame. His work is on quantitative methodology, where he focuses on the development, improvement, and evaluation of statistical methods and measurement issues. He is an Accredited Professional Statistician (PStat®); recipient of the Anne Anastasi early career award by the APA's Division of Evaluation, Measurement, and Statistics; a fellow of the American Psychological Association; elected member of the Society of Multivariate Experimental Psychology; and an award-winning teacher.

SEM... to Katy
HDD... to Nancy
KK... to Kenny and Corinne

DESIGNING EXPERIMENTS AND ANALYZING DATA

A Model Comparison Perspective

Third Edition

Scott E. Maxwell,
Harold D. Delaney, and Ken Kelley

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Third edition published 2018
by Routledge
711 Third Avenue, New York, NY 10017

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

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

First edition published by Wadsworth Pub Co 1990
Second edition published by Routledge 2003

Library of Congress Cataloging-in-Publication Data
Names: Maxwell, Scott E. | Delaney, Harold D. | Kelley, Ken (Professor of information technology)
Title: Designing experiments and analyzing data : a model comparison perspective.
Description: Third edition / Scott E. Maxwell, Harold D. Delaney, and Ken Kelley. | New York, NY : Routledge, 2017. | Includes bibliographical references and index.
Identifiers: LCCN 2017001424 | ISBN 9781138892286 (hard back : alk. paper) | ISBN 9781315169781 (ebook)
Subjects: LCSH: Experimental design.
Classification: LCC QA279 .M384 2017 | DDC 519.5/3—dc23
LC record available at https://lccn.loc.gov/2017001424

ISBN: 978-1-138-89228-6 (hbk)
ISBN: 978-1-315-16978-1 (ebk)

Typeset in Times New Roman
by Apex CoVantage, LLC

Visit the companion website *DesigningExperiments.com* and www.routledge.com/cw/maxwell

Contents

Preface

Designing Experiments and Analyzing Data: A Model Comparison Perspective is written to serve as a textbook or a reference book on designing experiments and analyzing experimental data. The methods we discuss are appropriate in a variety of scientific research areas, especially psychology and related disciplines. The book is centered around the view of data analysis as involving a comparison of models. We believe that such *a model comparison perspective* offers important advantages over the traditional variance partitioning approach often used to teach analysis of variance and related methods. Instead of approaching each experimental design in terms of its own unique set of computational formulas as if it were fundamentally different, the model comparison approach allows us to introduce a few basic formulas that can be applied with the same underlying logic to every experimental design. Our approach establishes an integrative framework that highlights how various designs and analyses are related to one another. The model comparison approach also allows us to cover topics that are often omitted in experimental design texts. For example, we are able to introduce the multivariate approach to repeated measures as a straightforward generalization of the approach used for between-subjects designs. Similarly, the analysis of nonorthogonal designs (designs with unequal cell sizes) fits nicely with our approach. Further, not only is the presentation of the standard analysis of covariance facilitated by the model comparison perspective, but we are also able to consider models that allow for heterogeneity of regression across conditions. In fact, the underlying logic can be applied directly to even more complex methods such as mixed-effects or hierarchical linear models, which we discuss, and also to other methods such as structural equation modeling.

The focus throughout the book is conceptual, with our greatest emphasis being on promoting an understanding of the logical underpinnings of design and analysis. This is perhaps most evident in the first part of the book dealing with the conceptual bases of design and analysis, which touches on relevant issues in philosophy of science and past and current controversies in statistical reasoning. But the conceptual emphasis continues throughout the book, in which our primary concern is with developing an understanding of the logic of statistical methods. This is why we present definitional instead of computational formulas, as we generally rely on statistical software to perform actual computations. This emphasis allows us to concentrate on the meaning of what is being computed instead of focusing on how to perform calculations. Nevertheless, we recognize the importance of doing hand calculations on occasion to better understand what it is that is being computed. Thus, we have included a number of exercises at the end of each chapter

that give the reader the opportunity to calculate quantities by hand on small data sets. We have also included many thought questions which are intended to develop a deeper understanding of the subject and to help the reader draw out logical connections in the materials. Finally, we provide larger actual or realistic data sets described in the published literature that allow the reader to experience an analysis of data from each design in its entirety.

There is a companion website for the book, *DesigningExperiments.com*, which contains example SAS code, IBM SPSS Statistics instructions (syntax and graphically illustrating point-and-click options), and step-by-step R code for replicating many of the analyses presented in the book. The data sets used in the chapters are also available at *DesigningExperiments.com/Data* as well as in the accompanying R package, AMCP (for "A Model Comparison Perspective"), which is available from CRAN (the Comprehensive R Archive Network). We have not provided SAS, IBM SPSS Statistics, or R code for end-of-chapter exercises because we believe that most instructors would prefer that students have the opportunity to develop appropriate computer code for these exercises themselves based on examples from the chapters instead of being given all of the answers. The data sets for the chapter exercises are also available at *Designing Experiments.com/Data* and in the AMCP R package. Solutions to numerous selected (marked by asterisks in the book) exercises are provided at *DesigningExperiments.com/Solutions*. Answers for the remaining exercises as well as other resources such as PowerPoint slides are available for instructors who adopt the book for classroom use.

Despite the inclusion of advanced topics, the only necessary mathematical background for the book is high school algebra. However, we do assume that readers will have had at least one undergraduate statistics course. For those readers needing a refresher of statistics, a review of basic statistics is also included at *DesigningExperiments.com/Supplements*. Even those who have had more than a single statistics course may find the *Review of Basic Statistics* helpful, particularly in conjunction with beginning the development of our model comparison approach in Chapter 3. We also provide another statistical tutorial, a discussion of regression that is also included on the website at *DesigningExperiments.com/Supplements*. The regression tutorial is most profitably read upon the completion of Chapter 3, as it provides a basic discussion of regression for those who have not previously studied or need a review of regression.

ORGANIZATION

The organization of the book allows chapters to be covered in various sequences or omitted entirely.

Part I (Chapters 1 and 2) explains the logic of experimental design and the role of randomization in the conduct of behavioral research. These two chapters attempt to provide the philosophical and historical context in which the methods of experimental design and analysis may be understood. Although Part I is not required for understanding statistical issues in the remaining chapters of the book, it does help the reader see the "big picture."

Part II provides the core of the book. Chapter 3 introduces the concept of comparing full and restricted models. Most of the formulas used throughout the book are introduced in Chapters 3 and 4. Although most readers will want to follow these two chapters by reading at least Chapters 5, 7, and 8 in Part II, it would be possible for more advanced readers to go straight to Chapters 13 and 14 on the multivariate approach to repeated measures. Chapter 9, on analysis of covariance, is written in such a way that it can be read either immediately following Chapter 8 or deferred until after Part III.

Part III describes design and analysis principles for within-subjects designs (that is, repeated measures designs). These chapters are written to provide maximum flexibility in choosing an approach to the topic. For a one-semester experimental design course, instructors may choose

to omit one of the four chapters on repeated measures. Covering only Chapters 11, 13, and 14 introduces the univariate approach to repeated measures but covers the multivariate approach in greater depth. Alternatively, covering only Chapters 11, 12, and 13 emphasizes the univariate approach. Advanced readers might skip Chapters 11 and 12 entirely and read only Chapters 13 and 14.

Part IV, consisting of Chapters 15 and 16, presents a basic introduction to mixed-effects models (also called hierarchical models or multilevel models). Chapter 15 extends Chapters 11 through 14 by developing additional models for longitudinal data. Chapter 16, an extension of Chapter 10, applies mixed-effects models to nested designs. This type of model has several advantages over traditional ANOVA approaches, including the possibility of modeling data at individual and group levels simultaneously, as well as permitting the inclusion of participants with incomplete data in analyses of repeated measures designs. We explicitly describe how these models are related to the traditional ANOVA and MANOVA models covered in previous chapters. This contrasts with many other presentations of such models, which either relate these models to regression but not ANOVA or present them in isolation from any form of more traditional models. In a two-quarter or two-semester course, one might cover not only all four chapters on ANOVA approaches to repeated measures, but also Chapters 15 and 16. Alternatively, these final two chapters might be used in the first part of a subsequent course devoted to mixed-effects models.

As in the first and second editions, discussion of more specialized topics is included but is now made available in a variety of ways. Brief sections explicating specific ideas within chapters are marked with an "Optional" heading; the optional sections we deemed more critical are included in the book; others are included at *DesigningExperiments.com/Supplements*, with a listing of any such material pertinent to a chapter given at the end of the chapter. A more involved discussion of methods relevant to a whole chapter is denoted as an Extension to the chapter; the extension on heterogeneity of regression is included at the end of Chapter 9, while other chapter extensions are available on *DesigningExperiments.com/Supplements*. Detailed notes on individual ideas presented in the text are provided in the chapter endnotes.

We have taken several steps to make key equations interpretable and easy to use. The most important equations are numbered consecutively in each chapter as they are introduced. If the same equation is repeated later in the chapter, we use its original equation number followed by the designation "repeated," to remind the reader that this equation was already introduced and to facilitate finding the point where it was first presented.

Finally, we have frequently provided tables that summarize important equations for a particular design or concept, to make equations easier to find and facilitate direct comparisons of the equations to enhance understanding of their differences and similarities.

CHANGES IN THIS EDITION

Especially for those who used the first or second editions of the book, we want to highlight important changes included in this edition.

Important pedagogical and organizational changes include:

- We begin each chapter with an overview that introduces the types of questions that can be addressed with the methods of the chapter.
- We cite a specific example of published research illustrating the chapter's content.
- We have reworked and added additional end-of-chapter exercises.
- Detailed instructions are provided online to illustrate applications of methods discussed in the book.

- We include at the end of each chapter a summary listing of important formulas used in the chapter.
- We also list at the end of each chapter resources included at *DesigningExperiments.com* relevant to that chapter.
- We provide an improved website, *DesigningExperiments.com*, with example R code to implement many of the analyses in the book and the AMCP R package.
- *DesigningExperiments.com* also includes a number of web apps that allow easy computation of, for example, confidence intervals for effect sizes.
- The companion website, which is more robust and informative than the version that accompanied the second edition, is itself a learning tool. We have leveraged *DesigningExperiments.com* to make the book smaller than it would have been if everything were included in the physical book; numerous optional sections and supplementary materials, such as the appendix on the relationship between ANOVA and regression models and general principles of formulating models that appeared in the second edition, now have been moved to *DesigningExperiments.com/Supplements*.

Some of the more important changes in content include the following:

- Discussion of the historical and philosophical context of experimental design and analysis has been updated, and now includes consideration of concerns over the reproducibility of psychological science.
- Expanded treatment of confidence intervals, including confidence intervals for population effect sizes.
- Coverage of methods for correcting for bias in some effect size measures like *d* and *f*.
- Expanded discussion of power analyses including the value of considering varying estimates of effect size in power analyses and determining the probability that power will be at least a specified value.
- Inclusion of a new section on propensity score analysis.
- Expanded discussion of heterogeneity of regression including introduction of methods for determining regions of the covariate where there is evidence for a significant difference between treatment conditions.
- Expanded discussion of how seemingly different models are related.
- Improved clarity throughout.
- Updated citing and discussion of relevant scholarly research in statistics and methodology.

ACKNOWLEDGMENTS

The number of individuals who contributed either directly or indirectly to this book's development defies accurate estimation. The advantages of the model comparison approach were first introduced to two of us (SEM and HDD) by Elliot Cramer when we were graduate students at the University of North Carolina at Chapel Hill. The excellent training we received there provided a firm foundation on which to build. Much of the philosophy underlying our approach can be traced to Elliot and our other mentors at the L. L. Thurstone Psychometric Lab (Mark Appelbaum, John Carroll, Lyle Jones, and Tom Wallsten). More recently we have benefited from insightful comments from colleagues who used the first or second edition, as well as many current and former students and teaching assistants. One former teaching assistant, Eric Kruger, was of great help in simulating data sets used in exercises and suggesting R code for selected analyses for the current edition. Tessa Cappelle's careful work in updating the index was also

much appreciated. We are also indebted to the Department of Psychology at the University of Notre Dame (SEM) and the Department of Psychology (HDD) at the University of New Mexico for providing sabbatical leaves to work on the book. The encouragement of our colleagues must be mentioned, especially that of David Cole, George Howard, Tim Goldsmith, Bill Miller, and Katie Witkiewitz. We appreciate the support of the staff at Taylor & Francis/Routledge, including former editor Debra Riegert and current editor Paul Dukes. We are still indebted to those that provided comments and support for the first two editions. Excellent support for this edition was rendered by Maggie Neenan-Michel and her team in the Mendoza College of Business Faculty Support office, especially Tamara (Tami) Springer, Diane Stauffer, and Laura Gerber. Additionally, Heather Denton, also from the Mendoza College of Business, provided helpful support.

The current edition, like the first two, builds on the many worthwhile suggestions of a number of reviewers, including the 17 anonymous reviewers who provided feedback on our plan for this edition. We remain indebted to those who provided comments either on the first or second editions, including: David A. Kenny, University of Connecticut; David J. Francis, University of Houston; Richard Gonzalez, University of Michigan; Sam Green and Stephen West, Arizona State University; Joe Rodgers and Howard M. Sandler, Vanderbilt University; András Vargha, Károli Gáspár University (Budapest); Ron Serlin, University of Wisconsin; James E. Carlson, Auburn University at Montgomery; James Jaccard, State University of New York at Albany; Willard Larkin, University of Maryland, College Park; K. J. Levy, State University of New York at Buffalo; Marjorie Marlin, University of Missouri, Columbia; Ralph G. O'Brien, Cleveland Clinic; Edward R. Stearns, California State University, Fullerton; Rand Wilcox, University of Southern California; Rhonda K. Kowalchuk, University of Wisconsin, Milwaukee; Keith F. Widaman, University of California, Davis; and Jon Williams, Kenyon College. We would also like to thank the following individuals who graciously shared data from their published studies with us: Lara Aknin, James Bray, Nathan Carnes, Shirley Crotwell, Brad Gibson, Nicole McNeil, Rolf Zwaan, and Anita Zwaan-Eerland. Finally, we also utilized data that had been made publicly available by the following authors: Emily Holmes, Xiaoquing Hu, Marijn C. W. Kroes, and Robert Rosenthal.

DESIGNING EXPERIMENTS AND ANALYZING DATA

A Model Comparison Perspective

Third Edition

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

I

Conceptual Bases of Experimental Design and Analysis

Man, being the servant and interpreter of Nature, can do and understand so much, and so much only, as he has observed, in fact or in thought, of the course of Nature. . . . Human knowledge and human power meet in one; for where the course is not known, the effect cannot be produced. Nature, to be commanded, must be obeyed.

—FRANCIS BACON, *NOVUM ORGANUM*, 1620

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1

The Logic of Experimental Design and Analysis

OVERVIEW OF CHAPTER: RESEARCH QUESTIONS ADDRESSED

Methods of experimental design and data analysis derive their value from the contributions they make to the more general enterprise of science. To appreciate what design and analysis can and cannot do, it is necessary to understand something of the logic of science. Although a comprehensive introduction to the philosophy of science is beyond the scope of this work, we believe it is appropriate to provide in this opening chapter some historical and philosophical context for the statistical methodology to be developed in subsequent parts of the book.

The first half of this chapter deals with philosophy of science, and opens with a discussion of the traditional view of science. We next mention some of the difficulties inherent in this view, and consider various responses that have been offered to the critique of the traditional view. In the second half of the chapter we develop how statistical decisions can be regarded as part of an organized argument warranting an inductive inference about some part of reality. Thus, the first half of the chapter attempts to address the question of "How does mathematical and statistical modeling relate to the more general enterprise of science?" In the second half of the chapter our focus is on the logic of statistical reasoning per se. Here we address questions like: What does a *p* value mean and what justifies its use? Can statistical inferences based on convenience samples rather than random sampling from a population be legitimate? What factors contribute to the fact that attempts to replicate published findings often either are deemed failures or yield estimated effect sizes that are smaller than those originally reported?

PUBLISHED EXAMPLE

Rodgers (2010) discusses both historical and recent developments in quantitative methodology in the behavioral sciences in his insightful *American Psychologist* paper titled "The Epistemology of Mathematical and Statistical Modeling: A Quiet Methodological Revolution." Rodgers argues that the role of statistics in science is appropriately understood in the framework of a broad philosophy of science. In particular, given that a major task of science is the development of theory, and that theories can often be helpfully and precisely instantiated in mathematical

models, a key feature of scientific epistemology (or how we can come to know and understand aspects of reality through science) is through developing mathematical models and evaluating them statistically. Such models highlight some aspects of reality and ignore others, and are evaluated primarily by comparison with competing models. Rodgers argues that the controversy about null hypothesis significance testing during the 1990s was in part unnecessary because of the quiet modeling revolution already underway within methodology to view statistics as an aid to building and evaluating models. Indeed, the model comparison approach we have taken in the various editions of this book is, as Rodgers noted, consistent with this quiet revolution and hopefully will prepare readers for the types of model comparisons underlying more advanced methodologies such as structural equation modeling.

PHILOSOPHY OF SCIENCE

The Traditional View of Science

The perspective on science that emerged in the West around 1600 and that profoundly shaped and defined the modern era (Whitehead, 1932) can be identified in terms of its methodology: empirical observation and, whenever possible, experimentation. The essence of experimentation, as Shadish, Cook, & Campbell (2002) note, is an attempt "to discover the effects of presumed causes" (p. 3). It is because of their contribution to the understanding of causal processes that experiments play such a central role in science. As Schmidt (1992) suggests, "The major task in any science is the development of theory. . . . Theories are causal explanations. The goal in every science is explanation, and explanation is always causal" (p. 1177). The explication of statistical methods that can assist in the testing of hypothesized causes and estimating their effects via experiments is the primary concern of this book. Such an emphasis on technical language and tools is characteristic of modern science and perhaps contributes to the popular perception of science as a purely objective, rule-governed process. It is useful to review briefly how such a view arose historically and how it must be qualified.

Many trace the origins of modern science to the British statesman and philosopher Sir Francis Bacon (1561–1626). The context in which Bacon was writing was that of a culture that for centuries had been held in the grips of an Aristotelian approach to obtaining knowledge. Although Aristotle had considered induction or making inferences from particular observations, the "predominant mode of his logic was deduction, and its ideal was the syllogism" (Durant & Durant, 1961, p. 174). Bacon recognized the stagnation that had resulted in science because of this stress on deduction rather than observation and because the ultimate appeal in scientific questions was to the authority of "the Philosopher," Aristotle. Bacon's complaint was thus not so much against the ancients as against their disciples, particularly the Scholastic philosophers of the late Middle Ages (Robinson, 1995, p. 155). Bacon's *Novum Organum* (1620/1928a) proposed that this old method be replaced with a new organ or system based on the inductive study of nature itself. In short, what Bacon immodestly attempted was to "commence a total reconstruction of sciences, [practical] arts, and all human knowledge, raised upon the proper foundations" (Bacon, 1620/1928b, p. 4). The critical element in this foundation was the method of experimentation. Thus, a deliberate manipulation of variables was to replace the "noting and naming" kind of empiricism that had characterized the Aristotelian approach when it did lower itself to observation (Robinson, 1995, p. 158).

The character of Bacon's reconstruction, however, was to have positive and negative consequences for the conception of science that predominated for the next three centuries. The Baconian ideal for science was as follows: at the start of their research, experimenters are to remove

from their thinking all the " 'idols' or time-honored illusions and fallacies, born of [their] personal idiosyncrasies of judgment or the traditional beliefs and dogmas of [their] group" (Durant & Durant, 1961, p. 175). Thus, in the Baconian view, scientific observations are to be made in a purely objective fashion by individuals having no loyalties to any hypotheses or beliefs that would cause them to be blind to any portion of the empirical evidence. The correct conclusions and explanatory principles would then emerge from the evidence relatively automatically, and without the particular philosophical presuppositions of the experimenter playing any part. Thus, the "course of Nature" could be observed clearly if the experimenter would only look at Nature as it is. Nature, as it were, unambiguously dictated the adoption of true theories. The whole process of science, it was thought, could be purely objective, empirical, and rational.

Although this view of science is regarded as passé by some academics (cf. Gergen, 2001), particularly in the humanities, its flaws need to be noted because of its persistence in popular thought and even in the treatment of the scientific method in introductory texts in the sciences. Instead of personal judgment playing no role in science, it is critical to the whole process. Whether one considers the data collection, data analysis, or interpretation phases of a study, the process is not purely objective and rule governed. First, the scientist's preexisting ideas about what is interesting and relevant undeniably guide decisions about what data are to be collected. For example, if one is studying the effects of drug treatments on recovery of function following brain injury, one has decided in advance not just that recovery of function after brain injury is important to study but that the drugs present in the bloodstream may be a relevant factor, and one has likely also decided that the day of the week on which the drug treatment is administered is probably not a relevant factor. Data cannot be collected without some preexisting ideas about what may be relevant, because it is those decisions that determine the variables to be manipulated or assessed in a particular experimental design. There are no logical formulas telling the scientist which particular variables must be examined in a given study.

Similarly, the patterns observed in a set of data are influenced by the ideas the investigator brings to the research. To be sure, a great deal can be said about what methods of analysis are most appropriate to aid in this pattern-detection process for a particular experimental design. In fact, much of this book is devoted to appropriate ways of describing causal relationships observed in research. However, both experiments in cognitive psychology and examples from the history of science suggest that, to a large extent, what one sees is determined by what one expects to see (see Kuhn, 1970, esp. chap. 6). Although statistical analysis can objectify to some extent the process of looking for patterns in data, statistical methods, as Koch (1981) and others point out, even when correctly applied, do not assure that the most appropriate ways of organizing the data will be found. For example, in a simple four-group experimental design, there are, at least in theory, an infinite number of comparisons of the four group means that could be tested for statistical significance. Thus, even assuming that the most appropriate data had been collected, it is entirely possible that a researcher might fail to examine the most illuminating comparison. Admittedly, this problem of correctly perceiving at least approximately what the patterns in your data are is less serious than the problem of collecting the relevant data in the first place or the problem of what one makes of the pattern once it is discerned. Nonetheless, there are no absolutely foolproof strategies for analyzing data.

The final step in the inductive process is the most troublesome. Once data relevant to a question are collected and their basic pattern noted, how should the finding be explained? The causal explanations detailing the mechanisms or processes by which causes produce their effects are typically much harder to come by than facts to be explained (cf. Shadish et al., 2002, p. 9). Put bluntly, "there is no *rigorous logical* procedure which accounts for the birth of theories or of the novel concepts and connections which new theories often involve. There is no 'logic of discovery' " (Ratzsch, 2000, p. 19). As many a doctoral candidate knows from painful experience after

puzzling over a set of unanticipated results, data sometimes do not clearly suggest any theory, much less dictate the "correct" one.

Responses to the Criticisms of the Idea of Pure Science

Over the years, the pendulum has swung back and forth regarding the validity and implications of this critique of the allegedly pure objectivity, rationality, and empiricism of science. We consider various kinds of responses to these criticisms. First, it is virtually universally acknowledged that certain *assumptions* must be made to do science at all. Next, we consider three major alternatives that figured prominently in the shaping of *philosophy of science* in the 20th century. Although there were attempts to revise and maintain some form of the traditional view of science well into the 20th century, there is now wide agreement that the criticisms were more sound than the most influential revision of the traditional view. In the course of this discussion, we indicate our views on these various perspectives on philosophy of science and point out certain of the inherent limitations of science.

Assumptions

All rational argument must begin with certain assumptions, whether one is engaged in philosophical, scientific, or competitive debating. Although these assumptions are typically present only implicitly in the practice of scientific activities, there are some basic principles essential to science that are not subject to empirical testing but that must be presupposed for science to make sense. As Gauch (2003, chap. 4) has suggested, science's presuppositions are essentially that nature is "orderly and comprehensible." We will unpack these ideas by delineating two assumptions we consider to be most fundamental: the lawfulness of nature and finite causation (Underwood, 1957, pp. 3–6).

Lawfulness of Nature Although possibly itself a corollary of a more basic philosophical assumption, the assumption that the events of nature display a certain lawfulness is a presupposition clearly required by science. This is the belief that nature, despite its obvious complexity, is not entirely chaotic: regularities and principles in the outworking of natural events exist and wait to be discovered. Thus, on this assumption, an activity like science, which has as its goal the cataloging and understanding of such regularities, is conceivable.

There are a number of facets or corollaries to the principle of the lawfulness of nature that can be distinguished. First, at least since the ancient Greeks, there has been agreement on the assumption that *nature is understandable*, although not necessarily on the methods for how that understanding should be achieved. In our era, with the growing appreciation of the complexities and indeterminacies at the subatomic level, the belief that we can understand is recognized as not a trivial assumption. At the same time, the undeniable successes of science in prediction and control of natural events provide ample evidence of the fruitfulness of the assumption and, in some sense, are more impressive in light of current knowledge. As Einstein said, the most incomprehensible thing about the universe is that it is comprehensible[1] (Einstein, 1936, p. 351; see Koch, 1981, p. 265). The Hungarian Nobel laureate in physics, Eugene Wigner, agreed, writing "it is not at all natural that 'laws of nature' exist, much less that man is able to discover them," and further, the "regularities in the events in the world . . . can be formulated in terms of mathematical concepts with an uncanny accuracy" (Wigner, 1960, p. 5, 11).

A second facet of the general belief in the lawfulness of nature is that *nature is uniform*—that is, processes and patterns observed on only a limited scale hold universally. This is obviously required in sciences such as astronomy if statements are to be made on the basis of current

observations about the characteristics of a star thousands of years ago. However, the validity of the assumption is questionable, at least in certain areas of the behavioral sciences. Two dimensions of the problem can be distinguished. First, relationships observed in the psychology of 2020 may not be true of the psychology of 1970 or 2070. For example, the social psychology of attitudes in some sense must change as societal attitudes change. Rape, for instance, was regarded as a more serious crime than homicide in the 1920s but as a much less serious crime than homicide in the 1960s (Coombs, 1967). One possible way out of the apparent bind this places one in is to theorize at a more abstract level. Rather than attempting to predict attitudes about extending the privilege of voting to a rapist some time after a crime, one might instead theorize about the reinstatement of the right to vote of someone who had committed a crime of a specified level of perceived seriousness and allow which crime occupied that level to vary over time. Although one can offer such abstract theories, it is an empirical question as to whether the relationship will be constant over time when the particular crime occupying a given level of seriousness is changing.

A second dimension of the presupposition of the uniformity of nature that must be considered in the behavioral sciences pertains to the homogeneity of experimental material (e.g., individuals or families) being investigated. Although a chemist might safely assume that one hydrogen atom will behave essentially the same as another when placed in a given experimental situation, it is not at all clear that the persons studied by a psychologist can be expected to display the same sort of uniformity. Admittedly, there are areas of psychology—for example, the study of vision—in which there is sufficient uniformity across individuals in the underlying processes at work that the situation approaches that in the physical sciences. In fact, studies with very small numbers of subjects are common in the perception area. However, it is generally the case that individual differences among people are sufficiently pronounced that they must be reckoned with explicitly. This variability is, indeed, a large part of the need for those in psychology and related disciplines to be trained in the areas of experimental design and statistics, in which the focus is on methods for accommodating this sort of variability. We deal with the logic of this accommodation at numerous points, particularly in our discussion of randomization in this chapter and external validity in Chapter 2. In addition, Chapter 9 is devoted to methods for incorporating variables assessing individual differences among participants into one's research design and data analysis, and the succeeding chapters relate to methods designed to deal with the systematic variation among individuals.

A third facet of the assumption of the lawfulness of nature is the *principle of causality.* One definition of this principle, which was suggested by Underwood, is that "every natural event (phenomenon) is assumed to have a cause, and if that causal situation could be exactly reinstituted, the event would be duplicated" (1957, p. 4). At the time Underwood was writing, there was fair agreement regarding causality in science as a deterministic, mechanistic process. Since the 1950s, however, we have seen the emergence of a variety of views regarding what it means to say that one or more event(s) cause(s) another and, equally important, regarding how we can acquire knowledge about causal relationships. Fortunately, the field seems to have moved beyond the state of affairs of several decades ago when, as Cook and Campbell put it, "the epistemology of causation, and of the scientific method more generally, is at present in a productive state of near chaos" (1979, p. 10).

Cook and Campbell (1979, chap. 1) admirably characterized the evolution of thinking in the philosophy of science about causality, and Shadish et al. (2002, chap. 1) presented a helpful analysis of causal inference in different types of experiments. Pearl provides, in a delightfully illustrated lecture, what he terms a "gentle introduction" to the history of the idea of causation and a contemporary way of representing and analyzing causal relationships (2000, pp. xiv, 331–358). We can devote space here to only the briefest of summaries of this domain. Through most of its first 100 years as an experimental discipline, psychology was heavily influenced by

the view of causation offered by the Scottish empiricist philosopher David Hume (1711–1776). Hume argued that the inference of a causal relationship involving unobservables is never justified logically. Even in the case of one billiard ball striking another, one does not directly observe what caused the second ball to move because one does not know that its movement was a *necessary* result of the impact. Rather, one simply observes a correlation between the ball being struck and its moving. Thus, for Hume, correlation is all we can know about causality. These 18th-century ideas, filtered through the 19th-century positivism of Auguste Comte (1798–1857), pushed early 20th-century psychology toward an empiricist monism, a hesitancy to propose causal relationships between hypothetical constructs. Rather, the search was for functional relationships between observables or, only slightly less modestly, between theoretical terms, each of which was operationally defined by one particular measurement instrument or set of operations in a given study. Thus, in 1923, Boring would define *intelligence* as what a particular intelligence test measures. Science was to give us sure knowledge of relationships that had been confirmed rigorously by empirical observation.

These views of causality have been found to be lacking on a number of counts. First, although it is the case, as every elementary statistics text reiterates, that correlation does not necessarily imply causation, causal inferences based on properly designed experiments *are* warranted. The distinction between correlation and causation must be stressed again here, because in this book we describe relationships with statistical models that can be used for either correlational or causal relationships. This is potentially confusing, particularly because we follow the convention of referring to certain terms in the models as "effects." At some times, these effects are the magnitude of the change an independent variable causes in the dependent variable; at other times, the effect is better thought of as simply a measure of the strength of the correlational relationship between two measures. The strength of the support for the interpretation of a relationship as causal, then, hinges not on the statistical model used, but on the nature of the design used. In a correlational study, one of the variables may be treated as a dichotomous rather than a continuous variable, for example, as a result of classifying individuals as depressed or not based on their score on the Beck Depression Inventory. That one could carry out a t test[2] of the difference in anxiety between depressed and non-depressed groups, rather than computing a correlation between depression and anxiety, does *not* mean that you have a more secure basis for inferring causality than if you had simply computed the correlation. If the design of the study were such that depression was a measured trait of individuals rather than a variable independently manipulated by the experimenter, then that limits the strength of the inference rather than the kind of statistic computed.

Second, although the manipulated or measured variables involved in a causal relationship may reasonably be viewed as instantiations of theoretical constructs, using a single measurement device as definitional of one's construct entails a variety of difficulties, not least of which is that meters (or measures) sometimes are broken (invalid). We have more to say about such construct validity in Chapter 2. For now, we simply note that, in the social sciences, "one-variable, 'pure' measuring instruments are an impossibility. *All* measures involve many known theoretical variables, many as yet unknown ones, and many unproved presumptions" (Cook & Campbell, 1979, p. 14).

Finally, whereas early empiricist philosophers required causes and effects to occur in *constant conjunction*—that is, the cause was necessary and sufficient for the effect—current views are again more modest. At least in the behavioral sciences, the typical view is that all causal relationships are contingent or dependent on the context (cf. Shadish et al., 2002). The evidence supporting behavioral "laws" is thus probabilistic. In a randomized study, if 90 of 100 patients in a treatment group, as opposed to 20 of 100 in the control group, were to be cured according to some criterion, the reaction is to conclude that the treatment caused a very large effect, *instead* of

reasoning that, because the treatment was not sufficient for 10 subjects, it should not be regarded as the cause of the effect.

Most scientists, particularly those in the physical sciences, are generally realists; that is, they see themselves as pursuing theoretical truth about hidden but real mechanisms whose properties and relationships explain observable phenomena. Thus, the realist physicist would not merely say, as the positivist would, that a balloon shrinks as a function of time. Rather, he or she would want to proceed to make a causal explanation, for example, the leakage of gas molecules caused the observed shrinkage. This is an assertion that not just a causal relationship was constructed in the physicist's mind, but that a causal relationship really exists among entities outside of any human mind. Thus, in the realist view, theoretical assertions "have objective contents which can be either right or wrong" (Cook & Campbell, 1979, p. 29).

Science then is about uncovering and explaining causal relationships, the "cement of the universe" connecting causes with their effects (Mackie, 1980). The core of scientific argumentation about these relationships builds on presuppositions like the lawfulness of nature by employing both deductive and inductive logical principles (Gauch, 2003). Indeed, as Einstein asserted, "Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility of finding out causal relationships by systematic experiment" (Letter to J. S. Switzer, April 23, 1953, quoted in Newton, 1997, p. 9). Deduction is employed, for example, in deriving predictions from theory that can be tested in experiments. Induction is employed in drawing inferences from those experiments and in reasoning from data to an inferred model. Whereas Hume was notoriously skeptical that induction was ever justified, the statistical procedures on which this volume focuses illustrate the key role that statistics can play in such inferences, and in quantifying uncertainty, for example about the location of population parameters. Indeed, if induction, as the aphorism goes, is "the glory of science and the scandal of philosophy" (Gauch, 2003, p. 264), then the applied inductive logic of statistics contributes to science's glory.

Experiments are uniquely suited to determining the "effects of causes" (Dawid, 2000, 2002). Experiments thus allow *causal description* (Shadish et al., 2002, p. 9), that is, warranted conclusions about the presence and strength of causal relationships. Methods for arriving at such conclusions are not controversial and will be our primary focus. What is more debatable is the process by which one should arrive at *causal explanations*, that is, clarifying "the mechanisms through which and the conditions under which that causal relationship holds" (Shadish et al., 2002, p. 9). Much progress has been made in developing methodology for investigating hypothesized mediators, which are central to such causal explanations (MacKinnon, 2008; Pearl, 2014). However, research into mediators necessarily must deal with the difficult issues of inferring causality when purported causes cannot be directly manipulated as can the causes explored through true experiments. We will touch on some of the logical difficulties that arise and possible methods for analyzing designs with non-equivalent groups in Chapter 9.

A final issue regarding causation particularly relevant to the social sciences is whether to include human volition as a cause, at least in sciences studying people. For example, Collingwood (1940) suggested "that which is 'caused' is the free and deliberate act of a conscious and responsible agent, and 'causing' him to do it means affording him a motive for doing it" (p. 285). This is the kind of attribution for the cause of action presupposed throughout most of the history of Western civilization, but that came to represent only a minority viewpoint in 20th-century psychology, despite persisting as the prevailing view in other disciplines such as history and law. Nonetheless, several prominent researchers in modern psychology such as Albert Bandura (2001), Roy Baumeister (Baumeister, Bratslavsky, Muraven, & Tice, 1998), Joseph Rychlak (2000), and George Howard (Howard & Conway, 1986; Howard, Curtin, & Johnson, 1991) have

argued that research in experimental psychology can proceed from such an agentic or teleological framework as well.[3]

Thus, we see that a variety of views are possible about the kind of causal relationships that may be discovered through experimentation: the relationship may or may not be probabilistic, the relationship may or may not be regarded as referring to real entities, and the role of the participant may or may not be regarded as that of an active agent. This last point makes clear that the assumption of the lawfulness of nature does not commit one to a position of philosophical determinism as a personal philosophy of life (Eacker, 1972). Also, even though many regard choosing to do science as tantamount to adopting determinism as a working assumption in the laboratory, others do not see this as necessary even there. For example, Rychlak (2000) states that traditional research experiments provide a means of his putting his teleological theories of persons as free agents to the test. Similarly, George Howard and colleagues argue (Howard et al., 1991) that it is the individual's freedom of choice that results in the unexplained variation being so large in many experiments. Given that the algebraic models of dependent variables we use throughout this book incorporate both components reflecting unexplained variability and components reflecting effects of other variables, their use clearly does not require endorsement of a strictly deterministic perspective. Rather, the commitment required of the behavioral scientist, like that of the physicist studying subatomic particles, is to the idea that the consistencies in the data will be discernible through the cloud of random variation (see Meehl, 1970b).

It should perhaps be noted, before we leave the discussion of causality, that in any situation there are a variety of levels at which one could conduct a causal analysis. Both nature and science are stratified, and properties of entities at one level cannot, in general, be reduced to constellations of properties of entities at a lower level. For example, simple table salt (NaCl) possesses properties that are different from the properties of either sodium (Na) or chloride (Cl) (see Manicas & Secord, 1983). To cite another simple example, consider the question of what causes a room to suddenly become dark. One could focus on what causes the light in the room to stop glowing, giving an explanation at the level of physics by talking about what happens in terms of electric currents when the switch controlling the bulb is turned off. A detailed, or even an exhaustive, account of this event at the level of physics would not do away with the need for a psychological explanation of why a person flipped off the switch (see Cook & Campbell, 1979, p. 15). Psychologists are often quick to argue against the fallacy of reductionism when it is hinted that psychology might someday be reduced to physics or, more often, to biology. However, the same argument applies with equal force to the limitations of the causal relationships that behavioral scientists can hope to discover through empirical investigation. For example, a detailed, or even an exhaustive, psychological account of how someone came to hold a particular belief says nothing about the philosophical question of whether such a belief is true.

Having considered the assumption of the lawfulness of nature in some detail, we now consider a second fundamental assumption of science.

Finite Causation Science presupposes not only that there are natural causes of events, but also that these causes are finite in number and discoverable. Science is predicated on the belief that generality of some sort is possible; that is, it is not necessary to replicate the essentially infinite number of elements operating when an effect is observed initially in order to have a cause sufficient for producing the effect again. Now, it must be acknowledged that much of the difficulty in arriving at the correct interpretation of the meaning of an experimental finding is deciding *which* elements are critical to causing the phenomenon and *under what conditions* they are likely to be sufficient to produce the effect. This is the problem of causal explanation with which Chapter 2 is concerned (cf. Shadish et al., 2002).

A statistical analogy may be helpful in characterizing the principle of finite causation. A common challenge for beginning statistics students is mastering the notion of an interaction, whereby the effect of a factor depends or is contingent on the level of another factor that is present. When more than two factors are simultaneously manipulated (as in the designs we consider in Chapter 8), the notion extends to higher-order interactions, whereby the effect of a factor depends on combinations of levels of multiple other factors. Using this terminology, a statistician's way of expressing the principle of finite causation might be to say that "the highest-order interactions are not always significant." Because any scientific investigation must be carried out at a particular time and place, it is necessarily impossible to re-create exactly the state of affairs operating then and there. Rather, if science is to be possible, one must assume that the effect of a factor does not depend on the levels of all the other variables, measured or unmeasured, that are present when that effect is observed.

A corollary of the assumption of finite causation has a profound effect on how we carry out the model comparisons that are the focus of this book. This corollary is the bias toward simplicity. It is a preference we maintain consistently, in test after test, until the facts in a given situation overrule this bias.

Many scientists stress the importance of a strong belief in the ultimate simplicity of scientific laws. As Gardner points out, "this was especially true of Albert Einstein. 'Our experience,' he wrote, 'justifies us in believing that nature is the realization of the simplest conceivable mathematical ideas' " (Gardner, 1979, pp. 169–170; see Einstein, 1950, p. 64). However, as neuroscientists studying the brain know only too well, there is also an enormous complexity to living systems that at least obscures if not makes questionable the appropriateness of simple models. Indeed, the same may be true in some sense in all areas of science. Simple first approximations are, over time, qualified and elaborated: Newton's ideas and equations about gravity were modified by Einstein; Gall's phrenology was replaced by Flourens's views of both the unity and diversification of function of different portions of the brain.

Thus, we take as our guiding principle that set forward for the scientist by Alfred North Whitehead: "Seek simplicity and distrust it"; or again, Whitehead suggests that the goal of science "is to seek the simplest explanation of complex facts" while attempting to avoid the error of concluding nature is simpler than it really is (1920/1964, p. 163).

Admittedly, the principle of parsimony is easier to give lip service to than to apply. The question of how to measure the simplicity of a theory is by no means an easy one. Fortunately, within mathematics and statistics the problem is somewhat more tractable, particularly if you restrict your attention to models of a particular form. We adopt the strategy in this text of restricting our attention for the most part to various special cases of the general linear model. Although this statistical model can subsume a great variety of different types of analyses, it takes a fundamentally simple view of nature in that such models assume the effects of various causal factors[4] simply cumulate or are added together in determining a final outcome. In addition, the relative simplicity of two competing models in a given situation may easily be described by noting how many more terms are included in the more complex model. We begin developing these ideas in much greater practical detail in Chapter 3.

Modern Philosophy of Science

Having considered two fundamental assumptions of science, we continue our discussion of responses to the critique of the traditional view of science by considering four alternative philosophies of science. We begin by considering an attempt to revise and maintain the traditional view that has played a particularly important role in the history of psychology.

Positivism In our discussion of the principle of causality as an aspect of the assumption of the lawfulness of nature, we previously alluded to the influence of Humean empiricism and 19th-century positivism on 20th-century psychology. This influence was so dominant over the first 75 years of the 20th century that something more must be said about the principal tenets of the view of science that developed out of positivism and the opposing movements that in the latter part of the 20th century continued to grow in strength to the point of overtaking this view.

A positivistic philosophy of science was crystallized by the "Vienna Circle," a group of philosophers, scientists, and mathematicians in Vienna who, early in the 20th-century, set forth a view of science known as *logical positivism.* Rudolph Carnap and Herbert Feigl were two of the main figures in the movement, with Carl Hempel and A. J. Ayer also being among those whose writings heavily influenced psychology. Their logical positivism represented a wedding of Comte's positivism with the logicism of Whitehead and Russell's *Principia Mathematica.*

The aim of Auguste Comte's positive philosophy was to advance the study of society beyond a theological or metaphysical stage, in which explanations for phenomena were sought at the level of supernatural volition or abstract forces, to a "positive" stage. The stage was conceived to be positive in two distinct senses. First, all knowledge in the positive stage would be based on the positive (i.e., certain, sure) methods of the physical sciences. Rather than seeking a cause or an essence, one is content with a law or an empirical generalization. Second, Comte expected that the philosophical unity that would be effected by basing all knowledge on one method would result in a religion of humanity uniting all men and women (Morley, 1955).

The logical positivists combined this positivism with the logicism of Bertrand Russell's mathematical philosophy (Russell, 1919a). Logicism maintains that mathematics is logic. "All pure mathematics deals exclusively with concepts definable in terms of a very small number of fundamental concepts, and . . . all its propositions are deducible from a very small number of logical principles" (Russell, 1937, p. xv). Thus, all propositions in mathematics can be viewed as the result of applying truth functions to interpret various combinations of elementary or atomic propositions—that is, one determines the implications of the fundamental propositions according to a set of strictly logical rules. The meaning or content of the elementary propositions plays no role in the decision concerning whether a particular molecular proposition constructed out of elementary propositions by means of operators is true or false. Thus, like logic, mathematics fundamentally "is concerned solely with syntax, i.e., with formal relations between symbols in accordance with precise rules" (Brown, 1977, p. 21).

The modern logical positivism, which played such a dominant role in the way academic psychologists thought about their field, is a form of positivism that takes such symbolic logic as its primary analytic tool. This is seen in the central doctrine of logical positivism, known as the *Verifiability Criterion of Meaning.* According to this criterion, a proposition is meaningful "if and only if it can be empirically verified, i.e., if and only if there is an empirical method for deciding if it is true or false" (Brown, 1977, p. 21). (The only exception to this rule is the allowance for analytical propositions, which are propositions that assert semantic identities or that are true just by virtue of the terms involved, for example, "All bachelors are unmarried.") Thus, scientific terms that could not be defined strictly and completely in terms of sensory observations were regarded as literally meaningless. Any meaningful statement must reduce then to elementary propositions that can literally be seen to be true or false in direct observation. The bias against statistical tests and in favor of black-or-white, present-or-absent judgment of relationships in data was only one practical outworking of this philosophical view.

The goal of the logical positivists was then to subsume the rationale and practice of science under logic. The central difficulty preventing this was that scientific laws are typically stated as universal propositions that cannot be verified conclusively by any number of observations. One cannot show, for example, that all infants babble simply by observing some critical number

of babbling babies. In addition, there are a number of paradoxes of confirmation about which no consensus was ever achieved as to how they should be resolved (Brown, 1977, Chapter 2). Hempel's "paradox of the ravens" illustrates the most famous of these (1945). As Wesley Salmon succinctly summarized in *Scientific American*:

> If all ravens are black, surely non-black things must be non-ravens. The generalizations are logically equivalent, so that any evidence that confirms one must tend to confirm the other. Hence the observation of a green vase seems to confirm the hypothesis that all ravens are black. Even a black raven finds it strange.
>
> (1973, p. 75)

Such paradoxes were especially troublesome to a philosophical school of thought that had taken the purely formal analysis of science as its task, attempting to emulate Whitehead and Russell's elegant symbolic logic approach that had worked so well in mathematics.

Although the dilemmas raised because the contrapositive of an assertion is logically equivalent to the original assertion [e.g., using an arrow between terms A and B to indicate "A implies B" or "if A, then B" and using a double-headed arrow to indicate "A if and only if B," the logical equivalence illustrated by the first sentence of Salmon's quote could be written (raven → black) ↔ (non-black → non-raven)] may not seem relevant to how actual scientific theories come to be accepted, this is typical of the logical positivist approach. Having adopted symbolic logic as the primary tool for the analysis of science, then proposition forms and their manipulation became the major topic of discussion. The complete lack of detailed analysis of major scientific theories or research efforts is thus understandable, but unfortunate. When psychologists adopted a positivistic approach as the model of rigorous research in the physical sciences, they were, in fact, adopting a method that bore virtually no relationship to the way physicists actually approached research.

The most serious failing of logical positivism, however, was the failure of its fundamental principle of the Verifiability Criterion of Meaning. A number of difficulties are inherent in this principle (Ratzsch, 2000, p. 31ff.), but the most critical problems include the following: first, as we have seen in our discussion of the assumptions of science, some of the basic principles needed for science to make sense are not empirically testable. One cannot prove that events have natural causes, but without such assumptions, scientific research is pointless.

Second, attempts such as operationism to adhere to the criterion resulted in major difficulties. The operationist thesis, so compatible with behaviorist approaches, was originally proposed by P.W. Bridgman: "In general, we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations" (1927, p. 5). However, this was taken to mean that if someone's height, much less their intelligence, were to be measured by two different sets of operations, these are not two different ways of measuring height, but are definitional of different concepts, which should be denoted by different terms (see the articles in the 1945 Symposium on Operationism published in *Psychological Review,* especially Bridgman, 1945, p. 247). Obviously, rather than achieving the goal of parsimony, such an approach to meaning results in a proliferation of theoretical concepts and, in some sense, "surrender of the goal of systematizing large bodies of experience by means of a few fundamental concepts" (Brown, 1977, p. 40). Finally, the Verifiability Criterion of Meaning undercuts itself. The criterion itself is neither empirically testable nor obviously analytic. Thus, either it is itself meaningless, or meaningfulness does not depend on being empirically testable—that is, it is either meaningless or false.

Thus, positivism failed in its attempts to subsume science under formal logic, did not allow the presuppositions necessary for doing science, prevented the use of generally applicable theoretical terms, and was based on a criterion of meaning that was ultimately incoherent. Unfortunately, its influence on psychology long outlived its relatively brief prominence within philosophy itself.

Popper An alternative perspective that we believe holds considerably more promise for appropriately conceptualizing science is provided by Karl Popper's falsificationism (1968) and subsequent revisions thereof (Lakatos, 1978; Newton-Smith, 1981). These ideas have received increasing attention in the literature on methodology for the behavioral sciences (see Cook & Campbell, 1979, p. 20ff.; Dar, 1987; Gholson & Barker, 1985; Rosenthal & Rosnow, 1991, p. 32ff.; Serlin & Lapsley, 1985; Shadish et al., 2002, p. 15ff.; see also Klayman & Ha, 1987). Popper's central thesis is that deductive knowledge is logically possible. In contrast to the "confirmationist" approach of the logical positivists, Popperians believe progress occurs by falsifying theories. Although this may seem counterintuitive, it rests on the logic of the compelling nature of deductive as opposed to inductive arguments.

What might *seem* more plausible is to build up support for a theory by observing that the predictions of the theory are confirmed. The logic of the seemingly more plausible confirmationist approach may be expressed in the following syllogism:

Syllogism of Confirmation

If theory T is true, then the data will follow the predicted pattern P.
The data follow predicted pattern P.
Therefore, theory T is true.

This should be regarded as an invalid argument but perhaps not as a useless argument. The error of thinking that data prove a theory is an example of the logical fallacy known as "affirming the consequent." The first assertion in the syllogism states that T is sufficient for P. Although such if-then statements are frequently misunderstood to mean that T is necessary for P (see Dawes, 1975), that does not follow. This is illustrated in the Venn diagram in Figure 1.1(a). As with any Venn diagram, it is necessary to view the terms of interest (in this case, theory T and data

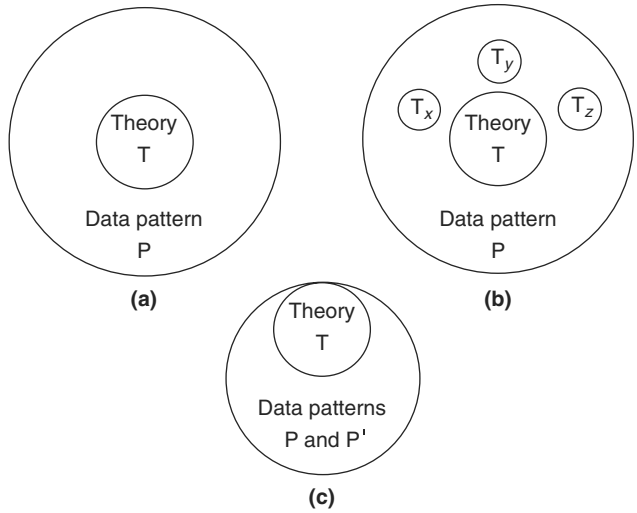


FIG. 1.1 Venn diagrams illustrating that theory T is sufficient for determining data pattern P [see (a)], but that data pattern P is not sufficient for concluding theory T is correct [see (b)]. The Venn diagram in (c), which illustrates that a smaller set of theories would be able to account for both data patterns P and P′, is discussed later in this section of the text.

pattern P) as sets, which are represented in the current diagram as circles. This allows one to visualize the critical difference between a theory being a sufficient explanation for a data pattern and its being necessarily correct. That theory T is sufficient for pattern P is represented by T being a subset of P. However, in principle at least, there are a number of other theories that also could explain the data, as illustrated by the presence of theories T_x, T_y, and T_z in Figure 1.1(b). Just being "in" pattern P does not imply that a point will be "in" theory T, that is, theory T is not necessarily true. In fact, the history of science provides ample support for what has been termed the *pessimistic induction*: "Any theory will be discovered to be false within, say 200 years of being propounded" (Newton-Smith, 1981, p. 14).

Popper's point, however, is that under certain assumptions, *rejection* of a theory, as opposed to confirmation, may be done in a deductively rigorous manner. The syllogism now is:

Syllogism of Falsification

If theory T is true, then the data will follow the predicted pattern P.
The data do not follow predicted pattern P.
Therefore, theory T is false.

The logical point is that although the converse of an assertion is *not* equivalent to the assertion, the contrapositive, as we saw in the paradox of the ravens, *is*. That is, in symbols $(T \rightarrow P) \not\leftrightarrow (P \rightarrow T)$, but $(T \rightarrow P) \leftrightarrow (\text{not } P \rightarrow \text{not } T)$. In terms of Figure 1.1, if a point is in P, that does not mean it is in T, but if it is outside P, it is certainly outside T. Thus, although one cannot prove theories correct, one can, by this logic, prove them false. For example, a theory saying that light will always follow a straight line could be proven false by a single observation, such as Eddington's during a 1919 eclipse, showing that light from a distant star bent when it went past the sun.

Although it is hoped that this example makes the validity of the syllogism of falsification clear, it is important to discuss some of the assumptions implicit in the argument and raise briefly some of the concerns voiced by critics of Popper's philosophy, particularly as it applies to the behavioral sciences. First, consider the first line of the falsification syllogism. The one assumption pertinent to this, about which there is agreement, is that it is possible to derive predictions from theories. Confirmationists assume this as well. Naturally, theories differ in how well they achieve the desiderata of good theories regarding predictions—that is, they differ in how easily empirical predictions may be derived and in the range and specificity of these predictions. Unfortunately, psychological theories, particularly in recent years, tend to be very restricted in scope. Also, unlike physics, the predictions that psychological theories do make are typically of a non-specific form ("the groups will differ") rather than being point predictions ("the light rays will bend by *x* degrees as they go past the sun") (see Meehl, 1967, 1986). However, whether specific or non-specific, as long as it is assumed that a rather confident judgment can be made—for example, by a statistical test—about whether the results of an experiment are in accord with the predictions, the thrust of the argument maintains its force.[5]

More troublesome than the lack of specificity or generality of the predictions of psychological theories is that the predictions depend not only on the core ideas of the theory, but also on a set of additional hypotheses. These often have to do with the particular way in which the theoretical constructs of interest are implemented in a given study and may actually be more suspect than the theory itself (cf. Smedslund, 1988). As expressed in the terminology of the brilliant psychologist and philosopher of science Paul Meehl, "[I]n social science the auxiliaries A and the initial and boundary conditions of the system C are frequently as problematic as the theory T itself" (1978, p. 819). For example, suppose a community or health psychologist wants to investigate the effect of perceived risk and response efficacy on self-protection. Funding is obtained to investigate

the effectiveness of such a theoretically driven intervention in decreasing the use of alcohol and illegal drugs as the criterion behavior in a study of at-risk youth, some of whom are randomly assigned to receive an experimental treatment. In her study, the psychologist attempts to impress upon groups of middle school youth from local economically disadvantaged areas the dangers of drug use by taking them to hospitals or detention centers to talk with young adults who have been injured or arrested as a result of their use of alcohol or illegal drugs. She also attempts to increase the middle schoolers' belief in their ability to avoid alcohol or drug use by having them participate in discussion groups on the subject led by undergraduate research assistants. A negative result (or worse yet, increased drug use in the treated group) causes one to question if the core substantive theory (T) of the impact of risk perception and response efficacy on self-protection has been falsified or if one or more of the auxiliary hypotheses (A) have been falsified. For example, perhaps the visits with the hospitalized or jailed youths served to tacitly validate them as role models to be emulated rather than increasing the students' perceived risk of drug use, or perhaps the fact that a large majority of the undergraduate assistants leading the discussions were themselves binge drinkers or users of illegal drugs did not facilitate their ability to persuade the middle schoolers of how easily and efficaciously they could make responses to avoid such risky behaviors. Or perhaps even the presumed boundary condition (C) that the motivation to avoid danger in the form of health or legal consequences was present at a high level, particularly in comparison to other motivations such as peer approval, was not satisfied in this population. We consider such difficulties further when we discuss construct validity in Chapter 2.

Turning now to the second line of the falsification syllogism, much also could be said about caveats. For one thing, some philosophers of science, including Popper, have philosophical reservations about whether one can know with certainty that a predicted pattern has not been obtained because that knowledge is to be obtained through the fallible inductive method of empirical observation (see Newton-Smith, 1981, chap. 3). More to the point for our purposes is the manner in which empirical data are to be classified as conforming to one pattern or another. Assuming one's theory predicts that the pattern of the data will be that people in general will perform differently in the treatment and control conditions, how does one decide on the basis of a sample of data what is true of the population? That, of course, is the task of inferential statistics and is the sort of question to which the bulk of this book is addressed. First, we show in the latter part of this chapter how one may derive probability statements rigorously for very simple situations under the assumption that there is no treatment effect. If the probability is sufficiently small (such as less than .05), the hypothesis of no difference is rejected. If the probability fails to reach a conventional level of significance, one might be tempted to conclude that the alternative hypothesis is false or, equivalently, that the null hypothesis is true. (More on this in a moment.) Second, we show beginning in Chapter 3 how to formulate such questions for more complicated experiments using standard parametric (e.g., t or F) tests. In sum, because total conformity with the exact null hypotheses of the social and behavioral sciences (or, for that matter, with the exact point predictions sometimes used—e.g., in some areas of physics) is never achieved, inferential statistics serves the function of helping scientists classify data patterns as being confirmed predictions, falsified predictions, or in some cases, ambiguous outcomes.

A final disclaimer is that Popper acknowledges that, in actual scientific practice, singular discordant facts alone rarely do or should falsify theories. Hence, in practice, as hinted at previously, a failure to obtain a predicted data pattern may not *really* lead to a rejection or abandonment of the alternative hypothesis the investigator wanted to support. In all too many behavioral science studies, the lack of statistical power is a quite plausible explanation for failure to obtain predicted results.[6] (What is more, as Maxwell, Lau, & Howard (2015) explain, having a sample size that provides adequate power (say, a 90% chance) of detecting the hypothesized effect size may by no means be adequate for concluding the true effect is essentially zero. This is one reason for

the maxim that a test yielding a statistically non-significant result means that one *fails to reject* the null rather than that one *accepts* the null hypothesis as true.) Also, such statistical reasons for failure to obtain predicted results are only the beginning. Because of the existence of the other explanations we have considered (e.g., "Some auxiliary theory is wrong") that are typically less painful to a theorist than rejection of the principal theory, in practice a combination of multiple discordant facts *and* a more viable alternative theory are usually required for the refutation of a theoretical conjecture (see Cook & Campbell, 1979, p. 22ff.).

We pause here to underscore some of the limitations of science that have emerged from our consideration of Popper and then highlight some of the general utility of his ideas. Regarding science's limitations, we have seen that not only is there no possibility of proving any scientific theory with logical certainty, but also that there is no possibility of falsifying one with logical certainty. That there are no proven theories is a well-known consequence of the limits of inductive logic. Such difficulties are also inherent to some extent in even the simplest empirical generalization (the generalization is not logically compelled, for reasons including the fact that you cannot be certain what the data pattern is because of limited data and potential future counterexamples to the current pattern and that any application of the generalization requires reliance on principles like uniformity). In short, "the data do not drive us inevitably to correct theories, and even if they did or even if we hit on the correct theory in some other way, we could not prove its correctness conclusively" (Ratzsch, 2000, pp. 76–77). Furthermore, theories cannot be proved false because of the possibility of explaining away purported refutations via challenges based on the fallibility of statistical evidence or of the auxiliary hypotheses relied on in testing the theory. In addition, there is the practical concern that despite the existence of discordant facts, the theory may be the best available.

On the positive side of the ledger, Popper's ideas have much to offer, both practically and philosophically. Working within the limitations of science, the practical problem for the scientist is how to eliminate explanations other than the theory of interest. We can see the utility of the Popperian conceptual framework in Figure 1.1. The careful experimenter proceeds, in essence, by trying to make the outer circle (i.e., the predicted data pattern) as small (i.e., as constrained or restrictive) as possible in order to refute the rival theories. We mentioned previously that the syllogism of confirmation, although invalid, was not useless. The way in which rival hypotheses are eliminated is essentially by confirming the predictions of one's theory in more situations, in at least some of which the rival hypotheses make contrary predictions. Figure 1.1(c) illustrates this. The outer circle now represents the intersection or joint occurrence of obtaining the predicted data P and also predicted data P′. For example, if a positive result had been obtained in the self-protection study with middle schoolers, the interpretation that increased perception of risk was the causal variable could be strengthened by including control conditions in which plausible other causes besides increased perception of risk were operating. One possible rival hypothesis (which might be represented by T_x in Figure 1.1) could be that the increased monitoring of the middle schoolers involved in the study might itself serve to suppress drug use regardless of the treatment received. Having a control group that was assessed as often and in as much detail as the treatment group but that did not manifest the decreased use seen in the treatment group essentially eliminates that rival explanation. The plausibility of the causal explanation would be enhanced further by implementing the construct in different ways, such as attempting to increase the perceived risk of smoking or sun exposure as a means of trying to induce other self-protective behaviors in other populations.

Indeed, part of the art of experimental design has to do with devising control conditions for which the theory of interest would make a different prediction than would a plausible rival hypothesis. (As another example, consider a study of recovery of function following ablation of a brain region where the rival hypothesis, "The deficit is a result of simply the operation, not

the brain area destroyed," is discounted by showing no deficit in a sham surgery condition.) If the rival hypothesis is false, part of the credo of science is that with sufficient investigation, ultimately, it will be discovered. As Kepler wrote regarding rivals to the Copernican hypothesis that made some correct predictions,

> And just as in the proverb liars are cautioned to remember what they have said, so here false hypotheses which together produce the truth by chance, do not, in the course of a demonstration in which they have been applied to many different matters, retain this habit of yielding the truth, but betray themselves.
>
> (Kepler, 1601)

Although in principle an infinite number of alternative hypotheses always remain, it is of little concern if no *plausible* hypotheses can be specified. We return to this discussion of how rival hypotheses can be eliminated in the discussion of validity in Chapter 2.

Regarding other, more philosophical considerations, for Popper the aim of science is truth. However, given that he concurs with Hume's critique of induction, Popper cannot claim to know the truth of a scientific hypothesis. Thus, the reachable goal for science in the real world is to be that of a closer approximation to the truth, or in Popper's terms, a higher degree of *verisimilitude.* The method of achieving this is basically a rational one by way of the logically valid refutation of alternative conjectures about the explanation of a given phenomenon. Although the details of the definition of the goal of verisimilitude and the logic of the method are still evolving (see Meehl, 1978; Newton-Smith, 1981; Popper, 1976), we find ourselves in basic agreement with a neo-Popperian perspective, both in terms of ontology and of epistemology. However, we postpone further discussion of this until we have briefly acknowledged some of the other major positions in contemporary philosophy of science.

Kuhn Thomas Kuhn, perhaps the best-known contemporary philosopher of science, is perceived by some as maintaining a position in *The Structure of Scientific Revolutions* (1970) that places him philosophically at the opposite pole from Karl Popper. Whereas Popper insists that science is to be understood logically, Kuhn maintains that science should be interpreted psychologically (Robinson, 1981, p. 24) or sociologically. Once a doctoral student in theoretical physics, Kuhn left the field to carry out work in the history and philosophy of science. Spending 1958–1959 at the Center for Advanced Studies in the Behavioral Sciences helped crystallize his views. Whereas his major work is based on the history of the physical sciences, his rationale draws on empirical findings in behavioral science, and others (e.g., Gholson & Barker, 1985; see also Gutting, 1980) apply Kuhn's views to psychology in particular. Kuhn's *Structure* was one of the most cited works in academic journals in the second half of the 20th century (e.g., Garfield, 1987) and has had an enduring impact on psychology (Driver-Linn, 2003).

Kuhn's basic idea is that psychological and sociological factors are the real determinants of change in allegiance to a theory of the world, and in some sense actually help determine the characteristics of the physical world that is being modeled. The notion is quasi-Kantian in that characteristics of the human mind, or at least of the minds of individual scientists, determine in part what is observed.

Once we have described four of Kuhn's key ideas—paradigms, normal science, anomalies, and scientific revolutions—we point out two criticisms commonly made of his philosophy of science.

For Kuhn, *paradigms* are "universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners" (Kuhn, 1970, p. viii). Examples include Newton's *Principia* and Lavoisier's *Chemistry*, "works that served for a time

implicitly to define the legitimate problems and methods of a research field" (1970, p. 10). The period devoted to solving the unresolved puzzles within an area following publication of such landmark works as these is what constitutes *normal science.* Inevitably, such periods of normal science turn up *anomalies,* or data that do not fit perfectly within the paradigm (1970, chap. 6). Although such anomalies may emerge slowly because of the difficulties in perceiving them shared by investigators working within the *Weltanschauung* or "worldview" of a given paradigm (1970, chap. 10), eventually a sufficient number of anomalies are documented to bring the scientific community to a crisis state (1970, chap. 7). The resolution of the crisis eventually may require a shift to a new paradigm. If so, the transition to the new paradigm is a cataclysmic event. Although some may view the new paradigm as simply subsuming the old, according to Kuhn, the transition—for example, from "geocentrism to heliocentrism, from phlogiston to oxygen, or from corpuscles to waves . . . from Newtonian to Einsteinian mechanics"—necessitated a "revolutionary reorientation," a conceptual transformation that is "decisively destructive of a previously established paradigm" (1970, p. 102).

Although his contributions have been immensely useful in stressing the historical development of science and certain of the psychological determinants of the behavior of scientists, there are, from our perspective, two major related difficulties with Kuhn's philosophy. Kuhn, it should be noted, has attempted to rebut such criticisms [see especially points 5 and 6 in the postscript added to *The Structure of Scientific Revolutions* (1970, pp. 198–207)]; however, in our view, he has not done so successfully. First, paradigm shifts in Kuhn's system do not occur because of the objective superiority of one paradigm over the other. In fact, such cannot be demonstrated, because for Kuhn, paradigms are incommensurable. Thus, attempts for proponents of different paradigms to talk to each other result in communication breakdowns (Kuhn, 1970, p. 201). Although this view is perhaps not quite consensus formation via mob psychology, as Lakatos (1978) characterizes it, it certainly implies that scientific change is not rational (see Manicas & Secord, 1983; Suppe, 1977). We are too committed to the real effects of psychological variables to be so rash as to assume that all scientific change is rational with regard to the goals of science. In fact, we readily acknowledge not only the role of psychological factors, but also the presence of a considerable amount of fraud in science[7] (see Broad & Wade, 1982; Stroebe, Postmes, & Spears, 2012). However, we believe that these are best understood as deviations from a basically rational model (see Newton-Smith, 1981, pp. 5–13, 148ff.).

Second, we share with others concerns regarding what appears to be Kuhn's relativism. The reading of his work by a number of critics is that Kuhn maintains that there is no fixed reality of nature for science to attempt to more accurately describe. For example, he writes:

> [W]e may . . . have to relinquish the notion, explicit or implicit, that changes of paradigm carry scientists and those who learn from them closer and closer to the truth. . . . The developmental process described in this essay has been a process of evolution *from* primitive beginnings—a process whose successive stages are characterized by an increasingly detailed and refined understanding of nature. But nothing that has been or will be said makes it a process of evolution *toward* anything.
>
> (Kuhn, 1970, pp. 170–171)

Kuhn elaborates on this in his postscript:

> One often hears that successive theories grow ever closer to, or approximate more and more closely to, the truth. Apparently generalizations like that refer not to the puzzle-solutions and the concrete predictions derived from a theory but rather to its ontology, to the match, that is, between the entities with which the theory populates nature and what is "really there."

> Perhaps there is some other way of salvaging the notion of "truth" for application to whole theories, but this one will not do. There is, I think, no theory-independent way to reconstruct phrases like "really there"; the notion of a match between the ontology of a theory and its "real" counterpart in nature now seems to me illusive in principle.
>
> (Kuhn, 1970, p. 206)

Although Kuhn in later publications claims he attempts to avoid the dangers of relativism and the "excesses of postmodernist movements" (2000, p. 91) by the rather vague, wistful suggestion that communities can agree to play a "language game" whose rules, for example, of noncontradiction, would constrain what might be asserted, he is much more clear in his assertion that "anything at all like a correspondence theory of truth" must be rejected (2000, p. 99). However, even if it is the case, as the pessimistic induction suggests, that all theories constructed in this world are false, it seems clear that some are less false than others. Does it not make sense to say that the earth revolves around the sun corresponds more closely to the truth of how things really are than to assert that the sun revolves around the earth or that the moon is made of blue cheese? Is it not reasonable to believe that the population mean score on the Wechsler Adult Intelligence Scale is really closer to 100 than it is to 70 or 130? In Kuhn's system, there is no paradigm-independent standard to allow such judgments (cf. Kuhn, 2000, p. 15). We concur with Newton-Smith (1981, pp. 34–37, 102–124) and Popper (1972) that this relativism about the nature of the world is unreasonable. In recent years, it has been the postmodernists who have advanced arguments against an objectively knowable world and against a view of science as attempting to use language, including numerical language, to make true statements about the world (Gergen, 2001). Yet the very advancing of an argument for the truth of the position that there is no truth undercuts itself. One is reminded of Socrates's refutation of the self-stultifying nature of the Sophists' skepticism (cf. Robinson, 1995, p. 26); in effect, if you claim each person is the measure of all things and that no one has any superior right to determine whether any assertion is true or false, why should I accept your position as authoritative?

Although the relativistic position of the postmodernists has certainly attracted numerous followers since the early 1980s, particularly in the humanities, for the most part the sciences, including academic psychology, continue to reject such views (see Haig, 2002; Hofmann, 2002) in favor of the realist perspective we now consider.

Realism Although there is a multitude of different realist positions in the philosophy of science, certain core elements of realism can be identified (Alston, 1996, p. 7ff.; Fine, 1987, p. 359ff.). First, realism holds that a definite world exists, a world populated by entities with particular properties, powers, and relations, and "the way the world is" is largely independent of the observer (Harré & Madden, 1975). Second, realist positions maintain that it is possible to obtain a substantial amount of accurate, relatively observer-independent information about the world (Rosenthal & Rosnow, 1991, p. 9), including information about structures and relations among entities as well as what may be observed more superficially. Third, the aim of science is to achieve such knowledge. Fourth, as touched on in our earlier discussion of causality, realist positions maintain that scientific propositions are true or false by virtue of their correspondence or lack of correspondence with the way the world is, independently of ourselves (Newton-Smith, 1981, pp. 28–29). Finally, realist positions tend to be optimistic in their view of science by claiming that the historically generated sequence of theories of a mature science reflect an improvement in terms of the degree of approximation to the truth (Newton-Smith, 1981, p. 39).

These tenets of realism can be more clearly understood by contrasting these positions with alternative views. Although there have been philosophers in previous centuries (e.g., Berkeley, 1685–1753) and in more modern times (e.g., Russell, 1950) who question whether the belief in

the existence of the physical world was logically justified, not surprisingly, most find arguments for the existence of the world compelling (Russell's argument and rebuttals thereof are helpfully juxtaposed by Oller, 1989). As Einstein tells it, the questioning of the existence of the world is the sort of logical bind one gets oneself into by following Humean skepticism to its logical conclusion (Einstein, 1944, pp. 279–291). Hume correctly saw that our inferences about causal connections, for example, are not logically necessitated by our empirical experience. However, Russell and others extended this skepticism to any knowledge or perception we might have of the physical world. Russell's point is that, assuming causality exists (even though we cannot know it does), our perception represents the end of a causal chain. Trying to reconstruct what "outside" caused that perception is a hazardous process. Even seeing an object such as a tree, if physics is correct, is a complicated and indirect affair. The light reaching the eye comes ultimately from another source such as the sun, not the tree, yet you do not say you are seeing the sun. Thus, Russell concludes that

> from what we have been saying it is clear that the relation of a percept to the physical object which is supposed to be perceived is vague, approximate and somewhat indefinite. There is no *precise* sense in which we can be said to perceive physical objects.
>
> (Russell, 1950, p. 206)

And, not only do we not know the true character of the tree we think we are seeing, but also "the colored surfaces which we see cease to exist when we shut our eyes" (Russell, 1914, p. 64). Here, in effect, Russell throws the baby out with the bathwater. The flaw in Russell's argument was forcefully pointed out by Dewey (1916). Dewey's compelling line of reasoning is that Russell's questioning is based on the analysis of perception as the end of a causal chain; however, this presupposes that there is an external object that is initiating the chain, regardless of how poorly its nature may be perceived.

Moving to a consideration of the other tenets of realism, the emphasis on accurate information about the world and the view that scientific theories come, over time, to more closely approximate a true description of the world clearly contrasts with relativistic accounts of science that see it as not moving toward anything. In fact, one early realist, C. S. Peirce, developed an influential view of truth and reality that hinges on there being a goal toward which scientific investigations of a question must tend (see Oller, 1989, p. 53ff.). Peirce wrote:

> The question therefore is, how is true belief (or belief in the real) distinguished from false belief (or belief in fiction). . . . The ideas of truth and falsehood, in their full development, appertain exclusively to the scientific method of settling opinion. . . . All followers of science are fully persuaded that the processes of investigation, if only pushed far enough, will give one certain solution to every question to which it can be applied. . . . The opinion which is fated to be ultimately agreed to by all who investigate, is what we mean by the truth and the object represented in this opinion is the real. . . . Our perversity and that of others may indefinitely postpone the settlement of opinion; it might even conceivably cause an arbitrary proposition to be universally accepted as long as the human race should last. Yet even that would not change the nature of the belief, which alone could be the result of investigation, that true opinion must be the one which they would ultimately come to.
>
> (Peirce, 1878, pp. 298–300)

Thus, in Peirce's view, for any particular scientific question that has clear meaning, there was one certain solution that would be obtained if only scientific investigation could be carried far

enough. This view of science is essentially the same as Einstein's, who likened the process of formulating a scientific theory to the task facing

> a man engaged in solving a well designed word puzzle. He may, it is true, propose any word as the solution; but, there is only one word which really solves the puzzle in all its forms. It is an outcome of faith that nature—as she is perceptible to our five senses—takes the character of such a well formulated puzzle.
>
> (Einstein, 1950, p. 64)

Scientific realism may also be contrasted with instrumentalist views. Instrumentalists argue that scientific theories are not intended to be literally true, but are simply convenient summaries or calculational rules for deriving predictions. This distinction is illustrated particularly well by the preface that Osiander added to Copernicus's *The Revolutions of the Heavenly Spheres*:

> It is the duty of the astronomer to compose the history of the celestial motions through careful and skillful observation. Then turning to the causes of these motions or hypotheses about them, he must conceive and devise, since he cannot in any way attain to the true causes, such hypotheses as, being assumed, enable the motions to be calculated correctly from the principles of geometry, for the future as well as the past. The present author [Copernicus] has performed both these duties excellently. For these hypotheses need not be true nor even probable; if they provide a calculus consistent with the observations that alone is sufficient.
>
> (Rosen, 1959, pp. 24–25)

Osiander recognized the distinction between factual description and a convenient formula for making predictions and is suggesting that whether the theory describes reality correctly is irrelevant. That is the instrumentalist point of view. However, many scientists, particularly in the physical sciences, tend to regard their theories as descriptions of real entities. This was the case for Copernicus and Kepler regarding the heliocentric theory and more recently for Bohr and Thomson regarding the electron. Besides the inherent plausibility of the realist viewpoint, the greater *explanatory power* of the realist perspective is a major argument offered in support of realism. Such explanatory power is perhaps most impressive when reference to a single set of entities allows predictions across different domains or allows predictions of phenomena that have never been observed but that, subsequently, are confirmed.

Some additional comments must be made about realism at this point, particularly as it relates to the behavioral sciences. First, scientific realism is not something that is an all-or-nothing matter. One might be a realist with regard to certain scientific theories and not with regard to others. Indeed, some have attempted to specify the criteria by which theories should be judged, or at least have been judged historically, as deserving a realistic interpretation (Gardner, 1987; Gingerich, 1973). Within psychology, a realistic interpretation might be given to a brain mechanism that you hypothesize is damaged on the basis of the poor memory performance of a brain-injured patient. However, the states in a mathematical model of memory, such as working memory, may be viewed instrumentally, as simply convenient fictions or metaphors that allow estimation of the probability of recall of a particular item.

A second comment is that realists tend to be emergentists and stress the existence of various levels of reality. Nature is viewed as stratified, with the higher levels possessing new entities with

powers and properties that cannot be explained adequately by the lower levels (Bhaskar, 1982, esp. secs. 2.5 and 3.3).

> From the point of view of emergence, we cannot reduce personality and mind to biological processes or reduce life to physical and chemical processes without loss or damage to the unity and special qualities of the entity with which we began.
>
> (Titus, 1964, p. 250)

Thus, psychology from the realist perspective is not in danger of losing its field of study to ardent sociobiologists any more than biologists would lose their object of inquiry if organic life could be produced by certain physical and chemical manipulations in the laboratory. Neither people nor other living things would cease to be real, no matter what the scientific development. Elements of lower orders are just as real, no more or less, than the comprehensive entities formed out of them. Both charged particles and thunderstorms, single cells and single adults exist and have powers and relations with other entities at their appropriate levels of analysis.

Because of the many varieties of realism—for example, critical realism (Cook & Campbell, 1979), metaphysical realism (Popper, 1972), and transcendental realism (Bhaskar, 1975)—and because our concern regarding philosophy of science is less with ontology than with epistemological method, we do not attempt to summarize the realist approach further. The interested reader is referred to the article by Manicas and Secord (1983) for a useful summary and references to the literature.

Conclusion Regarding Philosophy of Science Our own perspective is to hold to a realist position ontologically and a temperate rationalist position epistemologically of the neo-Popperian variety. The perspective is realist because it assumes phenomena and processes exist outside of our experience and that theories can be true or false, and among false theories, false to a greater or lesser extent, depending on the degree of correspondence between the theory and the reality. Naturally, however, our knowledge of this reality is limited by the nature of induction—thus, it behooves us to be critical of the strength of our inferences about the nature of that reality (see Cook & Campbell, 1979).

We endorse a rational model as the ideal for how science should proceed. Given the progress associated with the method, there is reason to think that the methodology of science has, in general, resulted in choices between competing theories primarily on the strength of the supporting evidence. However, our rationalism is temperate in that we recognize that there is no set of completely specifiable rules defining the scientific method that can guarantee success and that weight should be given to empirically based inductive arguments even though they do not logically compel belief (see Newton-Smith, 1981, especially p. 268ff.).

We believe the statistical methods that are the primary subject matter of this book are consistent with this perspective and more compatible with this perspective than with some others. For example, thinking it is meaningful to attempt to detect a difference between fixed-population means seems inconsistent with a relativistic perspective. Similarly, using statistical methods rather than relying on one's ability to make immediate judgments about particular facts seems inconsistent with a logical positivist approach. In fact, *one can view the primary role of statistical analysis as an efficient means for summarizing evidence* (see Abelson, 1995; Rosenthal & Rubin, 1985; Scarr, 1997): Rather than being a royal road to a positively certain scientific conclusion, inferential statistics is a method for accomplishing a more modest but nonetheless critical goal, namely quantifying the evidence or uncertainty relevant to a particular statistical conclusion. Doing this well is certainly not all there is to science, which is part of what we are trying to make clear, but it is a first step in a process that must be viewed from a broader perspective. Because

there is no cookbook methodology that can take you from a data summary to a correct theory, it behooves the scientist to think through the philosophical position from which the evidence of particular studies is to be viewed. Doing so provides you with a framework within which to decide if the evidence available permits you to draw conclusions that you are willing to defend publicly. The result of a statistical test is only one, albeit important, consideration in this process of reaching substantive conclusions and making generalizations, a point we attempt to underscore further in Chapter 2.

INTRODUCTION TO THE FISHER TRADITION

Discussion of issues relating to philosophy of science may, at first blush, seem unrelated to statistics. And, in fact, some presentations of statistics may border on numerology—whereby certain rituals performed with a set of numbers are thought to produce meaningful conclusions, with the only responsibility for thought by the investigator being the need to avoid errors in the calculations. This non-thinking attitude is perhaps made more prevalent by the ready availability of computers and statistical software. Even more extreme, perhaps, is that in the context of "big data" many actions are triggered based on black box calculations that can result in automated actions (cf. Grimmer, 2015). For all their advantages in terms of computational speed and accuracy, software and automation may mislead some into thinking that, because calculations are no longer an issue or that data sets can be "big," there is nothing more to statistics than learning the syntax for your software or which options to "click." It thus becomes easier to avoid facing the central issue squarely: how do I defend my answers to the scientific questions of interest in this situation?

Statistical decisions, appropriately conceived, are essentially organized arguments. This is perhaps most obvious when the derivations of the statistical tests themselves are carried out in a mathematically rigorous fashion. (Although the point of the argument might be totally obscure to all but the most initiated, that it is a highly structured deductive argument is clear enough.) Thus, in a book on linear models, one could begin from first principles and proceed to prove the theorems necessary for use of the *F* tests and the associated probability tables. That is the approach taken in some mathematical statistics texts. However, rigorous treatment of linear models requires mastery of calculus at a reasonably high level that not many students of the behavioral sciences have achieved. Fortunately, this does not preclude acquiring a thorough understanding of how statistics in general and linear models in particular can be used effectively in behavioral science research.

The view of statistics as a kind of rational argument was one that the prime mover in the area, Sir Ronald A. Fisher (1890–1962), heartily endorsed. In fact, Fisher reportedly was dismayed that, by the end of his life, statistics was being taught "essentially as mathematics" with an over-elaborate notation apparently designed to make it appear difficult (Cochran, 1967, p. 1461). Fisher, however, saw statistics as being much more closely related to the experimental sciences in which the methods actually were to be used. He developed new methods in response to the practical needs he saw in serving as a consultant to researchers in various departments related to the biological sciences. A major portion of Fisher's contributions to mathematical statistics and to the design and analysis of experiments came early in his career, when he was chief statistician at the Rothamsted agricultural research center. Fisher, who later served as Galton Professor at the University of London and as professor of genetics at the University of Cambridge, was responsible for laying the foundations for a substantial part of the modern discipline of statistics (and genetics). Certainly, the development and dissemination of the analysis of variance and the *F* test named for him were directly due to Fisher. His writings, which span half a century, provide masterful insights into the process of designing and interpreting experiments. His *Design of*

Experiments (1935/1971) in particular can be read with great profit, regardless of mathematical background, and illustrates very effectively the close link that should exist between logical analysis and computations. It is the purpose of the remainder of this chapter to provide a brief introduction to the kind of statistical reasoning that characterizes the tradition that Fisher set in motion.

We should note that the Fisherian approach has not been without its detractors, either in his day or in ours. Although current widely used procedures of testing statistical hypotheses represent an amalgam of Fisher's approach with that of others (namely Jerzy Neyman and Egon Pearson; see Gigerenzer, 1993), Fisher was arguably the most important figure in the modern development of statistics (cf. Huberty, 1991), and thus it is useful to gain an appreciation for some of his basic ideas regarding statistical reasoning. One purpose in tracing the rationale of hypothesis testing to its origins is to place our presentation of statistical methods in some broader historical context, in something of the same way that the first part of this chapter attempted to locate statistical reasoning within a broader philosophical context. By highlighting some of the past and present controversy regarding statistical reasoning, we hope to communicate something of the dynamic and evolving nature of statistical methodology.

We begin by examining one of the most fundamental ideas in statistics. A critical ingredient in any statistical test is determining the probability, assuming the operation of only chance factors, of obtaining a result as extreme or more extreme than that indicated by the observed value of the test statistic. For example, in carrying out a one-sample z test manually in an elementary statistics course, one of the final steps is to translate the observed value of z into a probability (e.g., using a table like that in the Review of Basic Statistics in the Supplements section of *DesigningExperiments.com*, which is the website that accompanies this book). The probability being sought, which is called a *p value*, is the probability of obtaining a z score as extreme or more extreme than that observed. Whenever the test statistic follows a continuous distribution like the z, t, or F, any treatment of this problem that goes deeper than "you look it up in the table" or "your software will provide the value" requires the use of rather messy mathematical derivations. Fortunately, the same kind of argument can be developed in detail quite easily if inferences are based on a discrete probabilistic analysis of a situation rather than by making reference to a continuous distribution. Thus, we illustrate the development of a statistical test by using an example relying on a discrete probability distribution.[8] First, however, let us consider why any probability distribution is an appropriate tool for interpreting experiments.

"Interpretation and Its Reasoned Basis"

Fisher aspired to contribute to scientific epistemology, or how we can come to know through science; he believed that an integrated methodology of experimental design and statistical procedures together satisfied "all logical requirements of the complete process of adding to knowledge by experimentation" (Fisher, 1935/1971, p. 3). Thus, Fisher was a firm believer in the idea that inductive inferences, although uncertain, could be made rigorously and based on specified levels of quantitative evidence. Probability distributions were used in this specification of the evidence.[9] However, as we have indicated, in Fisher's view, statistics was not a rarefied mathematical exercise. Rather, it was intimately related to experimentation, which in turn was viewed not merely as the concern of laboratory scientists, but also as the prototypical avenue by which people learn from experience. Given this, Fisher believed that an understanding of scientific inference was the appropriate concern of any intelligent person.

Experiments, Fisher wrote, "are only experience carefully planned in advance and designed to form a secure basis of new knowledge" (1935/1971, p. 8). The goal is to design experiments in such a way that the inferences drawn are fully justified and logically compelled by the data, as Fisher explained in *Design of Experiments.* When Fisher advised experimenters in a section

entitled "Interpretation and Its Reasoned Basis" to know in advance how they will interpret any possible experimental outcome, he was not referring to the theoretical or conceptual mechanism responsible for producing an effect. The theoretical explanation for why a particular effect should be observed in the population is quite different from the statistical conclusion itself. Admittedly, the substantive interpretation is more problematic in the behavioral sciences than in the agricultural sciences, where the experimental manipulation (e.g., application of kinds of fertilizer) is itself the treatment of substantive interest rather than being only a plausible representation of a theoretical construct (Chow, 1988, p. 107). However, the details of the preliminary argument from sample observations to general statistical conclusions about the effectiveness of the experimental manipulation had not been worked out prior to Fisher's time. His key insight, which solved the problem of making valid statistical inferences, was that of randomization. In this way, one is assured that no uncontrolled factor would bias the results of the statistical test. The details of how this works out in practice are illustrated in subsequent sections.

For the moment, it is sufficient to note that the abstract random process and its associated probabilities are merely the mathematical counterparts of the use of randomization in the concrete experimental situation. Thus, in any true experiment, there are points in the procedure when the laws of chance are explicitly introduced and are in sole control of what is to be done. For example, one might flip a coin (or simulate such a process by using a pseudo-random number generator) to determine what treatment a particular participant receives. The probability distribution used in the statistical test makes sense only because of the use of random assignment in the conduct of the experiment. By doing so, one assures that, if the null hypothesis of no difference between treatments is correct, the results of the experiment are determined entirely by the laws of chance (Fisher, 1935/1971, p. 17). One might imagine, for example, a wide variety of factors that would determine how a particular phobic might respond on a posttest of performance in the feared situation after receiving one of an assortment of treatments. Assuming the treatments have no effect, any number of factors—such as the individual's conditioning history, reaction to the experiment, or indigestion from a hurried lunch—might in some way affect performance. If, in the most extreme view, the particular posttest performance of each individual who could take part in your experiment was thought to be completely determined from the outset by a number of, for your purposes, irrelevant factors, the random assignment to treatment conditions assures that, in the long run, these would balance out. That is, *randomization* implies that the population means in the various treatments are, under these conditions, exactly equal, and that even the form of the distribution of scores in the various conditions is the same.

Next, we show how this simple idea of control of irrelevant factors by randomization works in a situation that can be described by a discrete probability distribution. Thus, we are able to derive (by using only simple counting rules) the entire probability distribution that can be used as the basis for a statistical test.

A Discrete Probability Example

Fisher introduced the principles of experimentation in his *Design of Experiments* (1935/1971) with an appropriately British example that has been used repeatedly to illustrate the power of randomization and the logic of hypothesis testing (see, e.g., Kempthorne, 1952, pp. 14–17, 120–134; Salsburg, 2001). We simply quote the original description of the problem:

> A lady declares that by tasting a cup of tea made with milk, she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested.
>
> (Fisher, 1935/1971, p. 11)

[Some who belittle single-subject designs compared to group experimentation might be bemused to realize that the principles of group experimentation were originally introduced with an *N*-of-1 (i.e., single subject) design. In fact, to be accurate in assigning historical priority, it was the distinguished American philosopher and mathematician Charles S. Peirce, working on single-subject experiments in psychophysics in the 1880s, who first discussed the advantages of randomization (Stigler, 1999, p. 192ff.). However, it was a half-century later before Fisher tied these explicitly to methods for arriving at probabilistic inferences.] If you try to come up with an exemplary design appropriate for this particular problem, your first thought might be of the variety of possible disturbing factors over which you would like to exert experimental control. That is, you may begin by asking what factors could influence her judgment and how could these be held constant across conditions so that the only difference between the two kinds of cups is whether the milk or tea was added first. For example, variation in the temperature of the tea might be an important clue, so you might carefully measure the temperature of the mixture in each cup to attempt to assure they were equally hot when they were served. Numerous other factors could also influence her judgment, some of which may be susceptible to experimental control. The type of cup used, the strength of the tea, the use of sugar, and the amount of milk added merely illustrate the myriad potential differences that might occur among the cups to be used in the experiment. The logic of experimentation until the time of Fisher dictated that to have a valid experiment here, all the cups to be used "must be exactly alike," except for the independent variable being manipulated. Fisher rejected this dictum on two grounds. First, he argued that exact equivalence was logically impossible to achieve, both in the example and in experimentation in general. The cups would inevitably differ to some degree in their smoothness, the strength of the tea and the temperature would change slightly over the time between preparation of the first and last cups, and the amounts of milk or sugar added would not be exactly equal, to mention only a few problems. Second, Fisher argued that, even if it were conceivable to achieve "exact likeness" or, more realistically, "imperceptible difference" on various dimensions of the stimuli, it would in practice be too expensive to attempt. Although one could, with a sufficient investment of time and money, reduce the irrelevant differences between conditions to a specified criterion on any dimension, the question of whether it is worth the effort must be raised in any actual experiment. The foremost concern with this and other attempts at experimental control is to arrive at an appropriate test of the hypothesis of interest. Fisher argued that, because the validity of the experiment could be assured by the use of randomization, it was not the best use of inevitably limited resources to attempt to achieve exact equality of stimuli on all dimensions. Most causes of fluctuation in participants' performance "ought to be deliberately ignored" (1935/1971, p. 19).

Consider now how one might carry out and analyze an experiment to test the British lady's claim. The difficulty with asking for a single judgment, of course, is that she might well correctly classify it just by guessing. How many cups then would be needed to constitute a test that provided a sufficient level of evidence that she could, indeed, tell if milk or tea was added first? The answer naturally depends on how the experiment is designed, as well as the criterion adopted for how strong the evidence must be in order to be considered compelling.

One suggestion might be that the experiment be carried out by mixing eight cups of tea, four with the milk added to the cup first (milk-first, or MF, cups) and four with the tea added first (tea-first, or TF, cups), and presenting them for classification by the subject in random order. Is this a sufficient number of judgments to request?

> In considering the appropriateness of any proposed experimental design, it is always needful to forecast all possible results of the experiment, and to have decided without ambiguity

> what interpretation shall be placed upon each one of them. Further, we must know by what argument this interpretation is to be sustained.
>
> (Fisher, 1935/1971, p. 12)

Thus, Fisher's advice translated into the current vernacular might be, "If you can't analyze an experiment, don't run it." To prescribe the analysis of the suggested design, we must consider what the possible results of the experiment are and the likelihood of the occurrence of each. To be appropriate, the analysis must correspond exactly to what actually went on in the experiment.[10] Assume the subject is told that the set of eight cups consists of four MF and four TF cups. The measure that indicates how compelling the evidence could be is the probability of a perfect performance occurring by chance alone. If this probability is sufficiently small, say less than 1 chance in 20, we conclude it is implausible that the lady has no discrimination ability. There are, of course, many ways the participant may divide the set of eight cups into two groups of four each, with the participant thinking that one group consists of MF cups and the other group, TF cups. However, if the participant cannot discriminate at all between the two kinds of cups, each of the possible divisions into two groups would be equally likely, as the participant would be dividing the eight cups into the two groups of four essentially at random.

Thus, the probability of a correct performance occurring by chance alone could be expressed simply as the proportion of the possible divisions of the cups that are correct:

$$\Pr\left(\text{being correct by chance}\right) = \frac{\text{Number of divisions that are exactly correct}}{\text{Total number of possible divisions}} \tag{1}$$

Naturally, only one division would match exactly the actual breakdown into MF and TF cups, which means the numerator of the fraction in Equation 1 would be 1. The only problem, then, is to determine the total number of ways of splitting up eight things into two groups of four each. Actually, we can solve this by determining only the number of ways the subject could select a particular set of four cups as being the MF cups; because once four are chosen as being of one kind, the other four must be put into the other category. Formulating the solution in terms of a sequence of decisions is easiest. Any one of the eight cups could be the first to be classified as an MF cup. For each of the eight possible ways of making this first decision, there are seven remaining cups from which to choose the second cup to be classified as an MF cup. Given the 8×7, or 56, ways of making the first two decisions, there are six ways of choosing the third MF cup. Finally, for each of these $8 \times 7 \times 6$ orderings of three cups, there would be five possible ways of selecting the fourth cup to be assigned to the MF category. Thus, there are $8 \times 7 \times 6 \times 5$, or 1,680 ways of choosing four cups out of eight *in a particular order*. However, each set of four particular cups would appear $4 \times 3 \times 2 \times 1$, or 24, times in a listing of the 1,680 orderings, because for any set of four objects, any one of the four could be the first chosen, any one of the remaining three could be second chosen, and either of the remaining two could be the third chosen, leaving just one way of "choosing" the fourth. We are not concerned with the particular sequence in which the cups in a set of four were selected, only with which set was selected. Thus, we can find the number of *distinct* sets of cups by dividing the number of orderings (1,680) by the number of ways (24) that each distinct set could be ordered. In summary,

$$\text{Total number of distinct sets of four cups} = \frac{8\times7\times6\times5}{4\times3\times2\times1} = \frac{1680}{24} = 70 \tag{2}$$

Those who have studied what are known as *counting rules*, or "permutations and combinations," may recognize this solution as *the number of combinations of eight things taken four at a time*, which may be denoted ${}_8C_4$. In general, if one is selecting r objects from a larger set n, by the reasoning followed previously, we write

$$ {}_nC_r = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r(r-1)(r-2)\cdots 1} = \frac{n!}{r!(n-r)!} \tag{3} $$

The solution here, of there being 70 different combinations or sets of four cups, which could possibly be designated as MF cups, is critical to the interpretation of the experiment. The result of Equation 3 can be understood as indicating "how many ways r items can be chosen from n items." Importantly, in this context selecting items (A, B) would not be counted differently than selecting items (B, A), as it is the combination of items and not the order that is important. Permutations, by the way, count different orderings separately (and thus A, B is a different permutation than B, A). Following Equation 1, because only 1 of these 70 possible answers is correct, the probability of the lady being exactly right by chance alone is 1/70. Because this is less than the 1/20, or .05, probability we adopted as our criterion for being *so unlikely as to be convincing*, if the lady were to correctly classify all the cups, we would have a sufficient basis for rejecting the null hypothesis of no discrimination ability.

Notice that in essence, we have formulated a statistical test of our null hypothesis, and instead of looking up a p value for an outcome of our experiment in a table, we have derived that value ourselves based on the situation. Because the experiment involved discrete events rather than scores on a continuous variable, we were able to simply use the definition of probability and a counting rule, which we also developed "from scratch" for our situation, to determine a probability that could be used to judge the statistical significance of one possible outcome of our experiment.

Although no small feat, we admittedly have not yet considered "all possible results of the experiment," deciding "without ambiguity what interpretation shall be placed on each one." One plausible outcome is that the lady might get most of the classifications correct, but fall short of perfect performance. In the current situation, this would necessarily mean that three of the four MF cups would be correctly classified. Note that, because the participant's response is to consist of putting four cups into each category, misclassifying one MF cup necessarily means that one TF cup was inappropriately thought to be a MF cup. Note also that the decision about which TF cup is misclassified can be made apart from the decision about which MF cup is misclassified. Each of these two decisions may be thought of as a combinatorial problem: How many ways can one choose three things out of four? and How many ways can one thing be selected out of four? Thus, the number of ways of making one error in each grouping of cups is

$$ \begin{aligned} \text{Number of ways of making one error of each kind} &= {}_4C_3 \cdot {}_4C_1 \\ &= \frac{4!}{3!1!} \cdot \frac{4!}{1!3!} = 4 \cdot 4 = 16 \end{aligned} \tag{4} $$

It may seem surprising that there are as many as 16 ways to arrive at three out of four correctly classified MF cups. However, any one of the four could be the one to be left out, and for each of these, any one of four wrong cups could be put in its place.

Making use again of the definition of the probability of an event as the number of ways that event could occur over the total number of outcomes possible, we can determine the probability of this near-perfect performance arising by chance. The numerator is what was just determined,

and the denominator is again the number of possible divisions of eight objects into two sets of four each, which we previously (Equation 2) determined to be 70:

$$\Pr(\text{three MF and one TF classified as MF}) = \frac{{}_4\mathrm{C}_3 \cdot {}_4\mathrm{C}_1}{{}_8\mathrm{C}_4} = \frac{4 \cdot 4}{70} = \frac{16}{70} \tag{5}$$

In general, if one selects n things at random from a larger set of N things, where the larger set consists of R things which are designated as the "right" things or events of interest and $N - R$ things designated as the "wrong" answers, the probability of including exactly r of the R things (and $n - r$ of the $N - R$ incorrect things) in the selected subset is:

$$\Pr(r \text{ out of } R \text{ included in } n \text{ selected}) = \frac{{}_R\mathrm{C}_r \cdot {}_{N-R}\mathrm{C}_{n-r}}{{}_N\mathrm{C}_n} \tag{6}$$

The fact that here this probability of 16/70, or .229, is considerably greater than our criterion of .05 puts us in a position to interpret not only this outcome, but all other possible outcomes of the experiment as well. Even though three out of four right represents the next best thing to perfect performance, the fact that performance that good or better could arise (16 + 1)/70 = .243, or nearly one-fourth, of the time when the subject had no ability to discriminate between the cups implies it would not be good enough to convince us of her claim. Also, because all other possible outcomes would be less compelling, they would also be interpreted as providing insufficient evidence to make us believe that the lady could determine which were the MF cups.

Let us now underscore the major point of what we have developed in this section. Although we have not made reference to any continuous distribution, we have developed, from basic principles of probability, a statistical test appropriate for use in the interpretation of a particular experiment. The test is in fact more generally useful and is referred to in the literature as the *Fisher-Irwin exact test* (Marascuilo & Serlin, 1988, p. 200ff.), or more commonly as *Fisher's exact test* (e.g., Hays, 1994, p. 863).

Many statistical packages include Fisher's exact test as at least an optional test in analyses of cross-tabulated categorical data. In SPSS, both one-tailed and two-tailed p levels for Fisher's exact test are computed to supplement chi-square tests for 2×2 tables in the Crosstabs procedure. Although our purpose in this section primarily is to illustrate how p values may be computed from first principles, we comment briefly on some other issues that we develop more fully in later chapters. In general, in actual data analysis situations it is desirable not just to carry out a significance test, but also to characterize the magnitude of the effect observed. There are usually multiple ways in which this can be done, and that is true in this simple case of analysis of a 2×2 table, as will be the case in more complicated situations. One way of characterizing the magnitude of the effect is by using the phi coefficient, which is a special case for a 2×2 table of the well-known Pearson product-moment correlation coefficient. For example, in the case in which one error of each kind was made in the classification of eight cups, the effect size measure could be computed as the correlation between two numerical variables, say Actual and Judged. With only two levels possible, the particular numerical values used to designate the level of TF or MF are arbitrary, but one would have eight pairs of scores [e.g., (1,1), (1,1), (1,1), (1,2), (2,1), (2,2), (2,2), (2,2)], which would here result in a correlation or phi coefficient between Actual and Judged of .50. Small, medium, and large effect sizes may be identified with phi coefficients of .10, .30, and .50, respectively (Cohen, 1988, chap. 7). We discuss pros and cons of such "benchmarks" of effect sizes in Chapter 3.

An alternative approach to characterizing the effect size is to think of the two rows of the 2×2 table as each being characterized, over replications of the experiment, by a particular probability of "success" or probability of an observation falling in the first column, say p_1 for row 1 and

p_2 for row 2. Then, one could describe the magnitude of the effect as the estimated difference between these probabilities, or $\hat{p}_1 - \hat{p}_2$. However, one difficulty with interpreting such a difference is that the relative chances of success can be very different with small as opposed to large probabilities. For example, a difference of .1 could mean the probability of success is 11 times greater in one condition than in the other if $p_1 = .11$ and $p_2 = .01$, or it could mean that one probability is only 1.2 times the other if $p_1 = .60$ and $p_2 = .50$. To avoid this difficulty, it is useful for some purposes to measure the effect size in terms of the ratio of the odds of success in the two rows. The odds ratio is defined as

$$\frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} \tag{7}$$

Methods for constructing confidence intervals around estimates of the odds ratio are discussed by Good (2000, p. 100) and Rosner (1995, pp. 368–370). The projected odds ratio in the population can also be used in planning a study to determine what sample size might be required to have a reasonable chance of detecting the expected effect.[11]

It perhaps should be mentioned that Fisher's exact test, besides illustrating how one can determine the probability of an outcome of an experiment, can be viewed as the forerunner of a host of other statistical procedures. These are particularly useful in those research areas—for example, some types of public health or sociological research—in which all variables under investigation may be categorical. A number of good introductions to such methods are available (see, e.g., Agresti, 2012; Bishop, Fienberg, & Holland, 1975; Nussbaum, 2014; Stokes, Davis, & Koch, 2001).

Although these methods have some use in the behavioral sciences, it is much more common for the dependent variable in experiments to be quantitative instead of qualitative. Thus, we continue our introduction to the Fisher tradition by considering another example from his writing that makes use of a quantitative dependent variable. Again, however, no reference to a theoretical population distribution is required.

Randomization Test

Assume that a developmental psychologist is interested in whether brief training can improve performance of 2-year-old children on a test of mental abilities. The test selected is the Mental Scale of the Bayley Scales of Infant Development, which yields a mental age in months. To increase the sensitivity of the experiment, the psychologist decides to recruit sets of twins and randomly assigns one member of each pair to the treatment condition. The treatment consists simply of watching a videotape of another child attempting to perform tasks similar to those making up the Bayley Mental Scale. The other member of each pair plays in a waiting area as a time-filling activity while the first is viewing the videotape. Then both children are individually given the Bayley by a tester who is blind to their assigned conditions. A different set of twins takes part in the experiment each day, Monday through Friday, and the experiment extends over a 2-week period, so that 10 twin pairs contribute data. Table 1.1 shows the data for the study in the middle columns.

Given the well-known correlations between twins' mental abilities, it would be expected that there would be some relationship between the mental ability scores for the two twins from the same family, although the correlation is considerably lower at age 2 than at age 18. (Behavior of any 2-year-old is notoriously variable from one time to another; thus, substantial changes in even a single child's test performance across testing sessions are common.) The measure of treatment effectiveness that would commonly be used then in such a study is simply the difference between

TABLE 1.1
SCORES ON BAYLEY MENTAL SCALE (IN MONTHS) FOR TEN PAIRS OF TWINS

	Condition		*Difference*
Twin Pair	*Treatment*	*Control*	(*Treatment* − *Control*)
Week 1 data			
1	28	32	−4
2	31	25	6
3	25	15	10
4	23	25	−2
5	28	16	12
Sum for Week 1	**135**	**113**	**22**
Week 2 data			
6	26	30	−4
7	36	24	12
8	23	13	10
9	23	25	−2
10	24	16	8
Sum for Week 2	**132**	**108**	**24**
Sum for 2 weeks	**267**	**221**	**46**
Mean for 2 weeks	**26.7**	**22.1**	**4.6**

the score of the child in the treatment condition and that of his or her twin in the control condition. These are shown on the right side of Table 1.1.

A *t* test would typically be performed to make an inference about the mean of these differences in the population. For this particular data set, some hesitation might arise because the sample distribution is U-shaped[12] rather than the bell-shaped distribution that would be expected if the assumption made by the *t* test of a normal population were correct. The *t* test might in practice be used despite this (see the discussion of assumptions at the end of Chapter 3). However, it is not necessary to make any assumptions about the form of the population distribution in order to carry out certain tests of interest here. In fact, one can use all the quantitative information available in the sample data in testing what Fisher referred to as "the wider hypothesis" (1935/1971, p. 43) that the two groups of scores are samples from the same, possibly non-normal population.

The test of this more general hypothesis is based simply on the implications of the fact that subjects were randomly assigned to conditions. Hence, the test is referred to as a *randomization test.* The logic is as follows: if the null hypothesis is correct, then subjects' scores in the experiment are determined by factors other than what treatment they were assigned (that is, the treatment did not influence subjects' scores). In fact, one may consider the score for each subject to be predetermined prior to the random assignment to conditions (i.e., there was no effect of the treatment). Thus, the difference between any two siblings' scores would have been the same in absolute value regardless of the assignment to conditions. For example, under the null hypothesis, one subject in Pair 1 was going to receive a score of 28 and the other subject a score of 32; the random assignment then simply determined that the higher-scoring subject would be in the control condition here, so that the difference of "treatment minus control" would be −4 instead of +4. Because a random assignment was made independently for each of the 10 pairs, 10 binary decisions were in effect made as to whether a predetermined difference would have a plus or minus sign attached

to it. Thus, there were 2^{10} possible combinations of signed differences that could have occurred with these subjects, and the sum of the signed differences could be used to indicate the apparent benefit (or harm) of the treatment for each combination. Note that the "2" in 2^{10} is due to the two sign options: positive or negative; and the "10" in 2^{10} is due to the 10 pairs. The distribution of these 2^{10} sums is the basis for our test. The sum of the differences actually observed, including the four negative differences, was 46. A randomization test is carried out simply by determining how many of the 2^{10} combinations of signed differences would have totals equal to or exceeding the observed total of 46. Because under the null hypothesis each of these 2^{10} combinations is equally likely, the proportion of them having sums at least as great as the observed sum provides directly the probability to use in assessing the significance of the observed sum.

In effect, one is constructing the distribution of values of a test statistic (the sum of the differences) over *all possible* reassignments of subjects to conditions. Determining where the observed total falls in this distribution is comparable to what is done when one consults a table in a parametric test (e.g., *t* test, *F* test) to determine the significance of an observed value of a test statistic. Indeed, a *t* test (for paired samples) is an alternative way of addressing this question (although more assumptions are made, as noted earlier). However, with the randomization test, the distribution is based directly on the scores actually observed rather than on some assumed theoretical distribution (such as the differences following a normal distribution).

Using all the quantitative information in the sample and obtaining a statistical test without needing to make any distributional assumptions makes an attractive combination. There are disadvantages, however. A major disadvantage that essentially prevented use of randomization tests until recent years in all but the smallest data sets is the large number of computations required. To completely determine the distribution of possible totals for even the set of 10 differences in Table 1.1 would require examining $2^{10} = 1{,}024$ combinations of the observed scores (i.e., all ways in which the signed differences could have been observed). With 20 pairs, more than a million random assignments would need to be considered; with 30 pairs, more than a billion. We summarize the results of this process later, but illustrate the computations for the smaller data set consisting only of the five scores from week 1.

With five scores, there are $2^5 = 32$ possible assignments of positive and negative signs to the individual scores. Table 1.2 lists the scores in rank order of their absolute value at the top left. Then, 15 other sets, including progressively more minus signs, are listed along with the sum for each. The sums for the remaining 16 sets are immediately determined by realizing that when the largest number of 12 is assigned a negative rather than a positive sign, the sum would be reduced by 24.

If the first week constituted the entire experiment, these 32 sums would allow us to determine the significance of the observed total Bayley difference for the first week of 22 ($= -4 + 6 + 10 - 2 + 12$, see Table 1.1). Figure 1.2 shows a grouped, relative frequency histogram for the possible sums, with the shaded portion on the right indicating the sums greater than or equal to the observed sum of 22. (An ungrouped histogram, although still perfectly symmetrical, appears somewhat less regular.) Thus, the probability of a total at least as large as and in the same direction as that observed would, under the null hypothesis, be 5/32 ($= 3/32 + 2/32$), or .16, which would not be sufficiently small for us to claim significance.

The same procedure could be followed for the entire set of 10 scores. Rather than listing the 1,024 combinations of scores or displaying the distribution of totals, the information needed to perform a test of significance can be summarized by indicating the number of totals greater than or equal to the observed sum of 46. Fortunately, it is clear that if five or more numbers were assigned negative signs, the total would necessarily be less than 46. Table 1.3 shows the breakdown for the other possible combinations.

TABLE 1.2
POSSIBLE SUMS OF DIFFERENCES RESULTING FROM REASSIGNMENTS OF FIRST-WEEK CASES

	Assignment															
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	10	10	10	10	10	10	10	10	−10	−10	−10	−10	−10	−10	−10	−10
	6	6	6	6	−6	−6	−6	−6	6	6	6	6	−6	−6	−6	−6
	4	4	−4	−4	4	4	−4	−4	4	4	−4	−4	4	4	−4	−4
	2	−2	2	−2	2	−2	2	−2	2	−2	2	−2	2	−2	2	−2
Sum	**34**	**30**	**26**	**22**	**22**	**18**	**14**	**10**	**14**	**10**	**6**	**2**	**2**	**−2**	**−6**	**−10**
	Assignment[a]															
	17	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*	*31*	*32*
	−12	−12	−12	−12	−12	−12	−12	−12	−12	−12	−12	−12	−12	−12	−12	−12
	10	10	10	10	10	10	10	10	−10	−10	−10	−10	−10	−10	−10	−10
	6	6	6	6	−6	−6	−6	−6	6	6	6	6	−6	−6	−6	−6
	4	4	−4	−4	4	4	−4	−4	4	4	−4	−4	4	4	−4	−4
	2	−2	2	−2	2	−2	2	−2	2	−2	2	−2	2	−2	2	−2
Sum	**10**	**6**	**2**	**−2**	**−2**	**−6**	**−10**	**−14**	**−10**	**−14**	**−18**	**−22**	**−22**	**−26**	**−30**	**−34**

[a] Note that assignments 17–32 are the same as assignments 1–16 except that 12 is assigned a negative sign rather than a positive sign, and so each sum is 24 less than the sum for the corresponding assignment.

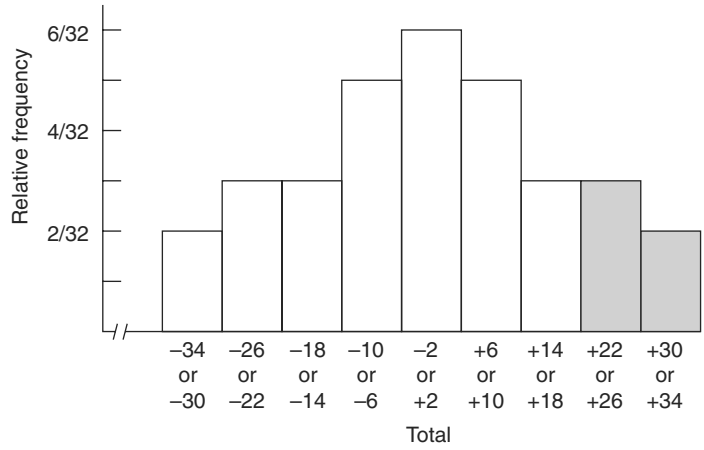


FIG. 1.2 Distribution of possible totals of difference scores using data from Week 1.

TABLE 1.3
NUMBER OF COMBINATIONS OF SIGNED DIFFERENCES WITH SUMS EQUAL TO OR GREATER THAN THE OBSERVED SUM

Number of Negative Values	*Total Number of Combinations*	*Number of Combinations with*		
		***Sum* > 46**	***Sum* = 46**	***Sum* < 46**
0	1	1		
1	10	8	2	
2	45	12	6	27
3	120	5	5	110
4	210		1	209
5	252			252
6	210			210
7	120			120
8	45			45
9	10			10
10	1			1
Totals	**1024**	**26**	**14**	**984**

We now have the needed information to address the question with which we began this section: does brief training improve the performance of 2-year-olds on a test of mental abilities? Under the null hypothesis that the scores from the subjects receiving training and those not receiving training represent correlated samples from two populations having identical population distributions, the random assignment to conditions allows us to generate a distribution of possible totals of 10 difference scores based on all the data actually observed. As shown in Table 1.3, we find that only 40 of 1,024, or .039, of the possible combinations of signed differences result in totals as large or larger than that actually observed. Thus, we conclude that we have significant evidence that our training resulted in improved performance among the children tested in the experiment.

Two points about this conclusion are noteworthy. First, we performed a one-tailed test. A one-tailed test might be warranted in an applied setting in which one is interested in the treatment only if it helps performance. If a two-tailed test had been performed, a different conclusion would have been reached. To see this, we make use of the symmetry of the distributions used in randomization tests (every combination of signed differences is matched by one in which every sign is reversed, so every positive total has a corresponding negative total of the same absolute value). Thus, there would be exactly 40 cases totaling −46 or less. This yields a combined probability (i.e., a two-sided *p* value) of 80/1024, or .078, of observing a total as extreme or more extreme *in either direction* than that observed; hence, we would fail to reject the null hypothesis in favor of a non-directional alternative hypothesis (using the standard Type I error rate of .05).

Second, it should be pointed out that the hypothesis tested by the randomization test is not identical to that tested by the *t* test. The hypothesis in the *t* test concerns the population mean of a continuous random variable. The hypothesis in the randomization test concerns the presumption that each of the observed difference scores could have been preceded by a positive or negative sign with equal likelihood. The *p* value yielded by performing a *t* test would be exact only if the theoretical distribution prescribed by its density formula were perfectly matched by the actual distribution of the test statistic given the current population, which it certainly is not here.[13] However, in part because of the factors summarized by the Central Limit Theorem (discussed in the next section), the *p* value in the table often is a very good approximation to the exact *p* value from a randomization test even with non-normal data. Note that the *p* value in the randomization test is the exact probability only for the distribution arising from hypothetical reassignments of the particular cases used in the study (Edgington, 1966, 1995). However, the closeness of the correspondence between the *p* value yielded by the randomization test and that yielded by the *t* test can be demonstrated mathematically under certain conditions (Pitman, 1937).

We can illustrate this correspondence in the current example as well. If we perform a *t* test of the hypothesis that the mean difference score in the population is 0, we obtain a *t* value of 2.14 with 9 degrees of freedom. This observed *t* value is exceeded by .031 of the theoretical *t* distribution, which compares rather closely with the .039 we obtained from our randomization test previously. The correspondence is even closer if, as Fisher suggested (1935/1971, p. 46), we correct the *t* test for the discontinuous nature of our data.[14] Hence, with only 10 cases, the difference between the probabilities yielded by the two tests is on the order of 1 in 1000. In fact, one may view the *t* test and the randomization test as very close approximations to one another (cf., Lehman, 1986, pp. 230–236). Deciding to reject the hypothesis of the randomization test is tantamount to deciding to reject the hypothesis of the *t* test.[15]

As with the Fisher's exact test, our purpose with the randomization test is primarily to emphasize the meaning of *p* values rather than to fully develop all aspects of the methodology. When such a method is used in actual research, one may want to construct a confidence interval around a parameter indicating the location or central tendency of the distribution in the population. Methods for doing so are discussed briefly in Good (2000, pp. 34–35) and in more theoretical detail in Lehmann (1986, pp. 245–248). Power of randomization tests is considered by Bradbury (1987), Robinson (1973), and Good (2000, p. 36), and is often similar to that of the standard *t* test. We consider measures of effect size and power for group comparisons in the context of the linear models introduced in subsequent chapters. Anderson (2001) provides a brief overview of how randomization tests, or permutation tests as they are sometimes called, can be used in various ANOVA and regression settings.

There are currently a variety of software options available for carrying out randomization tests. At *DesigningExperiments.com/chapter-1*, we provide syntax and instructions for carrying out approximate randomization tests for within-subject and between-subject (Hayes, 1998) designs via SPSS. The tests are approximate in that they are based on a large random sample

(e.g., 10,000) from the population of possible re-randomizations, because the total number of re-randomizations may be too large to exhaustively search for even a moderately large design. We also provide at *DesigningExperiments.com* code for using R to carry out an exact randomization test for a matched-pairs design like the Bayley numerical example discussed earlier. The R "coin" package[16] can perform exact randomization tests for the two-group case and approximate tests for larger designs (Hothorn, Hornik, van de Wiel, & Zeileis, 2006). Syntax has also been published for performing randomization tests in SAS (Chen & Dunlap, 1993). Some specialized, commercially available programs also perform such tests (e.g., StatXact, from Cytel Software at www.cytel.com), and one has the option of obtaining programs for free from published program listings (Edgington, 1995).

Of Hypotheses and *p* Values: Fisher Versus Neyman-Pearson

To this point, we have dealt with only a single hypothesis, namely the null hypothesis. This was Fisher's strong preference (Huberty, 1987). The familiar procedure of simultaneously considering a null and an alternative hypothesis, which became standard practice in psychology in the 1940s (Huberty, 1991; Rucci & Tweney, 1980), is actually a modification of Fisherian practice that had been advocated by statisticians Jerzy Neyman and Egon Pearson. One particularly memorable version of the historical debates regarding statistical methods and how they manifest themselves currently is that offered in Freudian terms by Gigerenzer (1993).

In the Neyman–Pearson view, statistical inference was essentially an exercise in decision making. Whereas Fisher had viewed significance testing as a means of summarizing data to aid in advancing an argument for a position on a scientific question, Neyman and Pearson emphasized the practical choice between two statistical hypotheses, the null hypothesis and its complement the alternative hypothesis. The benefit of this approach was to make clear that one could not only make a Type I error (with probability α or alpha) of rejecting the null hypothesis when it is true, but also a Type II error, or failing to reject the null hypothesis when it is false (with probability β or beta). In practical situations in business or medicine, or in exploratory research in science more generally, one could adjust the probabilities of these errors to reflect the relative costs and benefits of the different kinds of errors. Determining a particular value of β required one to specify an exact alternative hypothesis (e.g., $\mu = 105$, not just $\mu \neq 100$).

One disadvantage of the Neyman–Pearson approach was the overemphasis on the accept–reject decision. Although a 5% level of significance was acknowledged as "usual and convenient" by even Fisher (1935/1971, p. 13), thinking that an up-or-down decision is a sufficient summary of the data in all situations is clearly misguided. For one, an effect of identical size might be declared significant in one study but not another simply because of differences in the number of subjects used. Although abandoning significance tests, as some advocate (e.g., Cohen, 1994; Oakes, 1986), would avoid this problem, one thereby would lose this critical screen that prevents researchers from interpreting what could reasonably be attributed to chance variation (cf. Frick, 1996; Hagen, 1997). However, viewing the alpha level (or probability of a false positive decision) established before the experiment as the only probability that should be reported suppresses information. Some researchers apparently believe that what statistical correctness requires is to report all significant *p* values only as significant at the alpha level established before the experiment. Thus, the "superego" of Neyman–Pearson logic might seem to direct that if α is set at 5% before the experiment, then .049 and .003 should both be reported only as "significant at the .05 level" (Gigerenzer, 1993). But, as Browner and Newman (1987) suggest, all significant *p* values are not created equal. Although there is value in retaining the conventions of .05 and .01 for declaring results significant or highly significant, any published report of a statistical test, in our view and that of groups of experts asked to make recommendations on such issues, should

be accompanied by an exact *p* value (Greenwald, Gonzalez, Harris, & Guthrie, 1996, p. 181; Wilkinson & the APA Task Force on Statistical Inference, 1999, p. 599). As Fisher saw it, this is part of the information that should be communicated to others in the spirit of freedom that is the essence of the Western tradition. Reporting exact *p* values recognizes "the right of *other* free minds to utilize them in making *their own* decisions" [Fisher, 1955, p. 77 (italics Fisher's)].

Because we emphasize relying on and reporting *p* values, it is critical to be clear about what they are and what they are not. As we tried to make clear by our detailed development of the *p* values for the Fisher's exact and randomization tests, a *p* value is *the probability of data as extreme or more extreme as that obtained, computed under the presumption of the truth of the null hypothesis*. In symbols, if we let *D* stand for data as or more extreme as that obtained, and H_0 stand for the null hypothesis, then a *p* value is a conditional probability of the form *Probability* $(D \mid H_0) = p$ value.

Unfortunately, erroneous interpretations of *p* values by academic psychologists, including textbook authors and journal editors, are very common and have been well documented, often by those raising concerns about hypothesis testing. Two misunderstandings seem to be most prevalent. The first has been termed the *replication fallacy,* which is erroneously thinking that a significant *p* value is the complement of the probability (i.e., $1 - p$) that a replication of the study would also yield significance. However, the probability of obtaining significance in a replication when the null hypothesis is false refers to the concept of *power*, which can be computed only under the assumption of a specific alternative hypothesis, and in any event is only indirectly related to the *p* value. Gigerenzer (1993) provides a number of examples of the replication fallacy, including an example from Nunnally's (1975) *Introduction to Statistics for Psychology and Education,* which asserted " 'If the statistical significance is at the 0.05 level . . . the investigator can be confident with odds of 95 out of 100 that the observed difference will hold up in future investigations' (Nunnally, 1975, p. 195)" (Gigerenzer, 1993, p. 330). A study conducted by a British psychologist of 70 university lecturers, research fellows, and postgraduate students elicited endorsement by 60% of a statement to the effect that a result significant at $p = .01$ meant that "You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions" (Oakes, 1986, pp. 79–80). In point of fact, an experiment that yields a *p* value of .05 would theoretically lead to a probability of a significant replication of only about .50, not .95 (Greenwald et al., 1996; Hoenig & Heisey, 2001). So, neither the exact *p* value nor its complement can be interpreted as the probability of a significant replication. However, the point that some strident critics of null hypothesis testing overlook but that contributes to the enduring utility of the methodology is "replicability of a null hypothesis rejection is a continuous, increasing function of the complement of its *p* value" (Greenwald et al., 1996, p. 181). The exact probability of a successful replication depends on a number of factors, but some helpful guidance is provided by Greenwald et al. (1996), who show that under certain simplifying assumptions, *p* values can be translated into a probability of successful replication (power) at $\alpha = .05$ as follows: $p = .05 \rightarrow$ power $\approx .5$; $p = .01 \rightarrow$ power $\approx .75$; $p = .005 \rightarrow$ power $\approx .8$; and $p = .001 \rightarrow$ power $> .9$. A striking empirical demonstration of the relationship between *p* values and probability of successful replication was seen in the 2015 report in *Science* of replications of 100 cognitive psychology and social psychology studies published in three leading journals (Open Science Collaboration, 2015). Whereas 97% of the original studies reported significant ($p < .05$) results, 36% of the replications of the originally significant results obtained significance ($p < .05$). Importantly, when features of the original study that might correlate with reproducibility were examined (including the original effect size, the experience and expertise of the original researchers, the importance and surprisingness of the original result), the best single predictor of the replication being successful was the *p* value in the original study. When the original *p* value was > .04, only 18%

of the replications yielded $p < .05$; when the original p was $.02 < p < .04$, 26% of the replications had $p < .05$; when the original p was $< .02$, 41% of replications were significant; and when the original p was $< .001$, nearly two thirds (63%) of the replications were significant. Although the probability of successful replication was somewhat less in general than expected, likely in part because of publication bias favoring positive results operating in the original studies, the value of p values as an indicator of replicability was clearly demonstrated.

The second prevalent misinterpretation of p values is as indicating an inverse probability, that is, the probability that a hypothesis is true or false given the data obtained [e.g., $\Pr(H_0 \mid D)$], instead of the probability of data given the null hypothesis is assumed true. Again, textbooks as well as research psychologists provide numerous examples of this fallacy (Cohen, 1994, p. 999, lists various sources reporting examples). For example, when hypothesis testing was first being introduced to psychologists in the 1940s and 1950s, the leading text, Guilford's *Fundamental Statistics in Psychology and Education*, included headings such as " 'Probability of hypotheses estimated from the normal curve' (p. 160)" (cited in Gigerenzer, 1993, p. 323). That psychologists have gotten and believe this wrong message is illustrated by Oakes's (1986) study, which found that each of three statements of inverse probability, such as "You have found the probability of the null hypothesis being true" (p. 79), were endorsed by between 36% and 86% of academic psychologists, with 96% of his sample endorsing at least one of these erroneous interpretations of a p value of .01 (pp. 80, 82). Although one can construct plausible scenarios of combinations of power, alpha levels, and prior probabilities of the hypotheses being true, where the p value turns out to be reasonably close numerically to the posterior probability of the truth of the null hypothesis given the data (Baril & Cannon, 1995), the conceptual difference cannot be stressed too strongly.

Perhaps more serious than the problem of misinterpreting p values is the problem of undisclosed decisions researchers may exercise in data collection and analysis that can greatly inflate the probability of a false positive result so that the true probability is well above the nominal alpha level of .05. Simmons, Nelson, and Simonsohn (2011) report simulations of the impact of what they call "researcher degrees of freedom" on the likelihood of false positive results. Three examples of such flexibility are: performing tests on each of several dependent variables, testing for an effect repeatedly as data collection proceeds and stopping data collection when the test is significant, and testing for an effect both with and without including a covariate in the analysis. Any one of these strategies can double the actual alpha level, for example, from .05 to .10; employing all three simultaneously can increase the actual alpha sixfold, for example, from .05 to .30. Although the assertion of Ioannidis (2005b) that "It can be proven that most claimed research findings are false" may well not apply across the board,[17] there are different lines of evidence indicating that effect sizes may be exaggerated and significance levels distorted in some research literatures. For example, estimated effect sizes in replications of 100 psychology studies were only approximately half as large on average as those reported in the original study (Open Science Collaboration, 2015). Similarly, p values that are just barely below .05 are reported much more frequently than would be expected based on the number of p values reported in other ranges (Masicampo & Lalande, 2012). Such findings have led some to suggest that we are facing a "statistical crisis in science" (Gelman & Loken, 2014), resulting in what some would judge (Leek & Peng, 2015) to be an extreme or undue scrutiny on null hypothesis significance testing or on p values. Admittedly, the pressure to achieve statistical significance in order to publish may motivate some researchers to engage in data dredging or "p-hacking" (e.g., carrying out a large number of analyses but reporting only those that are significant; Head, Holman, Lanfear, Kahn, & Jennions, 2015; Simonsohn, Nelson, & Simmons, 2014). While we will address in detail the multiple comparison problem in Chapter 5, the "curse of multiplicities" (Maxwell, 2004; Wilkinson et al., 1999) extends far beyond simply controlling for the number of contrasts tested in a

given study. One proposed solution, namely, preregistering all planned analyses, may be desirable in some confirmatory research, but likely is not a general solution to the problem—valuable insights often result from statistical analyses that are arrived at only after an iterative process that is dependent to some extent on the data (cf. Gelman & Loken, 2014). Nonetheless, some simple steps to gain control over undisclosed researcher degrees of freedom are strongly recommended. These include the requirement that the rule for terminating data collection be decided in advance of a study and reported in subsequent publications; that authors report all variables collected in a study; and that when some observations (e.g., outliers or members of certain experimental groups) are excluded from an analysis, results of analyses including those observations also be reported (Simmons et al., 2011). Even more ambitiously, attempting to replicate results before publishing would do much to improve the credibility of findings and of the profession as a whole (Gelman & Loken, 2014).

Thus, in our view, the solution to the problem of misuse and misunderstanding of *p* values is not to abandon their use, but to work hard to get things correct. The venerable methods of null hypothesis testing need not be abandoned, but they can be effectively complemented by additional methods, such as effect sizes, confidence intervals, meta-analyses, and Bayesian approaches (Howard, Maxwell, & Fleming, 2000). The future holds the promise of the emergence of use of multiple statistical methodologies, including Bayesian procedures (see Kruschke, 2015) that allow statements regarding the truth of the null hypotheses—what the id, as Gigerenzer (1993) termed it, in statistical reasoning really wants.

Toward Tests Based on Distributional Assumptions

Although this chapter may in some ways seem an aside in the development of analysis of variance and related procedures, in actuality, it is a fundamental and necessary step. First, we have shown the possibility of deriving our own significance levels empirically for particular data-analysis situations. This is a useful conceptual development to provide an analogy for what follows, in which we assume normal distribution methods. Second, and perhaps more important, the close correspondence between the results of randomization and normal theory-based tests provides a justification for using the normal theory methods. This justification applies in two important respects, each of which we discuss in turn. First, it provides a rationale for use of normal theory methods regardless of whether subjects are, in fact, randomly *sampled* from a population. Second, it is relevant to the justification of use of normal theory methods regardless of the actual shape of the distribution of the variable under investigation.

Statistical Tests With Convenience Samples

The vast majority of psychological research uses participants that can be conveniently obtained rather than actually selecting participants by way of a random sampling procedure from the population to which the experimenter hopes to generalize. Subjects may be those people at your university who were in Psychology 101 and disposed to volunteer to participate in your experiment, or they may be clients who happened to come to the clinic or hospital at the time your study was in progress, or they may be individuals who self-select to respond to an online survey. In no sense do these individuals constitute a simple random sample from the population to which you would like to generalize, for example, the population of all students or all mental health clinic clients or all adults in the United States.

If your goal is to provide normative information that could be used in classifying individuals—for example, as being in the top 15% of all college freshmen on a reading comprehension test—then a sample obtained exclusively from the local area is of little help. You have no assurance

that the local students have the same distribution of reading comprehension scores as the entire population. Although one can compute standard errors of the sample statistics and perhaps maintain that they are accurate for the hypothetical population of students for which the local students *could* be viewed as a random sample, they do not inform you of what you probably want to know—for example, how far is the local mean from the national mean, or how much error is probable in the estimate of the score on the test that would cut off the top 15% of the population of all college freshmen? Such misinterpretations by psychologists of the standard errors of statistics from non-random samples have been soundly criticized by statisticians (see Freedman, Pisani, & Purves, 2007, pp. 388, A-84). Note also that this problem is not ameliorated by increasing sample size. An online survey posted on a popular website may attract many more participants than you could recruit from a campus subject pool. But there is no assurance that a self-selected online sample, no matter the size, would be more representative of the population to which you want to generalize (cf. Wainer, 2000).

The situation is somewhat, although not entirely, different with between-group comparisons based on a convenience sample in which subjects have been randomly assigned to conditions. When groups are randomly constituted, the situation in psychology and related disciplines with regard to such "randomized controlled trials" is likely similar to that in biomedical sciences. A survey of five widely cited biomedical journals found that 96% of all controlled experiments used random assignment of a convenience sample of participants (Ludbrook & Dudley, 1998). Thus, only 4% actually used the random sampling from a population that is presumed in most presentations of statistical tests, and all of those employed inbred strains of laboratory animals. Fortunately, with groups that are not randomly sampled but are randomly assigned, a randomization test could always be carried out in this situation and is a perfectly valid approach. The *p* value yielded by such a test, as we have shown, refers to where the observed test statistic would fall in the distribution obtained by hypothetical redistributions of participants to conditions. Because the *p* value for a *t* test or *F* test is typically very close to that yielded by the randomization test, and because the randomization test results are cumbersome to compute for any but the smallest data sets,[18] one may compute the more standard *t* or *F* test and interpret the inference as applying either to possible reassignments of the currently available subjects or to an imaginary population for which these subjects might be thought to be a random sample. The generalization to a real population or to people in general that is likely of interest is then made on non-statistical grounds. Thus, behavioral scientists in general must make use of whatever theoretical knowledge they possess about the stability of the phenomena under investigation across subpopulations in order to make accurate and externally valid[19] assertions about the generality of their findings.

The Assumption of Normality

The *F* tests that are the primary focus in the following chapters assume that the population distribution of the dependent variable in each group is normal in form. In part because the dependent-variable distribution is never exactly normal, the distribution of the test statistic is only approximately correct. However, as we discuss in Chapter 3, if the only assumption violated is that the shape of the distribution of individual scores is not normal, generally, the approximation of the distribution of the test statistic to the theoretical *F* is good if the distribution of scores is not very non-normal. Not only that, but the correspondence between the *p* value yielded by an *F* test and that derived from the exact randomization test is generally very close as well. Thus, the *F* tests that follow can actually be viewed as approximations to the exact randomization tests that could be carried out. The closeness of this approximation has been demonstrated both theoretically (Wald & Wolfowitz, 1944) and by numerical examples (Kempthorne, 1952, pp. 128–132; Pitman, 1937) and simulations (e.g., Boik, 1987; Bradbury, 1987). In the eyes of some, it is this

correspondence of *F* tests to randomization tests that is a more compelling rationale for their use than the plausibility of a hypothetical infinitely large population—for example, "Tests of significance in the randomized experiment have frequently been presented by way of normal law theory, whereas their validity stems from randomization theory" (Kempthorne, 1955, p. 947). Similarly, Scheffé (1959, p. 313) notes that the *F* test "can often be regarded as a good approximation to a permutation [randomization] test, which is an exact test under a less restrictive model."

Of course, if data tend to be normally distributed, either rationale could be used. Historically, there has been considerable optimism about the pervasiveness of normal distributions, buttressed by both empirical observations of bell-shaped data patterns as well as arguments for why it is plausible that data should be approximately normally distributed.

Researchers have been noting since the early 1800s that data in the behavioral sciences are often approximately normally distributed. Although the normal curve was derived as early as 1733 by Abraham De Moivre as the limit of the binomial distribution (Stigler, 1986, pp. 70–77), it was not until the work of Laplace, Gauss, and others in the early 1800s that the more general importance of the distribution was recognized. A first step in the evolution of the normal curve from a mathematical object into an empirical generalization of natural phenomena was the comparison with distributions of errors in observations (Stigler, 1999, p. 190ff.; Stigler & Kruskal, 1999, p. 407ff.). Many of the early applications of statistics were in astronomy, and it was an astronomer, F. W. Bessel, who in 1818 published the first comparison of an empirical distribution with the normal. [Bessel is known in the history of psychology for initiating the scientific study of individual differences by developing "the personal equation" describing interastronomer differences (Boring, 1950).] From a catalog of 60,000 individual observations of stars by British Astronomer Royal James Bradley, Bessel examined in detail a group of 300 observations of the positions of a few selected stars. These data allowed an empirical check on the adequacy of the normal curve as a theory of the distribution of errors. The observations were records of Bradley's judgments of the instant when a star crossed the center line of a specially equipped telescope. The error of each observation could be assessed; Table 1.4 portrays a grouped frequency distribution of the absolute value of the errors in tenths of a second. Bessel calculated the number of errors expected to fall in each interval by using an approximation of the proportion of the normal distribution in that interval. In short, the fit was good. For example, the standard deviation for

TABLE 1.4
BESSEL'S COMPARISON OF THE DISTRIBUTION OF THE ABSOLUTE VALUES OF ERRORS WITH THE NORMAL DISTRIBUTION FOR 300 ASTRONOMICAL OBSERVATIONS

	Frequency of Errors	
Range in Seconds	*Observed*	*Estimated* (*Based on Normal Distribution*)
0.0–0.1	114	107
0.1–0.2	84	87
0.2–0.3	53	57
0.3–0.4	24	30
0.4–0.5	14	13
0.5–0.6	6	5
0.6–0.7	3	1
0.7–0.8	1	0
0.8–0.9	1	0

these data was roughly 0.2 s, and thus approximately two-thirds of the cases (i.e., 200 of the 300 observations) were expected to fall within 1 standard deviation of the mean (i.e., absolute values of errors < .2), and in fact they did (see Stigler, 1986, p. 202ff.).

Two important figures in the history of psychology played pivotal roles in changing how the normal distribution was viewed. In 1873, C. S. Peirce was apparently the first to refer to the mathematical formula as the *normal curve*, with the connotation that it describes the way errors are usually or ordinarily distributed (Stigler & Kruskal, 1999, p. 411). However, the true believer in the ubiquity of normal distributions in nature was Francis Galton, who, extending the pioneering work of the Belgian sociological statistician Adolpe Quetelet, became an advocate of the remarkable fit of the normal distribution to distributions of human abilities and characteristics. At his Anthropometric Laboratory outside London in the late 1800s, Galton amassed data showing how both physical characteristics (e.g., height) and mental characteristics (e.g., examination scores) could be fit reasonably well with a normal curve (Stigler, 1986, chaps. 5, 8). Galton's "well-known panegyric" to the normal curve suggests the devotion felt by him and others:

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known it.
>
> (Galton, 1889a, p. 66, cited in Stigler & Kruskal, 1999, p. 414)

Later psychological research also revealed many situations in which normality is reasonably approximated (although in recent years many have argued that this is the exception rather than the rule). We cite two historically important examples to illustrate the point, one from experimental psychology, the other from clinical psychology.

One of the most frequently used measures in current human experimental psychology is that of reaction time. Reaction time is used, for example, in a chronometric approach to cognitive psychology to assess the effects of manipulations such as priming (presenting a cue word immediately before a trial) on the mean time it takes to detect the presentation of a target word. Although over repeated trials a single individual's reaction time tends to follow a positively skewed distribution (more on this in a moment), it has been known for many years that *across* individuals, the distribution of individuals' average reaction time conforms very closely to the normal distribution. Figure 1.3 presents data originally reported by Fessard (1926) and cited by Woodworth and Schlosberg (1954, p. 37). Fessard measured the reaction time to sound for each of a group of 1,000 men who were applicants for jobs as machinists in Paris. Each man was measured on 30 trials, and the mean of these was used in determining the frequencies shown in the figure. A few extreme cases (35 of 1,000) were excluded by Fessard (1926, p. 218) from the table reporting his data. Although the correspondence between the data as plotted and the normal distribution is quite close, the complete data may have provided an even better fit because of the long tails of the normal distribution. Nonetheless, allowing for sampling variability, the data as presented correspond reasonably well to the theoretical normal distribution.

A second empirical example of normally distributed data in psychology is provided by scores on the Minnesota Multiphasic Personality Inventory (MMPI). Figure 1.4 shows the distribution of scores of 699 Minnesotans on the Hypochondriasis scale of the MMPI, as reported by McKinley and Hathaway (1956). The respondents, originally described in Hathaway and McKinley (1940), were individuals who were not ill, but who accompanied relatives or friends to the University of Minnesota Hospital. Again, a distribution that corresponds quite closely to a theoretical normal distribution is yielded by these test scores from "Minnesota normals."

Although the data in these two examples are perhaps more nearly normal than most, many measures of aptitude, personality, memory, and motor skill performance are often approximately

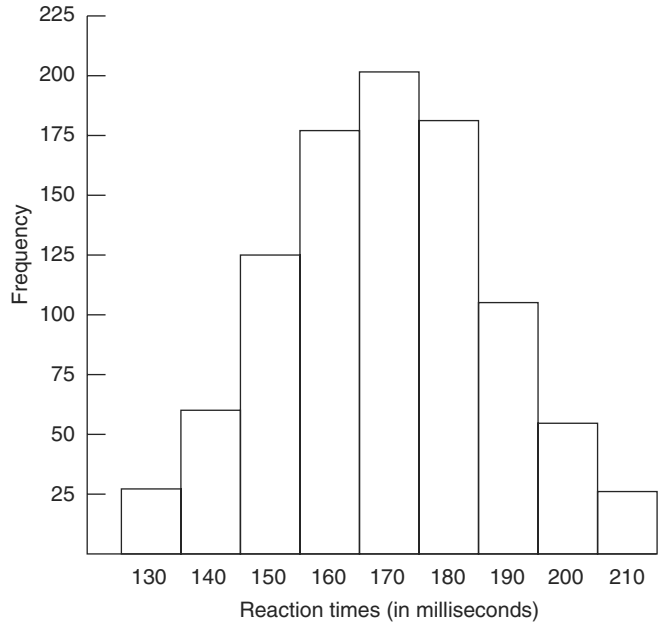


FIG. 1.3 Grouped frequency distribution of simple reaction times.

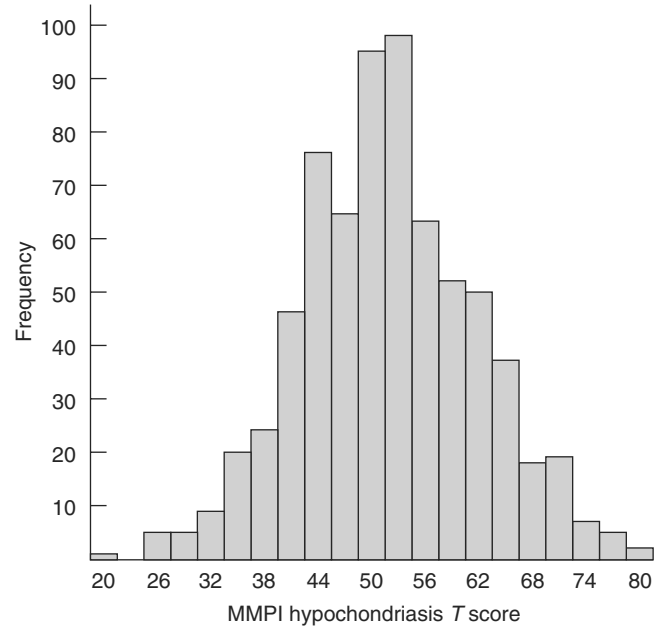


FIG. 1.4 MMPI hypochondriasis scores.

normally distributed. In part, this has to do with the global level at which constructs within the behavioral sciences are typically assessed. In a sense, the further the analysis of a phenomenon into its basic, elementary components has been carried, the less likely the data are to follow a normal distribution. Within some areas of physiological psychology, this is the case. The interest

may, for example, be simply in the occurrence or non-occurrence of a discrete event: did the neuron fire?

Perhaps the most extensively modeled non-normal, continuous processes are temporal ones. Mathematical psychologists theorize in detail about the specific non-normal form of, for instance, the distribution of simple reaction times within an individual to repeated presentations of a tone, or the distribution of interresponse times in the recordings of a single nerve fiber (see McGill, 1963). However, most areas of psychology have not progressed to having theories about the form of distributions. Nor do we have many valid binary measures of elementary processes. Instead, the dependent variable is most often a composite of a number of measures, for example, the total of the responses to 40 items on a questionnaire. Although the questionnaire may be of interest because it is thought to indicate the presence or absence of a particular psychological state such as clinical depression, the distribution of the observed variable probably is not such that it can be indicated by the frequency of two particular scores on the scale (for example, 0 and 40). Rather, its distribution is determined largely by the fact that the score on the questionnaire is the sum of the responses to 40 different items, which are far from all being perfectly correlated. Because it is not unusual for the dependent variable in a behavioral science study to be of this composite nature, a remarkable theorem can give a reasonable basis for expecting data in some situations to follow a bell-shaped curve.

This theorem, arguably the most important in statistics, is the *central limit theorem.* In its simplest form, the theorem states that the sum of a large number of independent random variables is approximately normally distributed. What is remarkable about the result is that there are almost no constraints placed on the individual distributions of the original random variables. Some could be discrete, others continuous; some could be U-shaped, some skewed, some flat; some could have large variances, some small; and still their sum would be normally distributed.

The central limit theorem can be relied on in two ways in constructing an argument for why broad classes of behavioral science data might be expected to be normally distributed[20] (Bailey, 1971, p. 199ff.). First, theory may suggest that numerous *independent factors* are the causes of a particular phenomenon. For example, for those without severe visual impairments, the keenness of an individual's vision may be viewed as the cumulative result of a series of partial causes, most of which are related to genetic background, although some environmental factors such as quality of diet or amount of eyestrain experienced might also be posited in a particular theoretical account. If these various partial causes were to occur independently in nature, be identically distributed, and summate to determine the quality of an individual's vision, then the central limit theorem tells us that the distribution of visual acuity over individuals would follow a bell-shaped distribution.

A second way in which the central limit theorem could be used to justify the expectation of a normal distribution is through conceptualizing behavioral observations for various individuals as being the result of a distribution of *errors around one true value.* This approach fits nicely with the way in which we express statistical models in Chapter 3. Instead of there being a distribution of true values across individuals as a result of specified causes, now there is assumed to be one true value around which individuals vary for unspecified reasons. To continue with another perceptual example, assume individuals are being asked to reproduce a line segment of a given length that they are shown briefly. Then, we might say that $Y_i = \tau + \varepsilon_i$, where Y_i is the measured length of the line drawn by individual i, τ is the true length of the line, and ε_i is the error term for individual i. Each of these ε_i scores may be viewed as each being a composite of a number of factors that cause the measured line length for an individual to depart from the true length. These would include both errors of measurement in recording the length of the line the subject draws and the momentary fluctuations in the individual that affect the perception of the length of the presented line and the exact length of the line the individual produces. This latter category might include the effects of slight changes in the point where the eyes are fixated at the time of

exposure, fluctuations in attention, and variations in the hosts of neural processes involved in programming a response and muscular actions required to execute it. If each of these small factors independently contributes to the composite error score for each of the individuals performing the task, then the central limit theorem shows that the composite error scores, and hence the observed *Y* scores, will be normally distributed. [This view of errors as themselves being composites, and hence approximately normally distributed according to the central limit theorem, was first conceived by Laplace in 1810, and played a major role in the development of inferential statistics (Stigler, 1986, p. 143).]

Either or both of these factors may be at work to make the data in any particular study tend toward a normal distribution. However, there are also a number of countervailing factors that may prevent this from happening. First, although it is the case that measures in the behavioral sciences are often composites or totals of numerous items, and those items are generally not perfectly correlated, they also are not independent. Indeed, psychological instruments generally are constructed so that items on a scale have at least moderate positive intercorrelations. Second, although there are situations in which individual observations are appropriately modeled as random variation around a group mean, in fact, it is probably much more common when the observations are coming from different people that they represent different true scores across people as well as random measurement error. For example, scores on a Beck Depression Inventory may be systematically different across different subgroups of participants. The random variation model may be most appropriate only when most important causal factors have been included in one's model. This is just one of many reasons for including relevant individual difference measures as predictors in one's statistical model in addition to any manipulated variables in a study (more on this in Chapter 9). Third, more mundane constraints may prevent reasonable approximations to normality, such as the fact that only a very small number of scale values are possible—say a 5-point scale is used to rate treatment outcome, or floor or ceiling effects are operating whereby a substantial proportion of participants receive either the lowest or highest value on the scale.

The point is that it is an empirical question as to whether data in any study are in fact drawn from a normally distributed population. One extensive study of large data sets in psychology argued for the conclusion that normal distributions are as hypothetical as the proverbial unicorn. Micceri (1989) examined 440 large-sample achievement and psychometric measures and found in every case that their distributions were significantly non-normal at the .01 α level. The majority of the distributions were moderately to extremely asymmetric, and most also had a greater proportion in the tails of the distribution than expected in a normal distribution. A variety of other problems such as distributions that were "lumpy" (relative frequencies not consistently increasing or decreasing) or multimodal were also noted. In short, the world certainly is not as universally normal as reading Galton might suggest.

Yet whatever the empirical and conceptual reasons or evidence for expecting data to be normally distributed, in the historical development of statistics, assuming normality made it easier to solve a number of difficult mathematical problems. This increased tractability no doubt contributed to the rise to prominence of statistical methods based on the normal distribution. For example, working independently, Gauss in 1809 showed that a particular estimation problem could be solved if errors were assumed to be normally distributed, and Laplace's central limit theorem of 1810 provided good reasons for expecting normal distributions to occur in a wide variety of situations. As Stephen Stigler tells the story in his excellent book on the history of statistics, "the remarkable circumstance that the curve that led to the simplest analysis also had such an attractive rationale was conceptually liberating" (1986, p. 145). The result was a synthesis of ideas and a development of techniques representing "one of the major success stories in the history of science" (1986, p. 158).

Although behavioral data often may not be closely approximated by the normal distribution, we have argued that normal theory-based tests are close approximations to randomization tests regardless of the shape of the distribution. Furthermore, to anticipate a related argument for the use of normal theory-based procedures that we explore in more detail at the end of Chapter 3 (when we discuss the statistical assumptions made in linear model tests), even when one is sampling from extremely non-normal distributions, such as some of those highlighted by Micceri (1989), tests assuming normality can often still perform well [e.g., when sample sizes are large and equal, and homogeneity of variance is satisfied (Sawilowsky & Blair, 1992)].

Even so, recent years have seen a profusion of so-called robust or sturdy statistical procedures, which are offered as an alternative to normal theory procedures. We consider some of these in the extension to Chapter 3 (on the website at *DesigningExperiments.com/Supplements*). However, for reasons such as those discussed regarding the reasonableness of the normal distribution assumption and the hard fact of a historical context in which normal theory-based procedures have been dominant (Huberty, 1987, 1991), statistical methods based on the general linear model assuming normally distributed data are expected to continue as some of the most important analytic methods in the behavioral sciences and many other fields. Also, although alternative methods such as robust methods are expected to continue to proliferate, one needs to understand normal theory-based methods both because they are most statistically powerful in situations in which their assumptions hold and as a point of departure for considering alternative methods when their assumptions are violated in important ways. Thus, in subsequent parts of the book, it is such normal theory-based methods that are our primary focus.

SUMMARY OF MAIN POINTS

We began with a brief discussion of philosophy of science. Although science in the Baconian tradition aspired to be purely objective, in the 20th century it was widely recognized that science, like other human endeavors, needed to rely on unproven assumptions. These include the assumptions that nature is orderly and understandable, which presumes that natural events result from a small finite number of causes, and that at least in some domains these causal mechanisms operate relatively uniformly throughout nature. Four modern schools of philosophy of science, namely the views of positivists, of Karl Popper, of Thomas Kuhn, and of realists, were briefly considered.

The second half of the chapter provided a historical context for the statistical methods we will consider in subsequent parts of the book. The critical role of randomization in experimentation was stressed, using examples and methods first introduced by Ronald Fisher early in the 20th century. The details of Fisher's exact test for discrete data and his randomization or permutation tests for continuous outcomes were presented. Debates about Fisherian statistics and the meaning and misunderstanding of *p* values were considered. The chapter concluded with a discussion of theoretical and empirical considerations that make plausible the use of normal-theory based statistics.

IMPORTANT FORMULAS

"*n* choose *r*," that is, the number of combinations of *n* things taken *r* at a time:

$$ {}_nC_r = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r(r-1)(r-2)\cdots 1} = \frac{n!}{r!(n-r)!} \tag{3} $$

Probability of r things from category R included in n things chosen out of N:

$$\Pr(r \text{ out of } R \text{ included in } n \text{ selected}) = \frac{{}_{R}C_{r} \cdot {}_{N-R}C_{n-r}}{{}_{N}C_{n}} \tag{6}$$

Odds ratio: $$\frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} \tag{7}$$

ONLINE MATERIALS AVAILABLE ON *DESIGNINGEXPERIMENTS.COM*

Data. Data sets in a variety of formats are available for download.

Computing: SAS and SPSS. We provide SPSS syntax, SAS syntax, and SPSS point-and-click directions for some of the analyses discussed in the chapter.

Computing: R. We provide detailed R syntax in a tutorial type fashion that details how to replicate essentially all analyses discussed in the chapter.

EXERCISES

Answers to exercises marked with an asterisk are available at DesigningExperiments.com/Solutions.

*1. Cite three flaws in the Baconian view that science can proceed in a purely objective manner.

2. a. Are there research areas in psychology in which the assumption of the uniformity of nature regarding experimental material is not troublesome? That is, in what kinds of research is it the case that between-subject differences are so inconsequential that they can be ignored?
   b. In other situations, although how one person responds may be drastically different from another, there are still arguments in favor of doing single-subject research. Cite an example of such a situation and suggest certain of the arguments in favor of such a strategy.

*3. Regarding the necessity of philosophical assumptions, much of 20th-century psychology was dominated by an empiricist, materialist monism, that is, the view that matter is all that exists and the only way one can come to know is by empirical observation. Some have even suggested that this position is necessitated by empirical findings. In what sense does attempting to prove materialism by way of empirical methods beg the question?

4. How might one assess the simplicity of a particular mathematical model?

5. Cite an example of what Meehl terms an *auxiliary theory* that must be relied on to carry out a test of a particular content theory of interest.

6. Explain why, in Popper's view, falsification of theories is critical for advancing science. Why are theories not rejected immediately on failure to obtain predicted results?

*7. A learning theorist asserts, "If frustration theory is correct, then partially reinforced animals will persist longer in responding during extinction than will continuously reinforced animals." What is the contrapositive of this assertion?

8. True or False: The observed value of a test statistic, and hence the observed p value, depend on the data collected in a study.

9. True or False: If a p value indicates the results of a study are highly statistically significant, the null hypothesis cannot be true.

10. True or False: Other things being equal, the smaller the p value, the stronger the evidence against the null hypothesis.

11. True or False: The p value in a randomization test can be 0.

*12. True or False: The p value associated with the observed value of a test statistic is the probability the results are due to chance.

13. True or False: A p value greater than .05 in a test of a null hypothesis means that no effect was observed, and thus that absence of an effect was shown or demonstrated.

14. True or False: Statistical significance indicates a scientifically important relation has been detected.

*15. True or False: Rejecting the null hypothesis because p was observed to be less than .05 implies that the chance you are making a wrong decision (i.e., the chance the "significant finding" is a false positive) is less than .05.

16. Assume a cognitive psychologist is planning an experiment involving brief presentations of letter strings that satisfy certain constraints. There are 14 such letter strings that satisfy the constraints, but only six of these can be used in a particular paradigm.
    a. How many combinations of six strings of letters can be chosen from the set of 14 potential strings?
    b. Given that six strings of letters have been selected, in how many different sequences could they conceivably be presented?

*17. Assume a staff member at the local state mental hospital who has been doing intake interviews for years claims that he can tell on the basis of his interviews whom the psychiatrists will judge to be sufficiently healthy to release from the hospital within the first week and whom the psychiatrists will require to stay longer than a week. As a young clinical intern at the hospital who is taken with actuarial as opposed to intuitive predictions, you are eager to prove the staff member wrong. You bet him that he will perform no differently than could be explained by chance (with alpha of .05, two-tailed) in his predictions about the next 12 patients. He agrees to the bet on the condition that you first provide him information at the end of the week about how many of the 12 patients were released so that he will know how many such patients to name. With this figure, he thinks he can determine who the released patients were, just on the basis of his earlier interview (he has no subsequent contact with the patients). To your surprise, he correctly names five of the six patients released early. Do you owe him any money? Would it have made any difference if he had named 5 of 6 early release patients out of a set of 15 intake interviews rather than 12? Support your answers.

18. A police officer in an urban police department alleges that minorities are being discriminated against in promotion decisions. The difference in promotion rates in 2014 is offered as evidence. In that year, among those eligible for promotion to the rank of sergeant, 20 officers, including 7 members of minority groups, passed an objective exam to qualify them for consideration by the review board. The number of officers that can be promoted is determined by the number of vacancies at the higher rank, and in 2014, there were 10 vacancies at the rank of sergeant that needed to be filled. Eight of the 13 non-minority officers were promoted, for a promotion rate of 61.5%, whereas only 2 of the 7 minority officers were promoted, for a promotion rate of 28.5%. If one assumes that the decisions about whom to promote were made independently of minority status, what is the probability that the discrepancy between proportions being promoted would be at least this different by chance alone, given the total number of officers under consideration and the total number of promotions possible?

    Data for the small data set presented in this exercise are also available at *DesigningExperiments.com/Data.* In later chapters, data for larger data sets will only be available online.

19. Fisher illustrated his exact test for a 2×2 table with data on criminal twins in his first paper read before the Royal Statistical Society [Fisher, R.A. (1934). The logic of inductive inference. *Journal of the Royal Statistical Society, 98*, 39–54]. The study identified 30 male criminals known to have a same-sex twin. The twin pairs were classified as monozygotic or dizygotic, and each of the 30 twin brothers of the identified criminals was then classified as to whether he was also a convicted

criminal. As shown in the following table, 10 of the 13 monozygotic criminals had brothers who had been convicted, whereas only two of 17 dizygotic criminals had brothers who had been convicted. What is the probability that so large a discrepancy in proportions would have arisen under the assumption that the difference observed is due to chance?

Convictions of Twin Brothers of Identified Criminals

	Monozygotic	*Dizygotic*	*Total*
Convicted	10	2	**12**
Not convicted	3	15	**18**
Total	**13**	**17**	**30**

Data for this exercise are also available at *DesigningExperiments.com/Data.*

20. Ioannidis (2005a) summarized results of attempted replications of clinical research studies that had originally been published in one of three major medical journals (*New England Journal of Medicine, Journal of the American Medical Association*, and *Lancet*) between 1990–2003, with each paper having been cited more than 1,000 times in the professional literature [Ioannidis (2005a). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association, 294*, 218–228]. Of the 45 original studies that had found a clinical intervention effective, 7 (16%) were contradicted by subsequent studies, 7 others (16%) reported an effect size that was at least twice as large as that found in a replication, the findings of 20 (44%) of the studies were successfully replicated, and the remaining 11 (24%) were "largely unchallenged." Ioannidis examined whether results of the replication seemed to vary depending on whether random assignment to conditions was used in the initial study. Five of the six highly cited non-randomized studies were contradicted by, or had found a larger effect size than, subsequent replications, whereas this was true of only 9 of the 39 randomized controlled trials.
    a. Carry out a Fisher's exact test to determine if the likelihood of a replication contradicting or finding a smaller effect was different in the non-randomized as opposed to the randomized studies. Compute both one-tailed and two-tailed *p* values, and compare the results of the two types of tests here.
    b. Why might non-randomized studies be less likely to be successfully replicated?

    Data for this exercise are also available at *DesigningExperiments.com/Data.*

*21. Biological changes that result from psychological manipulations, although typically not well understood, have captured attention in many areas such as health psychology. One early study examined the effects of the social environment on the anatomy of the brain in an effort to find evidence for the kinds of changes in the brain as a result of experience demanded by learning theories. The experiments are described in Bennett, Diamond, Krech, & Rosenzweig (1964). Chemical and anatomical plasticity of the brain. *Science, 146*, 610–619. Some of the raw data are presented in Freedman et al. (2007, p. 499). Pairs of male rats from a litter were used as subjects, with one member of each litter being chosen at random to be reared with other rats in an enriched environment, complete with playthings and novel areas to explore on a regular basis, whereas another member of the litter was randomly selected to be reared in isolation in a relatively deprived environment. Both groups were permitted to consume as much as they wanted of the same kinds of food and drink. After a month, the deprived environment animals were heavier and had heavier brains overall. Of critical interest, however, was the size of the cortex, or gray matter portion, of the brain in the two groups. The experiment was replicated a number of times. However, in the current exercise, we are considering the data from only one of the replications (labeled Experiment 3 in Freedman et al., 2007, p. 499). The

weights of the cortex (in milligrams) for the pairs of experimental (enriched) and control (deprived) subjects are shown in the following table:

Experiment #3	
Experimental	*Control*
690	668
701	667
685	647
751	693
647	635
647	644
720	665
718	689
718	642
696	673
658	675
680	641

Test for the effect of the treatment in this experiment by doing a randomization test. That is, perform a test of the hypothesis that the sum of the difference scores is no different than you would expect if the + and − signs had been assigned with probability .5 to the absolute values of the obtained difference scores. Although a large number of re-randomizations are possible with 12 pairs of subjects, the randomization test can be carried out here with even less computation than a *t* test by thinking a bit about the possibilities. To carry out the test, you should answer the following questions:

a. What is the observed sum of differences here?
b. How many assignments of signs to differences are possible?
c. What proportion of these would result in a sum at least as large in absolute value as that observed? To answer this question, use the following approach:
   (i) What is the largest possible positive sum that could be achieved, given the observed absolute values of the differences?
   (ii) By considering how much this largest sum would be reduced by changing one or two of the signs of the absolute differences from positive to negative, determine which assignments of signs to differences would result in sums between (or equal to) the maximal sum and the observed sum.
   (iii) Considering the symmetry of the distribution of sums resulting from re-randomizations, what is the total number of sums as extreme or more extreme, either positive or negative, as the observed sum?

Data for this exercise are also available at *DesigningExperiments.com/Data.*

*22. In 1876 Charles Darwin reported the results of a series of experiments on "The Effects of Cross- and Self-Fertilisation in the Vegetable Kingdom." The description of his experiment and the table of data for this problem are based on Fisher's discussion of "A Historical Experiment on Growth Rate" (Fisher, 1935/1971, chap. 3). The experimental method adopted by Darwin was to pit each self-fertilized plant against a cross-fertilized one under conditions that were as similar as possible for the two plants. Darwin emphasized this similarity by indicating "my crossed and self-fertilised plants . . . were of exactly the same age, were subjected from first to last to the same conditions, and were descended from the same parents" (as quoted in Fisher, 1935/1971, p. 28). One of the ways Darwin used to equalize conditions for the two members of a pair was to plant them in the same

pot. The dependent measure was the height of the plant. (Darwin did not specify when this was measured, other than to say that all plants were of the same age when their height was measured.) Although sample sizes were relatively small, Darwin indicated in his report that the experiment required 11 years to complete. To be certain that his analysis of these valuable data was correct, Darwin requested and obtained statistical consulting from his half-cousin, Francis Galton. Darwin's data and Galton's rearrangements of the data are shown in Table 1.5. Darwin's paired data are shown

TABLE 1.5
ZEA MAYS (YOUNG PLANTS)

			Arranged in Order of Magnitude				
As Recorded by Mr. Darwin			*In Separate Pots*		*In a Single Series*		
Column I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*
	Crossed	*Self-Fertilized*	*Crossed*	*Self-Fertilized*	*Crossed*	*Self-Fertilized*	*Difference*
	Inches	*Inches*	*Inches*	*Inches*	*Inches*	*Inches*	*Inches*
Pot I	$23\frac{4}{8}$	$17\frac{3}{8}$	$23\frac{4}{8}$	$20\frac{3}{8}$	$23\frac{4}{8}$	$20\frac{3}{8}$	$-3\frac{1}{8}$
	12	$20\frac{3}{8}$	21	20	$23\frac{2}{8}$	20	$-3\frac{2}{8}$
	21	20	12	$17\frac{3}{8}$	23	20	−3
					$22\frac{1}{8}$	$18\frac{5}{8}$	$-3\frac{4}{8}$
Pot II	22	20	22	20	$22\frac{1}{8}$	$18\frac{5}{8}$	$-3\frac{4}{8}$
	$19\frac{1}{8}$	$18\frac{3}{8}$	$21\frac{4}{8}$	$18\frac{5}{8}$	22	$18\frac{3}{8}$	$-3\frac{5}{8}$
	$21\frac{4}{8}$	$18\frac{5}{8}$	$19\frac{1}{8}$	$18\frac{3}{8}$	$21\frac{5}{8}$	18	$-3\frac{5}{8}$
					$21\frac{4}{8}$	18	$-3\frac{4}{8}$
Pot III	$22\frac{1}{8}$	$18\frac{5}{8}$	$23\frac{2}{8}$	$18\frac{5}{8}$	21	18	−3
	$20\frac{3}{8}$	$15\frac{2}{8}$	$22\frac{1}{8}$	18	21	$17\frac{3}{8}$	$-3\frac{5}{8}$
	$18\frac{2}{8}$	$16\frac{4}{8}$	$21\frac{5}{8}$	$16\frac{4}{8}$	$20\frac{3}{8}$	$16\frac{4}{8}$	$-3\frac{7}{8}$
	$21\frac{5}{8}$	18	$20\frac{3}{8}$	$16\frac{2}{8}$	$19\frac{1}{8}$	$16\frac{2}{8}$	$-2\frac{7}{8}$
	$23\frac{2}{8}$	$16\frac{2}{8}$	$18\frac{2}{8}$	$15\frac{2}{8}$	$18\frac{2}{8}$	$15\frac{4}{8}$	$-2\frac{6}{8}$

			Arranged in Order of Magnitude				
As Recorded by Mr. Darwin			*In Separate Pots*		*In a Single Series*		
Column I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*
	Crossed	*Self-Fertilized*	*Crossed*	*Self-Fertilized*	*Crossed*	*Self-Fertilized*	*Difference*
	Inches	*Inches*	*Inches*	*Inches*	*Inches*	*Inches*	*Inches*
					12	$15\frac{2}{8}$	$+3\frac{2}{8}$
Pot IV	21	18	23	18	12	$12\frac{6}{8}$	$+0\frac{6}{8}$
	$22\frac{1}{8}$	$12\frac{6}{8}$	$22\frac{1}{8}$	18	—	—	—
	23	$15\frac{4}{8}$	21	$15\frac{4}{8}$	—	—	—
	12	18	12	$12\frac{6}{8}$	—	—	—

in columns II and III, where the reader sees that varying numbers of pairs of plants were put in each pot. For example, there were three pairs in Pot I, five pairs in Pot III, and so on. Galton complained that the data had no "prima facie appearance of regularity." He attempted to rectify this problem by arranging the data by rank ordering according to heights, first within pots in columns IV and V, and then collapsing across pots in columns VI and VII. Galton's differences between the reordered lists are shown in column VIII.

a. Criticize Darwin's experimental design.
b. Perform appropriate analyses of these data.
   (i) Begin simply. Determine how many of the within-pair differences in heights in the original data of columns II and III favor cross-fertilization. If the cross-fertilization had no effect, how many differences would you expect on the average out of 15 to favor the cross-fertilized member of a pair? Is the observed number of differences favoring cross-fertilization significantly different from what you would expect by chance?
   (ii) Perform the simplest possible parametric statistical test appropriate for analyzing Darwin's data. How does the p value for this test compare to that in Part (i)? Why is the difference between the p values in this case in the direction it is?
   (iii) What assumptions are required for your analyses in Parts (i) and (ii)?
   (iv) One could, and Fisher in fact did, carry out a randomization test on these data. What assumptions does that test require, and what hypothesis would it test here?
c. Criticize Galton's analysis. How differently would the strength of the evidence have appeared if the data in columns VI and VII had been used for analysis rather than those in columns II and III?

Data for this exercise are also available at *DesigningExperiments.com/Data.*

23. In their article on randomization tests, Ludbrook and Dudley (1998) present hypothetical data on the impact of diet on cholesterol levels. Twelve men attending a fitness clinic agree to be randomly

assigned to a condition where they can eat fish but not meat ($n_1 = 7$) or eat meat but not fish ($n_2 = 5$). Plasma cholesterol concentrations at the end of the year-long study are shown here.

	Fish	*Meat*
	5.42	6.51
	5.86	7.56
	6.16	7.61
	6.55	7.84
	6.80	11.50
	7.00	
	7.11	
Mean	*6.414*	*8.204*

a. One way of analyzing these data would be to classify participants' cholesterol levels as being above or below the median for these data. Perform a Fisher's exact test of whether the diets resulted in different proportions of participants having cholesterol levels that were above the median cholesterol level observed.
b. A second way of analyzing these data would be by conducting a randomization test on the original cholesterol values to see if it is plausible that the observed difference in means could be attributed to chance.
   (i) How many distinct combinations of five men could have been assigned to the meat diet?
   (ii) How many of these combinations would have resulted in a greater mean difference in cholesterol levels than that actually observed? [Hint: note that the scores in the meat group include four of the five highest scores. Most (but not quite all) of the more extreme results would retain these all four highest scores and replace the one lower score with a higher score from the fish group. Similarly, a difference in the opposite direction might require all the highest scores to have been assigned to the fish group. Start with the most extreme results in either direction and make small adjustments to see if less extreme results would still exceed the observed mean difference in absolute value.]
   (iii) Alternatively, one could have analyzed these data by computer, performing a *t* test either assuming homogeneity of variance, or allowing for heterogeneity of variance. Determine the *p* values associated with such approaches.
   (iv) Which of the preceding tests would be preferred here and why?

Data for this exercise are also available at *DesigningExperiments.com/Data.*

NOTES

1. A more complete rendering of this statement in Einstein's own words is as follows:

> The very fact that the totality of our sense experiences is such that by means of thinking (operations with concepts, and the creation and use of definite functional relations between them, and the coordination of sense experiences to these concepts) it can be put in order, this fact is one which leaves us in awe, but which we shall never understand. One may say "the eternal mystery of the world is its comprehensibility." It is one of the great realizations of Immanuel Kant that the setting up of a real external world would be senseless without this comprehensibility.
>
> In speaking here concerning "comprehensibility," the expression is used in its most modest sense. It implies: the production of some sort of order among sense impressions, this order being produced by the creation of general concepts, relations between these concepts, and by relations between these concepts and sense experience, these relations being determined in any possible

> manner. It is in this sense that the world of our sense experiences is comprehensible. The fact that it is comprehensible is a miracle. (Einstein, 1936, p. 351)

2. As noted in the preface, we will be presuming some knowledge of elementary statistics in this book. If you need a review of methods such as the *t* test, see the first statistical tutorial, Review of Basic Statistics, on the website for this book at *DesigningExperiments.com/Supplements*.
3. "Agentic" connotes the capacity of human agents to "intentionally make things happen by one's actions" (Bandura, 2001, p. 2). In contrast to the positions of the radical behaviorists or eliminative materialists, agentic perspectives assert that people are initiators of action, conceiving desired ends and acting purposefully to achieve them. The term "teleological" derives from the Greek *telos*, meaning aim or purpose. Teleology figured centrally in Aristotle's analysis of causation, with the *final* cause or purpose of an action being most critical to understanding the action.
4. The causal factors could be represented by discrete or continuous variables, and by terms representing both the main effects and interactions of the factors, but the effects associated with the various variables or terms in the linear model simply will be added together to arrive at a final prediction.
5. We are for the moment setting aside considerations bearing on the validity of the statistical conclusions which will be a focus of our concern in Chapter 2 and repeatedly at other points. For now, suffice it to say that there are vagaries (cf. Schmidt, 1996) as well as clear benefits (cf. Wainer, 1999) in the binary accept–reject logic of testing statistical hypotheses. But in any event, such binary decisions fit well with the logic of Popperian falsificationism.
6. The helpful role of meta-analysis (e.g., Schmidt, 1992) in offsetting the decision-making errors in individual studies is not to be denied, and is one reason why we will be covering measures of effect size as well as statistical testing procedures throughout the current volume.
7. On the basis of a meta-analysis of surveys of anonymous admissions of having personally fabricated or falsified data and of a survey reporting instances of observed misconduct by others within an investigator's department, Stroebe et al. estimated more than 1,000 cases of research fraud "remain undetected each year in NIH-supported science alone" (2012, pp. 676, 683).
8. A discrete probability distribution is one with a countable (and typically a small finite) number of possible outcomes. An example would be the (flat) distribution of the probabilities of the six outcomes that can occur when you roll a (fair) die.
9. Although the goal was to make an inference about the population, Fisher's use of probability was in the deductive reasoning of determining the logical consequences of an exact hypothesis, and in particular the probability of occurrence of a given sample statistic. He consistently rejected the idea of attempting to determine the probability of the truth of a particular hypothesis—such attempts relied on the theory of "inverse probability," which Fisher argued was "founded upon an error, and must be wholly rejected" (1934b, p. 10). We will briefly address such Bayesian methods in a subsequent section of this chapter.
10. In attempting to formulate the probability of various outcomes, most students when faced with the tea-tasting problem begin searching their memories for a familiar discrete probability distribution. Most graduate students in the behavioral sciences have studied the binomial distribution, and so it is frequently suggested as the method of analysis. Whether it is appropriate depends again on how the experiment was run. The binomial distribution arises from a series of *independent* trials. If the subject were told there were four of each kind of cups, the successive judgments would clearly not be independent because once four cups had been classified as being of one kind, the remaining ones would have to be put into the other category to have any hope of the set of judgments being correct. If the subject were not told there were four cups of each kind, in order to make use of a binomial with probability of success equal to .5, it would be necessary to hypothesize not only that the lady had no discrimination ability but also that she had no bias for responding in favor of one cup over another. Thus, it is not clear that the binomial would be appropriate if the number of cups of each kind were determined in advance, regardless of what the subject was told. If, on the other hand, the subject understood that you determined what kind of cup each successive cup would be by the toss of a fair coin, the binomial could be used. However, in this situation, both experimenter and subject should realize that it is possible that all eight cups might be of a single kind, thus potentially allowing no comparison across kinds of cups.
11. The probability of detecting a projected effect is known as the power of a test. Although we will defer a more detailed introduction of power until Chapter 2, we make a few comments here for readers

particularly interested in the power of Fisher's exact test. Fisher's exact test may be regarded as the "uniformly most powerful among all unbiased tests for comparing two binomial populations" in a variety of situations such as where the marginals are fixed (Good, 2000, p. 99). By marginals we mean the totals in the rows and columns, which are usually written outside or in the margins of the 2×2 table. As is usually the case, one can increase the power of the test by increasing the total N and by maintaining equal numbers in the marginals under one's control, for example, the number of TF and MF cups presented. Power of Fisher's exact test against specific alternatives defined by a given odds ratio can be determined by computations based on what is called the *non-central hypergeometric distribution* (cf. Fisher, 1934a, pp. 48–51). The non-central hypergeometric can be defined in terms of the odds ratio thought to characterize the true probability of success for the two rows. For a table with fixed marginal frequencies of R, $N - R$, n, and $N - n$ as shown in the following table the probability of a particular outcome can be written in terms of the probability of the frequency in the first row and column, f_{11}, taking on a particular value, r. The power of a test can then be determined by summing the probabilities of the outcomes that are in the rejection region, with the probability of each particular outcome being computed according to the non-central hypergeometric distribution as follows:

r	$n - r$	n
$R - r$	$(N - R) - (n - r)$	$N - n$
R	$N - R$	N

$$Pr(f_{11} = r) = \frac{{}_R C_r \cdot {}_{N-R} C_{n-r} \theta^r}{\sum_{i=r_{\min}}^{r_{\max}} {}_R C_i \cdot {}_{N-R} C_{n-i} \theta^i}$$

where θ is the odds ratio of the hypothesized true probabilities of success in the two rows and $r_{\min}$ and $r_{\max}$ are the minimum and maximum possible values of r. In the case where the marginal frequencies are equal, r can range from 0 to n. When the marginal frequencies differ, the range may be restricted [specifically, $r_{\max}$ will be the smaller of R and n, and $r_{\min}$ will be the larger of 0 and $n - (N - r)$].

To illustrate the computation of power, if in the tea-tasting example the probability of classifying a cup as tea first were really .9 for tea-first cups and .3 for milk-first cups, then we would have an odds ratio of:

$$\theta = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} = \frac{.9 / (1 - .9)}{.3 / (1 - .3)} = \frac{9/1}{3/7} = 21$$

Because with four cups of each kind the only persuasive evidence of discrimination ability would be if all cups were correctly classified, the power of such a test could be computed as:

$$\begin{aligned} Pr(f_{11} = r) &= \frac{{}_4 C_4 \cdot {}_4 C_0 \theta^4}{\sum_{i=0}^{4} {}_4 C_i \cdot {}_4 C_{4-i} \theta^i} \\ &= \frac{1 \cdot 1 \cdot 21^4}{\sum_{i=0}^{4} {}_4 C_i \cdot {}_4 C_{4-i} 21^i} \\ &= \frac{194{,}481}{1 + 4 \cdot 4 \cdot 21 + 6 \cdot 6 \cdot 21^2 + 4 \cdot 4 \cdot 21^3 + 1 \cdot 1 \cdot 21^4} \\ &= \frac{194{,}481}{1 + 336 + 15{,}876 + 148{,}176 + 194{,}481} \\ &= \frac{194{,}481}{358{,}870} = .542 \end{aligned}$$

If more than one outcome could have led to rejection of the null hypothesis, the probabilities of those values of f_{11} occurring would be computed similarly and cumulated to determine the power.

A helpful discussion of Fisher's exact test with references to relevant literature is given in Good (2000, chap. 6). Alternative methods of estimating power illustrated with numerical examples are provided by Cohen (1977), O'Brien and Muller (1993), and Rosner (1995, p. 384ff.).

Readers wishing to determine power should be aware, as noted by O'Brien (1998), of the large number of different methods for carrying out computations of p values and power for the case of data from a 2×2 design. One common situation, different from the current case, is that in which one carries out analyses under the assumption that one is comparing two independent proportions, such as the proportion of successes in each of the two rows of the table. In contrast to situations such as the present one where the subject is constrained to produce equal numbers of responses in the two classifications, in many experimental situations the total number of responses of a given kind is not constrained. The appropriate power analysis can be considerably different under such an assumption.

Some authors following Berkson (1978) reject the idea of doing Fisher's exact tests, arguing that the marginal totals provide relevant information. While this is sometimes certainly the case, it is not always true (cf. Kempthorne, 1979), such as when both row and column marginals are constrained by the experimenter as in Fisher's tea-tasting example. What can be confusing is that some power analysis programs such as Power and Precision (www.power-analysis.com) provide an option labeled "Power Computation: Fisher's Exact Test," but in fact compute power assuming the two rows reflect two independent binomial distributions. For example, with data as in the tea-tasting example, the power to detect a difference between population proportions of .9 and .3 is determined by the Power and Precision program as:

$$\text{power} = \left({}_4C_4 \cdot .9^4 \cdot .1^0\right) \cdot \left({}_4C_0 \cdot .3^0 \cdot .7^4\right) = (.6561)(.2401) = .1575$$

clearly different from the value of .542 computed with the non-central hypergeometric appropriate for the case where the subject was required to give four responses of each kind. Other programs (e.g., UnifyPow; see O'Brien, 1998) use approximations to the conditional probability of the non-central hypergeometric appropriate for Fisher's exact, but with the warning that the approximation may not be very accurate with small sample sizes (cf. O'Brien & Muller, 1993, p. 336).

12. That is, a histogram showing the relative frequency of scores would be low in the middle range and high at either end; hence the distribution looks somewhat like a "U." In the current data, there are more scores below 0 and more scores greater than 8 than there are between 0 and 8.
13. The t test statistic would exactly follow a t distribution if one were randomly sampling difference scores from a normally distributed population. Ignoring for the moment that psychologists almost never draw a random sample from the population to which they would like to generalize, we can be confident here that even if the 10 twin pairs were a random sample that they would be coming from a population that was not exactly normally distributed. This is the case not just because the sample has a U-shaped distribution, which could be atypical, but because the dependent variable can here only take on certain discrete values, whereas the normal distribution is continuous. Whether the non-normality is sufficient to seriously affect the robustness of the test is a separate question from the point we are trying to make here.
14. The corrected value of t is slightly smaller, 2.046, and in fact is exceeded by .038 of the t distribution.
15. Exceptions to this rule are possible. As noted by Ludbrook and Dudley (1998), clear discrepancies between the p values yielded by randomization tests and those yielded by parametric tests can occur when sample sizes are small, particularly in multiple-group studies where the form of distributions varies across groups. An example suggested by Gerber and Green (2012) involved hypothetical data on charitable contributions, in a situation where 10 out of 20 participants had been randomly assigned to a treatment encouraging them to donate. A single very large contribution ($500) in the treatment condition resulted not only in a larger mean contribution ($80) in the treatment condition than that in the control condition ($10), but also in a drastically larger variance (more than 300 times larger than that in the control condition). A one-tailed randomization comparing the mean contributions yielded a p value of .032 as compared to a one-tailed p value of .082 from a t test (or $p = .091$ in a t test allowing for heterogeneous variances). In such a case, because of the extreme violation of the assumption of homogeneity of variance, the randomization test would be preferred. Exercise 23 at the end of the chapter explores another similar example.

16. As of the current writing, a very helpful overview of randomization or permutation tests is provided by David Howell (www.uvm.edu/~dhowell/StatPages).
17. Ioannidis considered different scenarios regarding the prior probability of the truth of the alternative hypothesis and for each derived the poststudy probability that the alternative hypothesis was true. In each of the scenarios where the poststudy probability that the alternative hypothesis was true turned out to be less than .5, the presumed prior probability of the truth of the alternative hypothesis was less than .5, and sometimes much less (Ioannidis, 2005b, Table 4). For example, one scenario Ioannidis considered would be applicable to a case where researchers are attempting to find a link between individual genes and risk of a disease, and where 30,000 genes might be tested, of which only 30 are causally linked to risk for the disease. Not surprisingly, if the prior odds are 1,000 to 1 that an alternative hypothesis is false (i.e., that the null hypothesis is true), evidence from a single study suggesting the contrary may be spurious. On the other hand, in most behavioral research, just the opposite may be true. That is, the prior probability that the null hypothesis is exactly true is practically zero (Cohen, 1994; Jones & Tukey, 2000), and conversely the alternative hypothesis, rather than being wildly implausible a priori, is in fact often reasonably viewed as being more likely true than false.
18. Just how cumbersome may be surprising. For example, if a total of 30 observations are to be assigned in equal numbers to the groups in a study, with two groups over 150 million assignments are possible, and with three groups over 3 trillion assignments are possible. Although the number of calculations for a complete specification of the distribution of a test statistic is clearly prohibitive in general, interest in randomization methods is increasing because of recent developments making such tests more practical. These developments include the design of computational algorithms and computer programs that take random samples from the distribution (for example, Edgington, 1995, pp. 50–51, 68ff.; Green, 1977; or the R routines mentioned previously), algebraic simplifications (Gabriel & Hall, 1983), and approximations (Gabriel & Hsu, 1983). Although the logic of randomization testing is important for gaining a fundamental understanding of *p* values, the specific procedures for more complex designs are not considered in subsequent chapters.
19. External and other types of validity will be discussed in Chapter 2.
20. Students in the behavioral sciences are often familiar with the central limit theorem for explaining why group means can be expected to be normally distributed. However, here we are considering the application of the theorem in the way conceived by its originator, Laplace (Stigler, 1986, p. 143), and that is to view an individual observation or even the error in an individual observation as a composite or summation of the effects of a number of variables.

2

Drawing Valid Inferences From Experiments

OVERVIEW OF CHAPTER: RESEARCH QUESTIONS ADDRESSED

We now focus on where attempts to draw inferences from experiments can flounder. We distinguish four types of validity: statistical conclusion validity, internal validity, construct validity, and external validity. After examining threats to each type of validity, the chapter concludes by giving an overview of the designs and analyses considered in the subsequent chapters of the book.

Thus, the sorts of research questions addressed in this chapter are:

- How can things go wrong in attempting to draw an inference from an experiment?
- What are the major types of experimental designs that will be considered in this book?

PUBLISHED EXAMPLE

West and Thoemmes (2010), in an article titled "Campbell's and Rubin's Perspectives on Causal Inference," provide a helpful summary and comparison of two of the more important contributions of the past half century to experimental design and analysis. Donald Campbell's analysis of threats to the validity of inferences (which we discuss in this chapter) has primarily been applied in psychology and education, whereas Donald Rubin's analysis of causal effects (which we discuss briefly in Chapter 9) has primarily been applied in public health and medicine. Campbell's approach helps the working scientist apply a synthesis of methodological insights, particularly within the behavioral sciences, to identify plausible threats and then incorporate design elements into a study to try to rule out those threats. For example, in an observational study of two teaching methods (e.g., online instruction vs. standard classroom instruction) used for a year of public school instruction in mathematics, the inference that the different teaching methods caused the difference in achievement at the end of the school year would plausibly be threatened if the methods were applied in two different intact groups of students. If online instruction were used with gifted students and standard classroom instruction with non-gifted students, the difference on the posttest may be due not to treatments but to differences in achievement at the beginning

of the school year. Incorporating a pretest (i.e., baseline measure) into the design and examining the *change* in achievement over the year would help somewhat. However, the gifted students might be expected to show greater gains than average students regardless of the teaching method (termed a Selection × Maturation interaction). If one had to compare the teaching methods with such different types of students, stronger evidence of a treatment effect would be provided if one had data allowing an estimation of the maturational trend across prior years of education for the two groups, which would allow one to assess if those trends were altered in the year that the two treatments were being compared. Campbell's approach thus emphasizes altering the design based on a qualitative analysis of identified threats.

Rubin's approach, in contrast, is more mathematical and focuses on precise assumptions underlying a formal model (the potential outcomes model). In observational studies or in "broken" randomized experiments (e.g., where not all individuals experience the assigned treatment or where there is differential attrition across conditions), Rubin's method involves matching participants in the treatment and control conditions on a composite index derived from a large number of predictors of group membership. Large-scale studies are typically required to have sufficient resources to measure the relevant covariates and to allow for selecting matched participants. As one example, West and Thoemmes cite a study of the effects of students being retained in first grade on their subsequent growth in math and reading. Rubin's method involved attempting to find, for each retained student, one of the normally promoted students who was a close match on a composite index (known as the "propensity score") of 72 covariates that together predicted membership in the retained or promoted group. Nonmatched students were then excluded from a statistical analysis, which suggested that, in fact, retention impaired subsequent growth in math and reading. Rubin's model relies on several assumptions, but if they are satisfied, the approach yields an unbiased estimate of the treatment effect.

THREATS TO THE VALIDITY OF INFERENCES FROM EXPERIMENTS

Having reviewed the perils of drawing inductive inferences at a philosophical level and having introduced Fisher's efforts to support inferences based on statistical theory, we now turn to a consideration of threats to the validity of inferences at a more practical level. The classic treatment of the topic of how things can go wrong in attempting to make inferences from experiments was provided in the monograph by Campbell and Stanley (1963). Generations of graduate students around the country memorized their "threats to validity." An updated and expanded version of their volume addressing many of the same issues, but also covering the details of certain statistical procedures, appeared 16 years later authored by Cook and Campbell (1979). More recently, the third instantiation of a volume on quasi-experimentation co-authored by Donald Campbell appeared (Shadish et al., 2002), which Campbell worked on until his death in 1996. Judd and Kenny (1981) and Krathwohl (1985) have provided very useful and readable discussions of these validity notions of Campbell and his colleagues. Cronbach's (1982) book also provides a wealth of insights into problems of making valid inferences, but like Cook and Campbell (1979), it presumes a considerable amount of knowledge on the part of the reader. (For a brief summary of the various validity typologies, see Mark, 1986).

For our part, we begin the consideration of the practical problems of drawing valid inferences by distinguishing among the principal types of validity discussed in this literature. Then, we suggest a way for thinking in general about threats to validity and for attempting to avoid such pitfalls.

Types of Validity

When a clinician reads an article in a journal about a test of a new procedure and then contemplates applying it in his or her own practice, a whole series of logical steps must all be correct for this to be an appropriate application of the finding. [Krathwohl (1985) offers the apt analogy of links in a chain for these steps.] In short, a problem could arise because the conclusion or design of the initial study was flawed or because the extrapolation to a new situation is inappropriate. Campbell and Stanley (1963) referred to these potential problems as threats to internal and external validity, respectively. Cook and Campbell (1979) subsequently suggested that, actually, four types should be distinguished: statistical conclusion validity, internal validity, construct validity, and external validity. Shadish et al. (2002) suggested further refinements but maintained this fourfold validity typology. We discuss each in turn, but first a word or two by way of general introduction.

Validity means essentially truth or correctness, a correspondence between a proposition describing how things work in the world and how they really work (see Russell, 1919b; Campbell, 1986, p. 73). Naturally, we never know with certainty if our interpretations are valid, but we try to proceed with the design and analysis of our research in such a way to make the case for our conclusions as plausible and compelling as possible.

The propositions or interpretations that abound in the discussion and conclusion sections of behavioral science articles are about how things work in general. As Shadish et al. (2002) quip, "Most experiments are highly local but have general aspirations" (p. 18). Typical or modal experiments involve particular people manifesting the effects of particular treatments on particular measures at a particular time and place. Modal conclusions involve few, if any, of these particulars. Most pervasively, the people (or patients, children, rats, classes, or most generally, units of analysis) are viewed as a sample from a larger population of interest. The conclusions are about the population. The venerable tradition of hypothesis testing is built on this foundational assumption: one unit of analysis differs from another. The variability among units, however, provides the yardstick for making the statistical judgment of whether a difference in group means is "real."

What writers such as Campbell stressed is that not just the units or subjects, but also the other components of our experiments should be viewed as representative of larger domains, in somewhat the same way that a random sample of subjects is representative of a population. Specifically, Cronbach (1982) suggested that there are four building blocks to an experiment: units, treatments, observations or measures, and settings. We typically want to generalize along all four dimensions, to a larger domain of units, treatments, observations, and settings, or as Cronbach puts it, we study "utos" but want to draw conclusions about "UTOS." For example, considering the dimension of treatments, a specific multifaceted treatment program (t) for problem drinkers could have involved the same facets with different emphases (e.g., more or less time with the therapist) or different facets not represented initially (e.g., counseling for family members and close friends) and yet still be regarded as illustrating the theoretical class of treatments of interest, controlled drinking (T). (In Chapter 10, we discuss statistical procedures that assume the treatments in a study are merely representative of other treatments of that type that could have been used, but more often the problem of generalization is viewed as a logical or conceptual problem instead of a statistical problem.)

Turning now to the third component of experiments—namely the observations or measures—it is perhaps easier because of the familiarity of the concepts of "measurement error" and "validity of tests" to think of the measures instead of the treatments used in experiments as fallible representatives of a domain. Anyone who has worked on a large-scale clinical research project has probably been impressed by the number of alternative measures available for assessing the

various psychological traits or states of interest in that study. Finally, regarding the component of the setting in which experiments take place, our comments in Chapter 1 about the uniformity of nature underscore what every historian or traveler knows but that writers of discussion sections sometimes ignore: what is true about behavior for one time and place may not be universally true. In sum, an idea to remember as you read about the various types of validity is how they relate to the question of whether a component of a study—such as the units, treatments, measures, or setting—truly reflects the domain of theoretical interest.

Statistical Conclusion Validity

The question to be answered in statistical conclusion validity is, "Was the original statistical inference correct?" That is, did the investigators reach a correct conclusion about whether a relationship between the variables exists in the population or about the extent of the relationship? Thus, statistical conclusions are about population parameters—such as means or correlations—whether they are equal or what their numerical values are. So in considering statistical conclusion validity, we are not concerned with whether there is a causal relationship between the variables, but whether there is any relationship, be it causal or not.

One of the ways in which a study might be an insecure base from which to extrapolate is that the conclusion reached by that study about a statistical hypothesis it tested might be wrong. As you likely learned in your first course in statistics, there are two types of errors or ways in which this can happen: Type I errors, or false positives—that is, concluding there is a relationship between two variables when, in fact, there is none—and Type II errors, or false negatives—that is, failing to detect a relationship that in fact exists in the population. One can think of Type I errors as being gullible or overeager, whereas Type II errors can be thought of as being blind or overly cautious (Rosnow & Rosenthal, 1989). Because the nominal alpha level or probability of a Type I error is fairly well established by convention within a discipline—for example, at .05 – the critical issue in statistical conclusion validity is power. The statistical power of a test is its sensitivity or ability to detect relationships that exist in the population, and so it is the complement of a Type II error. As such, *power* in a statistical sense means sensitivity or ability to detect, based on a study, what is present in the population. Studies with low power are like "trying to read small type in dim light" (Rosnow & Rosenthal, 1989). In conventional terms, power is the probability of rejecting the null hypothesis when it is false and equals 1 minus the probability of a Type II error.

The threats to the validity of statistical conclusions are then of two general kinds: a liberal bias, or a tendency to be overly optimistic about the presence of a relationship or exaggerate its strength; and a conservative bias, or a tendency to be overly pessimistic about the existence of a relationship or underestimate its strength.

As Cohen (1988) stressed, one of the most pervasive threats to the validity of the statistical conclusions reached in the behavioral sciences is low power. It is critical in planning experiments and evaluating results to consider the likelihood that a given design and sample size would detect an effect of a given size in the population. As discussed in detail beginning in Chapter 3, there are a variety of ways to estimate how strong the relationship is between the independent variable and the dependent variable, and using this, to compute a numerical value of the power of a study. Our concern here, however, is with why statistical conclusions are often incorrect; several reasons can be enumerated.

Studies typically have low power because sample sizes used are too small for the situation. Because the number required depends on the specifics of the research problem, one cannot specify in general a minimum number of subjects to have per condition. However, although other

steps can be taken, increasing the number of participants is the simplest solution, conceptually at least, to the problem of low power.

Another important reason for low power is the use of an unreliable dependent variable. Reliability, of course, has to do with consistency and accuracy in the sense of low error of measurement. Scores on variables are assumed to be the result of a combination of systematic or true score variation and random error variation. For example, your score on a multiple-choice quiz is determined in part by what you know and in part by other factors, such as your motivation and your luck in guessing answers you do not know. Variables are unreliable, in a psychometric sense, when the random error variation component is large relative to the true score variation component (see Judd & Kenny, 1981, p. 111ff., for a clear introduction to the idea of reliability).

We acknowledge, as Nicewander and Price (1983) point out, that there are cases in which the less reliable of two possible dependent variables can lead to greater power, for example, because a larger treatment effect on that variable may more than offset its lower reliability. However, other things being equal, the lower the reliability of a dependent measure is, the less sensitive it will be in detecting treatment effects. Solving problems of unreliability is not easy, in part because there is always the possibility that altering a test in an attempt to make it more reliable might change what it is measuring as well as its precision of measurement. However, the rule of thumb, as every standard psychometrics text makes clear (e.g., Nunnally, 1978; see Maxwell, 1994), is that increasing the length of tests increases their reliability. Thus, the longer the quiz, the less likely you can pass simply by guessing.

Other reasons why unexplained variability in the dependent variable and hence the probability of a Type II error may be unacceptably high include implementing the treatment in slightly different ways from one subject to the next and failure to include important explanatory variables in your model of performance for the situation. Typically, in behavioral science studies, who the participant happens to be is a more important determinant of how he or she performs on the experimental task than the treatment to which the person is assigned. Thus, including a measure of the relevant individual differences among participants in your statistical model, or experimentally controlling for such differences, can often greatly increase your power. (Chapters 9 and 11–15 discuss methods for dealing with such individual differences.)

Maxwell, Cole, Arvey, and Salas (1991) provide a helpful discussion of these issues, comparing alternative methods of increasing power. In particular, they focus on the relative benefits of lengthening the posttest and including a pretest in a design. These are complementary strategies for reducing unexplained variability in the dependent variable. When the dependent measure is of only moderate or low reliability, as may be the case with a locally developed assessment, greater gains in power are realized by using a longer and hence more reliable posttest. When the dependent measure has high reliability, then including a pretest that can be used to control for individual differences among subjects will increase power more.

The primary cause of Type I error rates being inflated over the nominal or stated level is that the investigator has performed multiple tests of the same general hypothesis. Statistical methods exist for adjusting for the number of tests you are performing and are considered at various points in this text (see, for example, Chapter 5 on multiple comparisons). Violations of statistical assumptions can also affect Type I and Type II error rates. As we discuss at the end of Chapter 3, violating assumptions can result in either liberal or conservative biases. Finally, sample estimates of how large an effect is, or how much variability in the dependent variable is accounted for, tend to overestimate population values. Appropriate adjustments are available and are covered in Chapters 3 and 7. A summary of these threats to statistical conclusion validity and possible remedies is presented in Table 2.1.

TABLE 2.1
THREATS TO STATISTICAL CONCLUSIONS AND SOME REMEDIES

Threats Causing Overly Conservative Bias	*Remedies and References*	
Low power as a result of small sample size	Increase sample size	Chapter 3 ff.; Cohen, 1988
Low power due to increased error because of unreliability of measures	Improve measurement (e.g., by lengthening tests)	Chapter 9; Maxwell, 1994; Maxwell et al., 1991
Low power as a result of high variability because of diversity of subjects	Control for individual differences: In analysis by controlling for covariates In design by blocking, matching, or using repeated measures	Chapters 9 and 11 ff.; Maxwell, Delaney, and Dill, 1984
Low power due to violation of statistical assumptions	Transform data or use different method of analysis	Chapter 3; McClelland, 2000
Threats Causing Overly Liberal Bias		
Repeated statistical tests	Use adjusted test procedures	Chapter 5
Violation of statistical assumptions	Transform data or use different method of analysis	Chapter 3
Biased estimates of effects	Use corrected values to estimate effects in population	Chapter 3ff.

Internal Validity

Statistical tests allow one to make conclusions about whether the mean of the dependent variable (typically referred to as variable *Y*) is the same in different treatment populations. If the statistical conclusion is that the means are different, one can then move to the question of what caused the difference, with one of the candidates being the independent variable (call it variable *X*) as it was implemented in the study. The issue of internal validity is, "Is there a causal relationship between variable *X* and variable *Y*, regardless of what *X* and *Y* are theoretically supposed to represent?" If variable *X* is a *true* independent variable and the statistical conclusion is valid, then internal validity is to a large extent assured (appropriate caveats follow). By a *true independent variable*, we mean one for which the experimenter can and does independently determine the level of the variable that each participant experiences—that is, assignment to conditions is carried out independently of any other characteristic of the participant or of other variables under investigation. Internal validity is, however, a serious issue in quasi-experimental designs in which this condition is not met. Most commonly, the problem is using intact or self-selected groups of subjects. For example, in an educational psychology study, one might select the fifth-grade class in one school to receive an experimental curriculum and use the fifth-grade class from another school as a control group. Any differences observed on a common posttest might be attributed to preexisting differences between students in the two schools rather than the educational treatment. This threat to internal validity is termed *selection bias* because subjects receiving different treatments were selected from different intact groups.[1] A selection bias is an example of the more general problem of a *confound*, defined as an extraneous variable that is correlated with, or whose levels are literally "found with," the levels of the variable of interest. Perhaps less obvious is the case in which an attribute of the subjects is investigated as one of the factors in an experiment. Assume that depressed and non-depressed groups of subjects were formed by scores on an instrument such as the Beck Depression Inventory; then, it is observed that the depressed group performs significantly worse on a memory task. One might like to claim

that the difference in memory performance was the result of the difference in level of depression; however, one encounters the same logical difficulty here as in the study with intact classrooms. Depressed subjects may differ from non-depressed subjects in many ways besides depression that are relevant to performance on the memory task.

Internal validity threats are typically thus "third" variable problems. Another variable besides *X* and *Y* may be responsible for either an apparent relationship or an apparent lack of a relationship between *X* and *Y*.

A number of other threats to internal validity arise when subjects are assessed repeatedly over time,[2] or participate in what is called a *longitudinal* or *repeated measures design*. The most intractable difficulties in making a causal inference here arise when there is just a single group whose performance is being monitored over time, in what Campbell has referred to as a *one-group pretest-posttest design*, denoted O_1 X O_2 to indicate a treatment intervenes between two assessments (observations). One of the most common threats to internal validity is *attrition*, or the problem that arises when possibly different types of people drop out of various conditions of a study or have missing data for one or more time periods. The threats to validity caused by missing data are almost always a concern in longitudinal designs. Chapter 15 presents methodology especially useful in the face of missing data in such designs. *Cross-sectional designs* or designs that involve only one assessment of each subject can often avoid problems of missing data, especially in laboratory settings. However, the internal validity of even cross-sectional designs can be threatened by missing data, particularly in field settings, for example, if a subject fails to show up for his or her assigned treatment or refuses to participate in the particular treatment or measurement procedure assigned. Attempts to control statistically for variables on which participants are known to differ can be carried out, but face interpretational difficulties, as we discuss in Chapter 9. West and Sagarin (2000) present a very readable account of possible solutions for handling missing data in randomized experiments, including subject losses that arise from noncompliance as well as attrition.

Other threats arising in longitudinal designs include *testing*. This threatens internal validity when a measurement itself might bring about a change in performance, such as when assessing the severity of participants' drinking problem affects their subsequent behavior. Such measures are said to be *reactive*. *Regression* is a particular problem in remediation programs in which subjects may be selected based on their low scores on some variable and then naturally move toward the mean for statistical reasons rather than because of the treatment. *History* threatens the attribution of changes to the treatment when events outside the experimental setting occur that might cause a change in subjects' performance. *Maturation* refers to changes that are not caused by some external event, but by processes such as fatigue, growth, or natural recovery. So, when only one group experiences the treatment, the appropriate attribution may be that "time heals." Thus, the potential remedy for these last four artifacts shown in Table 2.2 that are characteristic of one-group longitudinal designs is the addition of a similarly selected and measured but randomly assigned group of control participants who do not experience the treatment.

Estimating the internal validity of a study is largely a thought problem in which you attempt to systematically think through the plausibility of various threats relevant to your situation.[3] On occasion, one can anticipate a given threat and gather information in the course of a study relevant to it. For example, questionnaires or other attempts to measure the exact nature of the treatment and control conditions experienced by subjects as well as possible other experiences besides those manipulated in the study may be useful in determining whether extra-experimental factors differentially affected subjects in different conditions.

Finally, a term from Campbell (1986) is useful for distinguishing internal validity from the other types remaining to be considered. Campbell suggests it might be clearer to call internal validity "local molar (pragmatic, atheoretical) causal validity" (p. 69). Although a complex

TABLE 2.2
THREATS TO INTERNAL VALIDITY

Threats	*Definition*
Selection bias	Participant characteristics confounded with treatment conditions because of use of intact or self-selected participants; or more generally, whenever predictor variables represent measured characteristics as opposed to independently manipulated treatments.
Attrition	Differential drop out across conditions at one or more time points that may be responsible for differences.
Testing	Altered performance as a result of a prior measure or assessment instead of the assigned treatment.
Regression	The changes over time expected in the performance of subjects, selected because of their extreme scores on a variable, that occur for statistical reasons but might incorrectly be attributed to the intervening treatment.
Maturation	Observed changes as a result of ongoing, naturally occurring processes rather than treatment effects.
History	Events, in addition to an assigned treatment, to which subjects are exposed between repeated measurements that could influence their performance.

phrase, this focuses attention on points deserving of emphasis. The concern of internal validity is causal in that you are asking what was responsible for the change in the dependent variable. The view of causes is molar—that is, at the level of a treatment package, or viewing the treatment condition as a complex hodgepodge of all that went on in that part of the study—thus emphasizing that the question is *not* what the "active ingredient" of the treatment is. Rather, the concern is pragmatic, atheoretical—did the treatment, for whatever reason, cause a change, did it work? Finally, the concern is local: Did it work here? With internal validity, one is not concerned with generalization.

Construct Validity

The issue regarding construct validity is, "Given there is a valid causal relationship, is the interpretation of the constructs involved in that relationship correct?"[4] Construct validity pertains to both causes and effects. That is, the question for both the independent and dependent variables as implemented in the study is, "Can I generalize from this one set of operations to a referent construct?" What one investigator labels as construct *A* causing a change in construct *C*, another may interpret as an effect of construct *B* on construct *C*, or of construct *A* on construct *D* or even of *B* on *D*. Showing a person photographs of a dying person may arouse what one investigator interprets as death anxiety and another interprets as compassion. Threats to construct validity are a pervasive and difficult problem in psychological research. We addressed this issue implicitly in Chapter 1 in commenting on the meaning of theoretical terms. Since Cronbach and Meehl's (1955) seminal paper on construct validity in the area of assessment, something approaching a general consensus has been achieved that the specification of constructs in psychology is limited by the richness, generality, and precision of our theories. Given the current state of psychological theorizing, it is understandable why a minority continue to argue for strategies such as adopting a strict operationalism or attempting to avoid theorizing altogether. However, the potential for greater explanatory power offered by theoretical constructs places most investigators in the position of having to meet the problem of construct validity head-on rather than sidestepping it by abandoning theoretical constructs.

The basic problem in construct validity is the possibility "that the operations which are meant to represent a particular cause or effect construct can be construed in terms of more than one

construct, each of which is stated at the same level of reduction" (Cook & Campbell, 1979, p. 59). The qualifier regarding the level of reduction refers to the fact that alternative explanations of a phenomenon can be made at different levels of analysis, and that sort of multiplicity of explanation does not threaten construct validity. This is most clearly true across disciplines. One's support for a political position could be explained at either a sociological level or by invoking a psychological analysis, for example, of attitude formation. Similarly, showing there is a physiological correlate of some behavior does not mean the behavioral phenomenon is to be understood as nothing but the outworking of physiological causes.

Some examples of specific types of artifacts serve to illustrate the confounding that can threaten construct validity. A prime example of a threat to construct validity is the experimenter bias effect demonstrated by Rosenthal (1976). This effect involves the impact of the researcher's expectancies and, in particular, the transmission of that expectancy to the subject in such a way that performance on the dependent variable is affected. Thus, when the experimenter is not blind to the hypothesis under investigation, the role of experimenter bias must be considered, as well as the nominal treatment variable, in helping to determine the magnitude of the differences between groups. This is a rationale for the double-blind experiment, where not only does the subject not know the group he or she is in but neither do those collecting the data.

Another set of threats to construct validity arises in situations in which there are clear, unintended by-products of the treatment as implemented that involve causal elements that were not part of the intended structure of the treatment (cf. Shadish et al., 2002, p. 95). One example is *treatment diffusion*, which can occur when there is the possibility of communication during the course of a study among subjects from different treatment conditions. Thus, the mixture of effects of portions of different treatments that subjects functionally receive, filtered through their talkative friends, can be quite different from the single treatment they were nominally supposed to receive. This type of threat can be a particularly serious problem in long-term studies such as those comparing alternative treatment programs for clinical populations. Such treatment diffusion is more of an issue in psychological and educational settings where participants are typically aware, for example, of the cognitive strategies they are supposed to be practicing, than is the case in pharmaceutical or biomedical research where participants more often can be blind regarding the drug or other treatment they are receiving. Another such threat is termed *resentful demoralization.* For example, a waiting-list control group may be demoralized by learning that others are receiving effective treatments while they are receiving nothing, or at least a less preferred treatment. Furthermore, in a variety of other areas of psychology in which studies tend to involve brief treatment interventions but in which different people may participate over the course of an academic semester, the essence of a treatment can be affected greatly by dissemination of information over time. Students who learn from previous participants the nature of the deception involved in the critical condition of a social psychology study may experience a considerably different condition than naive subjects would experience. These participants may well perform differently than participants in other conditions, but the cause may have more to do with the possibly distorted information they received from their peers than the nominal treatment to which they were assigned.

Two major pitfalls to avoid in one's attempt to minimize threats to construct validity can be cited: *inadequate preoperational explication* of the construct and *mono-operation bias* or using only one set of operations to implement the construct (Cook & Campbell, 1979, p. 64ff.; Shadish et al., 2002, p. 73ff.). First, regarding explication, the question is, "What are the essential features of the construct for your theoretical purposes?" For example, if you wish to study social support, does your conceptual definition include the perceptions and feelings of the recipient of the support or simply the actions of the provider of the support? Explicating a construct involves consideration not only of the construct you want to assess, but also the other similar

constructs from which you hope to distinguish your construct (see Campbell & Fiske, 1959; Judd & Kenny, 1981). Second, regarding mono-operation bias, using only a single dependent variable to assess a psychological construct typically runs the risk of both underrepresenting the construct and containing irrelevancies. For example, anxiety is typically regarded as a multidimensional construct subsuming behavioral, cognitive, and physiological components. Because measures of these dimensions are much less than perfectly correlated, if one's concern is with anxiety in general, then using only a single measure is likely to be misleading. The structural equation modeling methods that have become popular since the early 1980s provide a means for explicitly incorporating such fallible indicators of latent constructs into one's analytical models (see Tutorial 4, "Principles of Formulating and Comparing Models" at *DesigningExperiments.com/Supplements*).

External Validity

The final type of validity we consider refers to the stability across other contexts of the causal relationship observed in a given study. The issue in external validity is, "Can I generalize this finding across populations, settings, or time?" As mentioned in our discussion of the uniformity of nature in Chapter 1, this is more of an issue in psychology than in the physical sciences.

A central concern with regard to external validity is typically the heterogeneity and representativeness of the sample of people participating in the study. Unfortunately, most research in the human sciences is carried out using the sample of participants that happens to be conveniently available at the time. Thus, there is no assurance that the sample is representative of the initial target population, not to mention some other population to which another researcher may want to generalize. The randomization tests we considered in Chapter 1 provide one perspective on analyzing data from convenience samples that, unlike most statistical procedures, does not rely on the assumption of random sampling from a population. Such tests allow one to arrive at a p value legitimized solely by the process of random assignment of subjects to conditions. Conclusions regarding generalizations to other populations in such a case would rely on conceptual arguments about what characteristics of the population might be relevant rather than statistical arguments.

The concern in brief with external validity is that the effects of a treatment observed in a particular study may not be obtained in other contexts, such as outside of the laboratory setting, or in other locations. For example, a classroom demonstration of a mnemonic technique that had repeatedly shown the mnemonic method superior to a control condition in a sophomore-level class actually resulted in worse performance than the control group in a class of students taking a remedial instruction course. Freshmen had been assigned to take the remedial course in part on the basis of their poor reading comprehension, and apparently failed to understand the somewhat complicated written instructions given to the students in the mnemonic condition.

One partial solution to the problem of external validity is, where possible, to take steps to assure that the study uses a heterogeneous group of persons, settings, and times. Note that this is at odds with one of the recommendations we made regarding statistical conclusion validity. In fact, what is good for the precision of a study, such as standardizing conditions and working with a homogeneous sample of subjects, is often detrimental to the generality of the findings. The other side of the coin is that although heterogeneity makes it more difficult to obtain statistically significant findings, once they are obtained, heterogeneity allows generalization of these findings with greater confidence to other situations. In the absence of such heterogeneity or with a lack of observations of the people, settings, or times to which you wish to apply a finding, your generalization must rest on your ideas of what is theoretically important about these differences from the initial study (Campbell, 1986). Much more in-depth discussion of the issues of causal generalization across settings is presented by Shadish et al. (2002).

Conceptualizing and Controlling for Threats to Validity

As discussed by Campbell (1969), a helpful way to think about most of the artifacts that we have considered is in terms of incomplete designs or of designs having more factors than originally planned. For example, consider a two-group study in which a selection bias was operating. Because the two treatment groups involved, in essence, subjects from two different populations, one could view the groups as but two of the four possible combinations of treatment and population. Similarly, when a treatment is delivered, there are often some incidental aspects of the experience that are not an inherent part of the treatment, but that are not present in the control condition. These instrumental incidentals may be termed the *vehicle* used to deliver the treatment. Once again, a two-group study might be thought of as just two of the four possible combinations: the "pure" treatment being present or absent combined with the vehicle being present or absent (Figure 2.1).

In the case of such confoundings, a more valid experimental design may be achieved by using two groups that differ along only one dimension, namely that of the treatment factor. In the case of selection bias, this obviously would mean sampling subjects from only one population. In the case of the vehicle factor, one conceivably could either expand the control group to include the irrelevant details that were previously unique to the experimental group or "purify" the experimental group by eliminating the distinguishing but unnecessary incidental aspects of the treatment (Figure 2.2). Both options may not be available in practice. For example, in a physiological

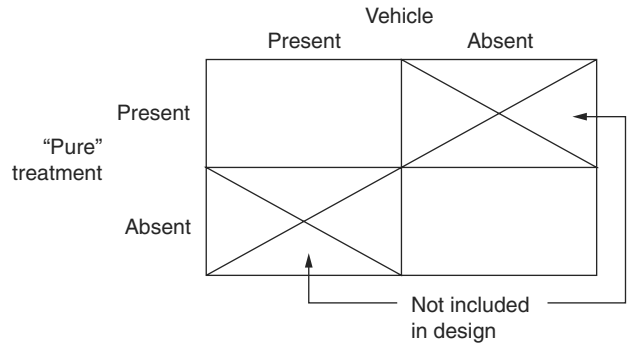


FIG. 2.1 Original design.

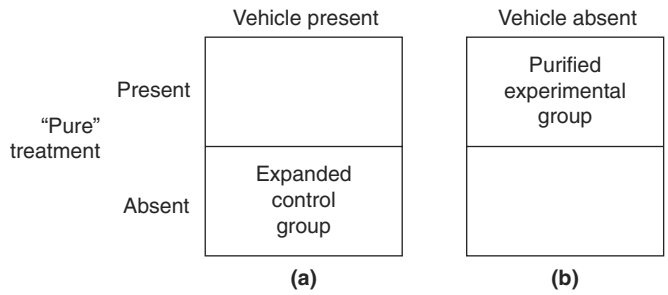


FIG. 2.2 Preferred designs.

study involving ablation of a portion of the motor cortex of a rat, the surgical procedure of opening the skull may be a part of the ablation treatment that cannot be eliminated practically. In such a case, the appropriate controls are not untreated animals, but an expanded control group: animals that go through a sham surgery involving the same anesthetic, opening of the skull, and so on, but that do not experience any brain damage.

Regarding the issues having to do with increasing the generality of one's findings, viewing simple designs as portions of potentially larger designs is again a useful strategy. One might expand a two-group design, for example, by using all combinations of the treatment factor and a factor having levels corresponding to subpopulations of interest (Figures 2.3 and 2.4). If, in your psychology class of college sophomores, summer school students behave differently on your experimental task than regular academic year students, include both types to buttress the generality of your conclusions.

Finally, with regard to both construct validity and external validity, the key principle for protecting against threats to validity is *heteromethod replication* (Campbell, 1969, p. 365ff.). Replication of findings is, of course, a desirable way of demonstrating the reliability of the effects of an independent variable on a dependent variable. Operationism would suggest that one should carry out the details of the original design in exactly the same fashion as was done initially. The point we are making, however, is that construct and external validity are strengthened if the details of procedure deemed theoretically irrelevant are varied from one replication to the next. Campbell (1969, p. 366) even went so far as to entertain the idea that every PhD dissertation in the behavioral sciences be required to implement the treatment in at least two different ways and measure the effects of the treatment using two different methods. Although methodologically a good suggestion for assuring construct and external validity, Campbell rejects this idea as likely being too discouraging in practice, because, he speculates, "full confirmation would almost never be found" (1969, p. 366).

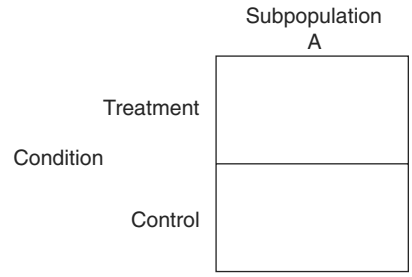


FIG. 2.3 Original design.

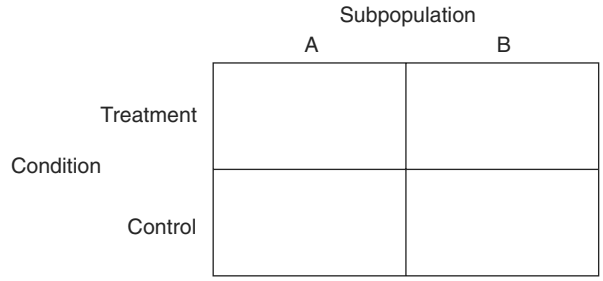


FIG. 2.4 Expanded design.

OVERVIEW OF EXPERIMENTAL DESIGNS TO BE CONSIDERED

Having surveyed some of the factors that threaten the validity of inferences from experiments, it is now time to provide a brief overview of the various types of designs we consider in this book.

First, however, a word is in order about the goals of scientific investigation and distinctions among the kinds of factors one might investigate. Science has to do with relationships among variables. At a descriptive level, the goal might be characterized as accounting for variability by predicting the values of one variable on the basis of the values of one or more other variables. At a conceptual level, however, the goal is explanation. Explanations of phenomena posit not only predictive but causal relations as well (cf. Schmidt, 1992, p. 1177ff.). Discovering causal relationships carries the greatest import for theory and also confers the practical power of insight into how a phenomenon may be controlled, or as Bacon termed it, commanded.

Predictor variables or factors may be manipulated by the experimenter or simply measured. In trying to predict scores of an undergraduate psychology major on the Psychology Area test of the Graduate Record Exam (GRE), one may find that such scores are predicted by variables such as the student's cumulative grade point average (GPA), GPA in psychology courses, and the quality of the student's undergraduate institution. Yet, from the point of view of controlling or increasing a student's score on the Psychology Area test, these do not immediately give insight into how that might be done. Perhaps much of the variance in these predictors is the result of the intelligence of the student, which might also independently contribute to the determination of the GRE score. Thus it may be the case either that some of these predictors could not be readily changed, or that changing one of them, such as the value of a student's GPA, would not cause a change in the score the student achieves on the GRE. However, if students randomly assigned to an intensive instructional program were shown to have significantly higher GRE Psychology Area test scores than students randomly assigned to a control condition, one has gained insight into how one could increase GRE psychology scores, even though the strength of the relationship with the dependent variable may be considerably weaker than the relationship between the dependent variable and the continuous individual difference variables. How to characterize such varying strength of effects is one of the major concerns of this book, with methods of assessing the strength or magnitude of effects being covered in detail in subsequent chapters.

Factors that are manipulated in studies are almost always discrete variables, whereas factors that are measured, although sometimes discrete, are more often relatively continuous. From the perspective only of accounting for variability in the dependent variable, the most important factors to include in a model are usually continuous measures of preexisting individual differences among subjects. We deal with considerations bearing on incorporating such variables into your models in Chapter 9. (For readers who have not been exposed previously to multiple regression, we have included a tutorial on the website to provide a brief introduction. For those who are familiar with multiple regression, a more in-depth discussion of the relationship between regression and analysis of variance, as well as how they relate to more advanced techniques, is also included at *DesigningExperiments.com/Supplements*.) Yet the effects of manipulated variables are clearer to interpret theoretically and apply practically, and constitute the primary focus of this book.

This last point regarding manipulated variables is of such importance that some elaboration is warranted. As we tried to suggest in Chapter 1 (see "Lawfulness of Nature"), causal relationships might be regarded as "the fundamental building blocks of physical reality and of human understanding of that reality" (Pearl, 2000, p. xiv). Now it may be the case, as some (e.g., Schmidt, 1992) argue that scientific theories constitute the epitome of such human understanding, but what is more basic and what has been the hallmark of modern science since its emergence in the

1600s is the experiment. As Kepler remarked, "Without proper experiments, I conclude nothing" (Kepler, *Astronomi Opera*, Bk. 8, quoted in Burtt, 1959, p. 50). Similarly, Galileo's first maxim of science was "description first, explanation second" (Pearl, 2000, p. 334). The meaning of experiment evolved from simply describing, that is, observing and recording facts, to the deliberate manipulation of nature. Eventually, since Fisher, to have a *true* experiment came to mean having the ability to randomly assign units to the levels of the independent variable. Indeed, the privileged status randomized experiments hold as the most secure basis for drawing causal inferences boils down to why the manipulated variable in such a study may truly be called "independent": random assignment assures that its levels will be *statistically independent* in the long run from not only explicitly controlled variables but of all causes, measured or unmeasured, known or unknown, that could influence the dependent variable. Thus, randomized experiments are relied upon both in initial exploratory studies to address questions like "What would happen if . . ." and also, and more typically, in later studies to test hypotheses derived from theory (Morey, Rouder, Verhagen, & Wagenmakers, 2014). This is not to say that there are not compelling reasons at times for conducting quasi-experiments or observational studies. Such reasons include ethical considerations (precluding, e.g., exposing individuals to severe trauma to observe the effects), practical considerations (e.g., investigating effects of public policies that an experimenter could not change), or conceptual considerations (e.g., laboratory analogues may lack critical ingredients of a construct of interest such as social support). Further, methodological advances continue to help one justify causal inferences in certain observational studies, provided the required assumptions of the method are satisfied. Nonetheless, whenever it is possible, the randomly controlled trial is the closest one can come to a universally accepted gold standard[5] for drawing causal inferences.

Some critical distinctions among types of experimental designs are introduced now that will structure much of the rest of the book. Designs differ in how many factors are being investigated, the number of levels of each factor and how those levels are selected, how the levels of different factors are combined, and whether participants in the study are repeatedly measured or not and whether they experience only one treatment or more than one treatment.

The simplest experimental design is one involving only a single factor. Among single-factor designs, the simplest situation to model, although not to interpret, occurs when there is only a single group of participants who may experience an experimental treatment, but there is no similar control group and no measured variable other than the dependent variable. This constitutes, to use Campbell and Stanley's (1963) terminology, a *one-shot case study* and permits only limited testing of hypotheses. For example, if one were to have available a sample of undergraduate psychology majors and have them experience a GRE psychology preparation course, one could compare their group mean to normative information on a typical score on the test. Because discrepancies from past averages might be the result either of the study program or because of differences between the participants in your study and the individuals in the norming group used to determine the typical score on the test, such one-shot case studies are seldom done. Instead, the more conventional design would include one or more control groups whose performance could be compared with that of the group of interest. When more than two groups are involved, one is typically interested not only in whether there are differences among the groups overall, but also in specific comparisons among combinations of group means. Designs in which the various groups are defined by the particular level of a single factor they experience are referred to as *one-way designs*, because the groups differ in one way or along one dimension. Note that this is the convention even when the levels of the factor correspond to conditions that differ qualitatively, not just quantitatively. A design with three groups that receive 5 hours, 10 hours, or 15 hours of classroom instruction is a one-way design, but so is a design that involves a group that receives classroom instruction, a group that receives a self-study manual, and a no-treatment control group.

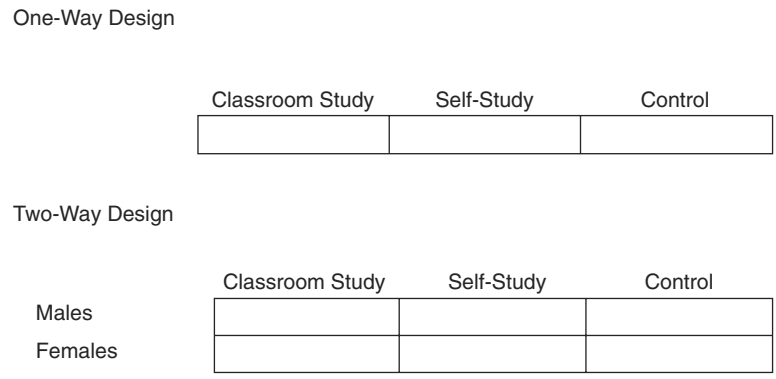


FIG. 2.5 Schematic diagrams of one-way and two-way designs.

For various practical or theoretical reasons, an experimenter may prefer to include multiple factors in a single study rather than in separate experiments. When an added factor represents a breakdown of participants previously ignored in a single-factor study of a treatment (e.g., including gender along with treatment condition in a two-factor study), typically the result is to increase power to detect the effect of the treatment factor, as well as to allow a check on the consistency of the effect across subgroups (e.g., do results differ for males as compared to females?). When the various conditions included in a study represent combinations of levels of two different factors, the design is referred to as a *two-way design.* One-way designs can be represented with a schematic involving a group of cells differing along one dimension, and in the usual case, two-way designs can be represented as a two-dimensional table (see Figure 2.5).

In cases of designs with multiple factors, designs differ in which combinations of levels of the different factors are used. In most cases, all possible combinations of levels of the factors occur. The factors in such a design are said to be *crossed*, with all levels of one factor occurring in conjunction with every level of the other factor or factors. Thus, if there are a levels of Factor A and b levels of Factor B, there would be $a \times b$ combinations of levels in the design. Each combination of levels corresponds to a different cell of the rectangular schematic of the design. Alternatively, in certain designs, not all of the possible combinations of levels occur. Among such *incomplete designs*, the most common is one where non-overlapping subsets of levels of one factor occur in conjunction with the different levels of the other factor. For example, in a comparison of Rogerian and Behavior Analytic therapies, therapists may be qualified to deliver one method or the other, but not both. In such a case, therapists would be said to be *nested* within therapy methods. In contrast, if all therapists used both methods, therapists would be said to be *crossed* with method. Diagrams of these structures are shown in Figure 2.6.

Although it is not apparent from the groups that are ultimately included in a design, one can also make distinctions based on how the levels of a particular factor were selected for inclusion. In most instances, the levels are included because of an inherent interest in that particular level or group. One might be interested in a particular drug treatment or patient group, and thus would include the same condition in any replication of the study. Such factors are said to be *fixed*, and any generalization to other levels or conditions besides those included in the study must be made on non-statistical grounds. Alternatively, if one wanted to provide a statistical argument for such generalizations, one could do so by selecting the levels for inclusion in a study at random from some larger set of levels. When this is done, the factor is designated as *random*, and how the statistical analysis of the data is carried out may be affected, as well as the interpretation.

Crossed Design

	Rogerian	Behavior Analytic
Therapist 1		
Therapist 2		
Therapist 3		
Therapist 4		

Nested Design

	Rogerian	Behavior Analytic
Therapist 1		missing
Therapist 2		missing
Therapist 3	missing	
Therapist 4	missing	

FIG. 2.6 Diagrams of crossed and nested designs.

Perhaps the most important distinction among types of design is between-subjects versus within-subjects designs. Here, the important point is whether each subject experiences only one or multiple experimental conditions. The basic advantage of the between-subjects design is that one need not be concerned about possible carryover effects from other conditions, because only one condition is experienced. Carryover effects include general effects such as practice or fatigue that result in improvements or decrements in performance for all participants regardless of condition. More troublesome are differential carryover effects, that is, where the carryover depends on which condition is experienced first. In some cases, within-subjects designs are essentially unworkable. For example, in a human memory study, if participants are taught a chunking strategy in an initial condition, they cannot validly serve as an untrained control in a subsequent condition. As a design strategy, counterbalancing the order of administration of different tasks in a within-subjects design may successfully avoid a confounding of a particular treatment with its position in a series of treatments. Even so, analysis strategies that account for the position or order effects will still typically be required.

Separate from whether or not there are carryover effects, one may be specifically interested in using the same subjects under different conditions, for statistical reasons or conceptual reasons. For example, one may want to use each participant as his or her own control, on the one hand, to achieve a more sensitive test or, on the other hand, to ask a question about how participants respond when they experience a contrast between two conditions. In many cases in psychology, the various conditions experienced by a given subject correspond to observations at different points in time. For example, a test of clinical treatments may assess clients at each of several follow-up time points. If so, the same subjects would serve in multiple conditions. Denoting the different subjects in an experiment by the letter "S" with a different subscript for each person, we can diagram a basic between-subjects design as in the top portion of Figure 2.7 and contrast that with the structure of a within-subjects design in the bottom portion of Figure 2.7.

Part II of this book, which includes Chapters 3–10, concerns various between-subjects designs, beginning with single-factor designs in Chapter 3, and considering tests of contrasts among the levels in Chapters 4 and 5. Chapter 6 considers the special case of a one-way design where the levels of the factor represent points along a single quantitative dimension, such as hours of treatment or concentration of a drug. Chapters 7 and 8 extend the discussion of between-subjects

Between-subjects design

Condition 1	Condition 2	Condition 3
S_1 S_2 S_3 S_4 S_5	S_6 S_7 S_8 S_9 S_{10}	S_{11} S_{12} S_{13} S_{14} S_{15}

Within-subjects design

Condition 1	Condition 2	Condition 3
S_1	S_1	S_1
S_2	S_2	S_2
S_3	S_3	S_3
S_4	S_4	S_4
S_5	S_5	S_5

FIG. 2.7 Between-subjects vs. within-subjects designs.

designs to studies involving multiple factors. Chapter 9 considers the implications of having a continuous predictor variable, as well as a grouping variable in the analysis. Chapter 10 concludes the discussion of between-subjects designs with a consideration of designs with random and nested factors.

Parts III and IV of the book, which include Chapters 11–16, focus primarily on designs involving within-subjects factors. Chapters 11 and 13 consider the case in which there is only one within-subjects factor. Chapters 12 and 14 consider cases in which there are multiple factors, either all within-subjects factors or one or more within-subjects factors in conjunction with one or more between-subjects factors. Chapters 15 and 16 present an introduction for models useful for correlated data, such as that obtained in repeated measures designs and with random factors to which you will be introduced in Chapter 10. Chapter 15 explains how these models, which have variously been called multilevel models, hierarchical linear models, or mixed effect models, can be used with repeated measures designs, and Chapter 16 develops how they can be used with nested designs.

Whether simple or complex, experimental designs require statistical methods for summarizing and interpreting data, and it is toward the development and explication of those methods that we move in subsequent chapters.

SUMMARY OF MAIN POINTS

Attempting to apply in another context a reported finding of a treatment effect relies on a whole chain of inferences, any one of which might be invalid. Four types of validity were distinguished. The question of statistical conclusion validity is whether the original inference about effects in the population were correct. In a two-group study, the conclusion in the initial study that the population means differ might be spurious, or a Type I error. The question of internal validity is whether the inference is correct that the treatment as implemented was responsible for any observed difference across groups. Threats to internal validity include selection bias, differential attrition, and the effects of testing, regression, maturation, and history. The question in construct

validity is whether the interpretation of the constructs involved in the purported causal relationship is correct. Finally, the question regarding external validity is whether the relationship will generalize across populations, settings, or time. A general strategy for controlling for threats to validity was presented that involved eliminating or varying design elements thought not to be critical to the hypothesized causal relationship.

Subsequent chapters will begin with the simplest possible experimental designs and then consider progressively more complex designs.

EXERCISES

1. As noted in Chapter 1, the assumption of the uniformity of nature is often questionable in the behavioral sciences. This fact is most relevant to which of the four types of validity? Explain briefly.

*2. A national study involving a sample of more than 2,000 individuals included a comparison of the performance of public and Catholic high school seniors on a mathematics achievement test [Summary data are reported by Wolfle, L. M. (1987). Enduring cognitive effects of public and private schools. *Educational Researcher, 16*(4), 5–11]. The statistics on the mathematics test for the two groups of students were as follows:

	High School	
	Public	*Catholic*
Mean	12.13	15.13
SD	7.44	6.52

Would you conclude from such data that Catholic high schools are doing a more effective job in educating students in mathematics? What additional information could make this explanation of the difference in mean scores more or less compelling?

3. A research study conducted in a rural New Mexico county investigated the effect of a program to treat drunk drivers [Delaney et al. (2005). Variations in jail sentences and the probability of re-arrest for driving while intoxicated. *Traffic Injury Prevention, 6*, 105–109]. Local judges agreed to randomly assign convicted first-time offenders to either a 28-day incarceration control condition, or to a treatment condition that involved additional cognitive behavioral therapy as well as 28 days of incarceration. An important indicator of severity of an individual's drinking problem was the total number of drinks consumed in the past 90 days. This was assessed at five time periods: pre-treatment, and at 6-month, 1-year, 2-year, and 3-year follow-ups.
   a. There was evidence that at least some judges failed to comply with the random assignment procedure. One indication of this was that the level of drinking at time 1 was significantly higher in the Treatment group than the Control group. Does this fact threaten the validity of the study? If so, what type of validity is threatened and why?
   b. Analyses of these data revealed two additional facts: (1) Although the drinking levels of participants in the treatment condition tended to be somewhat lower than those in the control conditions at the follow-up assessments, the difference between the two groups was not statistically significant at any of these assessments. (2) The decline from pre-treatment drinking levels to the average post-treatment drinking was significantly greater in the Treatment group than the Control group. A psychologist who examined these findings asserted, "There was a difference pre, but no difference post. All that's going on here is regression toward the mean." Comment on the psychologist's conclusion, indicating whether you believe regression toward the mean may have been operating in this situation, and specifically indicating whether and why you agree or disagree with the psychologist.

4. A researcher studying statistics education wants to know whether instructing students using conceptual formulas (i.e., what some refer to as "definitional" formulas), computational formulas (i.e., which minimize the steps in hand calculations), or a mix of both conceptual and computational formulas leads to better learning of statistics. To investigate this question, she conducts a study in which participants are randomly assigned to one of three study conditions. Participants assigned to the first condition are given 20 minutes to study a set of conceptual equations, those assigned to the second condition are given 20 minutes to study a set of computational equations, and those assigned to the third condition are given 20 minutes to study the set of conceptual equations and an additional 20 minutes to study the set of computational equations (the order of the kinds of equations was counterbalanced so that half of the participants in this third condition studied the conceptual equations first, while the other half studied the computational equations first). After the study period, all participants attempt to solve 10 statistics problems. Statistical tests revealed the difference in performance between the first two groups did not approach statistical significance, whereas the performance in the third group was significantly better than that in the other two groups. Based on these results, the researcher concluded in her write-up of the study that studying only conceptual or only computational equations did not make a difference in performance in solving statistics problems, but that studying both kinds of equations together allowed students to make connections that resulted in deeper understanding and better performance in solving statistics. If you were asked to review this manuscript, what type of validity would you say is most clearly threatened and why? What modification in the design of a replication of the study would you recommend?
5. Newspaper stories in 2006 reported on the risk of brain tumors among cell phone users. One such story ("Studies Find Cell Phone Link to Tumors," *South Florida Sun-Sentinel*, Feb. 4, 2006) stated "European research groups . . . have found an increased risk of brain tumors in people who have used the phones for 10 years or more," and in particular "found an increased risk of glioma, an often deadly brain cancer, in people who had used cell phones 10 years or more." Detailed data were reported in Lahkola et al. (2007). Mobile phone use and risk of glioma in 5 North European countries. *International Journal of Cancer, 120*, 1769–1775.
    a. Lahkola and her colleagues compared cell phone users to non-cell phone using controls who were matched on country, sex, and age group. In one comparison, they found risk of glioma was significantly ($p = .04$) higher in the cell phone group than the matched controls for those who had used cell phones for more than 10 years. Identify a plausible threat to the validity of the conclusion that the long-term cell phone use caused the increased risk of glioma, indicating which type of validity is threatened thereby.
    b. Two other findings of the Lahkola et al. study were that (1) the risk of glioma overall was actually significantly *lower* among all cell phone users (collapsing across years of use) than among the matched controls; and (2) when tumors were classified as being on the same or opposite side of the head as used for the mobile phone, the risk was significantly higher for long-term cell phone users relative to controls for glioma on the same side but not on the opposite side of the head. Do these two facts make the threats to the claim of that prolonged cell phone use caused an increased risk of glioma more or less plausible?

*6. In a series of studies, Emily Holmes and her colleagues have attempted to develop "a cognitive vaccine against traumatic flashbacks." In one recent article [James et al. (2015). Computer game play reduces intrusive memories of experimental trauma via reconsolidation-update mechanisms. *Psychological Science, 26*, 1201–1215], participants viewed a 12-min trauma film consisting of 11 different incidents portraying actual or threatened death or serious injury, for example, a child being hit by a car or a man drowning. Twenty-four hours later, participants in one experiment returned to the lab and were randomly assigned to either (1) a reactivation-plus-Tetris group, in which selected still images from all 11 trauma scenes were presented followed by playing the computer game Tetris for 12 minutes, or (2) a no-task control group who were not given the memory-reactivation images nor were asked to

play Tetris but simply rated classical music excerpts for pleasantness and then sat quietly for the same length of time the other group was playing Tetris. The investigators hypothesized that the memory of the film would be reactivated by the presented still images but that a taxing visuospatial task would create a capacity limitation that would interfere with reconsolidation of the traumatic memory, and hence lessen over the next week intrusive memories. Intrusive memories were defined as "scenes of the film that appeared spontaneously and unbidden in their mind" (James et al., 2015, p. 1204). The predicted difference across groups in intrusive memories was observed, however, it was not entirely clear that the memory reconsolidation task and the Tetris game were *both* necessary to reduce intrusive memories. What additional groups (e.g., expanded control groups) might be included in a subsequent study to make more compelling the claim that both memory reactivation and the Tetris game were needed to reduce intrusive memories?

7. A recent study reported an evaluation of an online mindfulness course for perceived stress [Krusche, A., Cyhlarova, E., King, S., & Wiliams, J.M.G. (2012). Mindfulness online: A preliminary evaluation of the feasibility of a web-based mindfulness course and the impact on stress. *BMJ Open, 2*, e000803. doi:10.1136/bmjopen-2011–000803]. Individuals self-selected to enroll and paid £40 (about $60) for the course, which lasted at least four weeks. Participants completed the Perceived Stress Scale (PSS) before the course, upon completion of the course, and at a 1-month follow-up. Completion of the course was self-paced, with the average time to complete the course being 6 weeks. The first 100 participants to complete the course and to complete the 1-month follow-up were included in the analysis. The average age of these participants was 48 years and 74% were women. The mean PSS pre-treatment was more than twice the mean of a probability sample of the United States, indicating the typical participant was a "highly stressed individual." An analysis of variance indicated the scores declined significantly from before to after the course. Individuals reported how often they practiced mindfulness and were classified as high ("every day or most days") or low ("sometimes" or "rarely"). There was no difference across these two groups on the amount of the PSS score decrease. The high practice group had a significantly higher PSS mean pre-treatment than the low practice group; PSS scores declined somewhat but not significantly more from pre-treatment to post-treatment for the high practice group than the low practice group. Based on these results, the authors concluded "participation in the online mindfulness course significantly reduced perceived stress upon completion and remained stable at follow-up" and "people who had higher PSS scores before the course reported engaging in significantly more mindfulness practice, which was in turn associated with greater decreases in PSS" (p. 1). Evaluate the validity of these conclusions, identifying specific threats to the different kinds of validity discussed in this chapter.

8. A psychology professor wants to claim that taking Psych 499 (an elective, independent study course available to psychology majors at his institution and typically taken in the junior or senior year as a means of receiving academic credit for working on a research project in a lab) increases the likelihood of undergraduate students at his university staying in school and graduating within 6 years. He has data on undergraduate students enrolled in his university for the past 15 years and finds that 80% of the 400 students who had taken Psych 499 graduated within 6 years of their initial enrollment at the university, whereas only 50% of the 22,500 students who did not enroll in Psych 499 during their time at the university graduated within 6 years of their initial enrollment. He believes this provides strong evidence in support of his claim.
   a. Is the validity of the claim that taking Psych 499 increases the probability of graduating within 6 years threatened here? If so, identify the kind of validity that is threatened, and specify two concrete, plausible threats to that kind of validity. If the validity of the claim is not threatened, explain why not.
   b. Even though it is not feasible to randomly assign undergraduates to take Psych 499, how might the design of his study be changed to make his claim more plausible?

*9. Assume a study finds that children who watch more violent television programs are more violent themselves in a playground situation than children who report watching less violent television programs. Does this imply that watching violence on television causes violent behavior? What other explanations are possible in this situation? How could the inference of the alleged causal relationship be strengthened?

10. Regarding statistical conclusion validity, sample size, as noted in the text, is a critical variable. Complete the following:

    a. Increasing sample size _____ the power of a test.
    increases decreases does not affect

    b. Increasing sample size _____ the probability of a Type II error.
    increases decreases does not affect

    c. Increasing sample size _____ the probability of a Type I error.
    increases decreases does not affect

NOTES

1. Note that as the term is used in the methodology literature, "selection bias" does *not* connote that the subjects in a study in general are not representative of the population to which one hopes to generalize. The lack of representativeness of participants generally is an external validity concern, as we will explain shortly. In internal validity, the concern is with differences *across* treatment conditions within a study besides the nominal treatment.
2. A major distinction among experimental designs is whether the same individuals are assessed only once or repeatedly in a given study. This is the distinction between Parts II and III of this book. Perhaps not surprisingly given that psychologists, educators, and others tend to be concerned with change, most behavioral science studies involve repeated measurements of the same units.
3. Huck and Sandler (1979) have an excellent (and fun) book, which is organized somewhat like a series of mysteries, that is designed for practicing your skills at this.
4. Shadish et al. (2002) have extended the notion of construct validity to include the problems of correctly naming or identifying not only the independent and dependent variables, but also the units and settings. While naming the units is an obvious concern when the focus of an investigation is on an individual difference variable, such as a diagnostic category in a clinical population, we prefer to treat such cases as a construct validity issue of the "independent" variable whose presumed effects or sequelae the investigator hopes to assess. Similarly, while characteristics of the setting could be regarded as representative of a larger category or kind of setting, and problems of the meaning of "setting constructs" be addressed, we believe that the meaning of the independent and dependent variable constructs should be of primary concern. Thus, we will here continue to treat such issues of generalizing beyond the setting of the current study to other locales or environments only as a problem of external validity per Cook and Campbell (1979).
5. Schmidt (1992) had strongly argued that meta-analyses provided the royal road for arriving at causal explanations. The next 25 years saw a geometric increase (of over 2,500%) in the number of published meta-analyses, resulting in the startling recent claim that "currently, probably more systematic reviews of trials than new randomized trials are published annually" (Ioannidis, 2016, pp. 485–486). The continuing emergence of software and automata streamlining the production of meta-analyses may result in even more proliferation in the future. Unfortunately, Ioannidis's (2016) evaluation of the 9,135 meta-analyses published and indexed in PubMed in 2014 concluded only 3% were "decent and clinically useful," with much larger proportions being classified as "misleading," "flawed beyond repair," or "redundant and unnecessary." Clearly they have not been the panacea that some hoped.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

II

Model Comparisons for Between-Subjects Designs

The aim of science is, on the one hand, a comprehension as complete as possible . . . and, on the other hand, the accomplishment of this aim by the use of a minimum of primary concepts and relations.

—ALBERT EINSTEIN, *PHYSICS AND REALITY,* 1936

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

3

Introduction to Model Comparisons
One-Way Between-Subjects Designs

OVERVIEW OF CHAPTER: RESEARCH QUESTIONS ADDRESSED

One of the most common questions motivating behavioral research is whether two or more conditions differ from each other in effectiveness. The conditions could be therapies delivered in a clinical setting, or types of masking of stimuli in a computer-administered task employed by cognitive psychologists, or instructional methods being compared by an educational psychologist. In this chapter, and in fact throughout all of Part II of the book, the assumption is that each of the various conditions is experienced by a different group of participants, hence the label "between-subjects" designs, as explained at the end of Chapter 2. Typically, the question of most interest is whether the difference between groups is statistically significant, that is, is the difference larger than would be expected to occur simply as a result of the variation induced by random assignment of participants to conditions.Typically, one would also like to be able to provide information about the magnitude of the difference between conditions. The current chapter will provide an introduction to methods for conducting statistical tests and characterizing the size of effects that will be generalized in subsequent chapters to apply to more complex designs. In addition, the current chapter will introduce methods for answering the critical question that arises in planning an experiment, namely, how large must the sample sizes be to make it likely that an effect of a given projected size will be detected.

PUBLISHED EXAMPLE

In a seminal study of motivational interviewing (MI), Brown and Miller (1993) assigned half of the eligible alcohol-dependent inpatients in an alcohol treatment program to receive two sessions of motivational interviewing prior to the abstinence-oriented treatment in a residential milieu program experienced by all participants. In the second of the two MI sessions, clients were given feedback about their current levels of alcohol consumption "in a supportive and empathic manner that encouraged open expression of reactions to the information" (Brown & Miller, 1993, p. 213). Three months following discharge, participants completed a follow-up interview with a research assistant who was unaware of group assignment—that is, "blinded," as discussed in

Chapter 2. Reports of quantity and frequency of drinking were corroborated via interviews of collaterals. Reports of alcohol consumption were converted into standard ethanol content (SEC) units (one "standard" drink as used in this article is equal, for example, to about 4 oz of wine or 10 oz of regular beer). The authors hypothesized the motivational treatment would affect both treatment participation and outcome. Analyses suggested that the mean level of drinking post-treatment as measured by SECs was lower in the group that received MI before the standard in-patient treatment than in the control group that only received the in-patient treatment. Ratings of treatment compliance by therapists who were unaware of group assignment were also higher in the MI group than in the control condition.

INTRODUCTION

Analysis of variance (ANOVA) has traditionally been viewed as a method of partitioning variability on a dependent variable in order to test hypotheses about differences in means. The model comparison approach we emphasize in the current volume views ANOVA from the more general perspective of being a method that researchers can use in deciding what linear model is appropriate for describing the data obtained in a study. Typically the models being compared will differ in whether certain means are presumed to be equal or to differ. The most appropriate model is one that is as simple as possible, yet still provides an adequate description of the data. Although the simplicity and adequacy of a particular model could be evaluated on an absolute basis, typically models are judged on a relative basis by comparisons with other possible models. This notion of searching for a simple yet adequate model is pervasive in statistics and in science more generally. It informs not only all applications of ANOVA, but also many other kinds of hypothesis testing.

We begin our discussion of ANOVA and linear models by approaching the problem from a purely descriptive point of view. We define a model in this context, as we develop shortly, as simply an algebraic statement of how the scores on the dependent variable arose. *Linear* is used in the sense of linear combination; that is, the models portray the dependent variable as being the result of the additive combination of various effects. We estimate the unknowns in each model in such a way that the model appears as adequate as possible; that is, the error of the model is minimized given a particular set of data. Statistical tests can then be developed as a comparison of the minimal errors associated with two competing models. To perform a hypothesis test is essentially to ask if a more complex model results in a substantially better fit to the data than does a simpler model.

To give an overview of the direction of our discussion, we first present the rationale and form of the *general linear model*, a very general framework that subsumes ANOVA models as a special case. In the remainder of the chapter, and indeed the book, we proceed from the simplest case of this general linear model to more and more complex forms. In this chapter, we consider a one-group situation, a two-group situation, and then situations involving three or more groups of subjects. In each situation, we formulate two models and compare them. To ensure that this model-comparison approach is clear, we begin with experimental designs that are one or two steps simpler than those considered in typical ANOVA texts. Besides easing the introduction to linear models, this illustrates the generality of the linear models approach.

When considering the situation involving a single population, typically the primary question to answer is, "Is the mean of the population equal to a particular value?" Naturally, any attempt to answer such a question involves estimating the population mean for the dependent variable on the basis of a sample of data, as the entire population has almost certainly not been assessed. After analyzing this situation descriptively, we develop an intuitively reasonable test statistic and relate this to a statistical test with which you are probably already familiar. (If you need a review