Statistical Applications for the Behavioral and Social Sciences

Statistical Applications for the Behavioral and Social Sciences

Second Edition

K. Paul Nesselroade, Jr.
Asbury University

Laurence G. Grimm[†]
University of Illinois at Chicago

WILEY

This edition first published 2019

Edition History
John Wiley & Sons Inc. (1e, 1993)

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
Editorial Office
111 River Street, Hoboken, NJ 07030, USA
For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.
Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data
Names: Grimm, Laurence G., author. | Nesselroade, K. Paul, Jr., author.
Title: Statistical applications for the behavioral and social sciences / K. Paul Nesselroade, Jr., Asbury University, Laurence G. Grimm, University of Illinois at Chicago.
Other titles: Statistical applications for the behavioral sciences
Description: 2nd edition. | Hoboken, NJ : John Wiley & Sons, Inc., 2019. | Includes index. | Earlier edition published in 1993 as: Statistical applications for the behavioral sciences [by] Laurence G. Grimm. |
Identifiers: LCCN 2018022259 (print) | LCCN 2018025247 (ebook) | ISBN 9781119355380 (Adobe PDF) | ISBN 9781119355366 (ePub) | ISBN 9781119355397 (hardcover)
Subjects: LCSH: Social sciences–Statistical methods.
Classification: LCC HA29 (ebook) | LCC HA29 .G7735 2019 (print) | DDC 300.1/5195–dc23
LC record available at https://lccn.loc.gov/2018022259

Cover design: Courtesy of Meg Sanchez
Cover image: Courtesy of Max Ostrozhinskiy on Unsplash

Set in 10/12pt Warnock by SPi Global, Pondicherry, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

For Cheryl, Andrew, Sarah, and Lisa
– each of you bring special meaning to life

Contents

Preface

This textbook is an outgrowth of our combined 40+ years worth of experience teaching undergraduate statistics for social and behavioral science students, an experience that has impressed us with the dread students face when entering the course and the frustration they voice in trying to understand statistics. The dread is most likely a result of the unimaginative manner in which mathematics is taught in the American grade school system. Unfortunately, there is nothing a statistics instructor can do about that. However, there is something the instructor of an undergraduate statistics course can (and must) do to combat this frustration. To be sure, most students may not find a statistics course as engaging as a course in social psychology or child development, but it need not rival the forced reading of the unending pages of terms and conditions associated with approving a new software program!

This book has been written with the typical student in mind – one who not only dislikes math but also has no confidence in their ability to "deal with numbers." Consequently, even the student with only a little background in algebra will be able to understand the computational flow of the formulas. A knowledge of algebraic derivations and proofs is unnecessary for mastering the material in this text.

Goals of the Text

The primary goal of this book is to teach students the conceptual foundations of statistical analyses, particularly inferential statistics. Where applicable, the conceptual foundation of statistical tests is explained in the context of *standardized* scores. Throughout the chapters on hypothesis testing, the surface mechanics of computing a test statistic are always related to the underlying sampling distribution of relevance. In this way, students learn *why* the formula for a test statistic looks as it does, and they gain an appreciation of the *statistical* meaning of each analysis.

Emphasis on the conceptual underpinnings of hypothesis testing distinguishes this textbook from those that offer a "cookbook" approach. In addition, this text places heavy emphasis on the research context of the statistical analysis under discussion. As a result, students will feel "connected" to the research activities of social and behavioral scientists and come to view formulas as tools to answer questions about human behavior.

Nonetheless, the arithmetic operations involved in arriving at problem solutions are not sacrificed. Indeed, another goal of this book is to teach students how to "work the formulas." Learning to "crunch the numbers" is accomplished by presenting definable, clearly specified steps in working through statistical problems. Despite the existence of numerous statistical software packages that can quickly and accurately arrive at the solution to problems, we believe that the initial introduction to a statistical tool should utilize a hand calculation. This number-crunching process provides the student with a deeper understanding of the inner workings of statistical formulas. Once familiarity is achieved by crunching through small sample versions of the mathematics of statistical tools, then the introduction of a computer software program becomes a welcome timesaving aid, and not a method of obscuring what is going on. For this reason, at the end of most chapters, brief tutorials are presented, showing the user how to use Microsoft® Excel and SPSS® to compute various descriptive and inferential statistical values.

Organization and Flexibility

The text has 18 chapters organized into seven parts: (1) "Introduction," (2) "Descriptive Statistics," (3) "Inferential Statistics: Theoretical Basis," (4) "Inferential Statistics: z Test, t Tests, and Power Analysis," (5) "Inferential Statistics: Analysis of Variance," (6) "Inferential Statistics: Bivariate Data Analysis," and (7) "Inferential Statistics: Nonparametric Tests." The breadth of coverage of topic areas makes this book suitable for a semester course, a two-quarter course, or a one-quarter course. If students have had exposure to research design, Chapter 1 may be skipped or used as a brief summary of research concepts.

Because earlier chapters build the conceptual foundation for later chapters, there is only a modest amount of leeway in assigning chapters out of sequence. Nonetheless, the chapters covering chi-square and other nonparametric tests may be assigned before the chapter on one-way ANOVA. The chapters covering two-way and repeated-measures ANOVA can be omitted without hampering the students' understanding of subsequent chapters. The chapters covering linear correlation and regression treat these data analytic procedures in the context of inferential statistics. Consequently, it is not recommended that they be presented immediately after the section on descriptive statistics. The chapter on

probability can be left out if there is limited time or a desired lack of emphasis on the theoretical underpinnings of inferential statistical tests. Finally, the chapter on power may be omitted without sacrificing the students' understanding of hypothesis testing. The concept of power is defined simply whenever it is mentioned in chapters covering hypothesis testing.

Student Aids in the Text

Because most students approach statistics with considerable foreboding, we have included several pedagogical features in the text to enhance learning and maintain motivation:

1) The application of formulas is illustrated in step-by-step computational procedures so that students can master the sequential process of arriving at the correct answer to sample problems.
2) Boxes that highlight the topic under discussion using published and unpublished research are presented in each chapter. The material is selected for its interest value to students.
3) One series of boxes addresses, head-on, a major issue in the social and behavioral sciences, the so-called replication crisis. These boxes bring students into this larger discussion, help them understand the underlying issues, and empower them to think critically about their own and others' research.
4) Spotlights present biographical sketches of some of the luminaries in the field of statistics. Interesting aspects of the person and their times are provided to bring the material to life.
5) A list of selected formulas can be found in the back of the text for easy reference.
6) Recognizing that many instructors want their students to be able to communicate the outcomes of statistical analyses in written form, most chapters contain a section informing the students how to present statistical findings in sentence form.
7) Each chapter ends with an extended summary of the chapter (where applicable), brief tutorials for how to use Microsoft® Excel and SPSS® to generate statistical values associated with that chapter, a presentation of the key formulas, and numerous questions and exercises for concept checks and practice. Many of the questions and exercises are based on published research findings of high interest to students, so that students not only receive practice in data analyses but also increase their knowledge of the content of psychology.
8) A glossary of terms is provided at the end of the text.
9) A list of symbols is provided in the back matter.

Appendix A contains helpful tables for determining various critical values needed for determining probability and testing null hypotheses. (Although the tables are incomplete, they will provide the appropriate values for almost all of the exercises. However, students may need to reference tables online to find the critical values needed to answer a few questions.)

It is our experience that students overwhelmingly prefer that *all* the answers to work problems be provided, and so they are, in Appendix B. In addition, for computation problems, the answers are provided along with the interim steps, thereby allowing students to locate the source of potential errors in the use of formulas. Most of the chapters also include short data sets that can be used with any statistical software program. The answers to these problems are also provided in Appendix B.

Appendix C presents brief instructions for basic data entry procedures for Microsoft® Excel and SPSS®. This resource further supports student's ability to use these software products for statistical calculation purposes.

Acknowledgments

Many people at John Wiley & Sons, Mindy Okura-Marszycki, Kathleen Pagliaro, Vishnu Narayanan, and Grace Paulin, S., have contributed to this textbook.

We are grateful to the literary executor of the late Sir Ronald A. Fisher, F. R. S., to Dr. Frank Yates, F. R. S., and to Longman Group Ltd., London, for permission to reprint Tables III, IV, and VII from their book *Statistical Tables for Biological, Agricultural, and Medical Research* (6th edition, 1974).

A special thanks is also extended to Emma Nesselroade Miller for the graphic design work, Tricia Taylor for her help with the PowerPoint slides, Meg Sanchez for her help with the cover design, and Daniel Nesselroade for consultation and wording advise.

Our largest debt of gratitude goes to the reviewers of the manuscript. For many years they remained anonymous, yet we came to know many through their styles of criticism, their preferences for how to teach statistics, and their thoroughness. Some went beyond the call of duty for time spent on the manuscript – to each of you, a special thanks.

Judy Britt performed an accuracy check on the entire manuscript, worked all of the statistics problems, arranged the index, and continually amazed me with her "eagle eye." Her diligence is appreciated beyond words.

To the student, as well as the instructor, please send any suggestions or comments that you think ought to be considered for the next edition to Paul Nesselroade, Asbury University, Psychology Department, 1 Macklem Drive, Wilmore, KY, 40390.

About the Companion Website

This book is accompanied by a companion website:

http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

The Instructor Companion Site includes:

1) **PowerPoint slides** for all the chapters of the book for instructors
2) **Kick Start Quizzes** for instructors

Part 1

Introduction

1

Basic Concepts in Research

1.1 The Scientific Method

This is a textbook about statistics. Simply defined, statistics are the mathematical tools used to analyze and interpret data gathered for scientific study. It is paramount to remember that statistical analyses and interpretations do not exist in a vacuum. They occur within the larger scientific research process. Both *how to analyze* and *how to interpret* the data are quite dependent upon the surrounding research context. While the subject of statistics can be singled out and studied in isolation (as this textbook demonstrates), it is inextricably linked to the larger scientific enterprise. As such, it is appropriate to review the basic features of "doing" science before we delve into statistics proper.

The scientific method can be conceptualized as a three-step recursive process. Each can be summarized as follows:

Theory. Theories are an attempt to explain and organize collections of data observed about the topic (or "phenomenon") under scrutiny by appealing to general principles and relationships that are independent of the topic itself. Take, for example, a line of research on the endurance of friendships. In theorizing why some acquaintances lead to enduring friendships while others do not, one could propose that personalities are a bit like magnets; similar ones repel one another, while dissimilar personalities are drawn together (i.e. opposites attract). Clearly, this theory appeals to the prior concepts "magnets," "personality," "similarity," and "dissimilarity" in purporting to explain why certain friendships pass the test of time while others do not.

Not all theories can be considered "scientific." For a theory to qualify as properly "scientific," it must be testable. By testable we mean: is it potentially falsifiable? Can it be placed into jeopardy and potentially observed to be untrue? If it cannot, it still remains a theory, but it is not considered to be properly "scientific." Using testability as a criterion, for example, the theory that each of our choices, past and future, is actually predetermined by some combination of

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

our DNA and behavioral conditioning through our previous experiences, could hardly be considered "scientific." While many people believe it to be true, how exactly would we go about testing it? And chiefly, how exactly could we place this theory in jeopardy and observe it to be true or untrue?

Hypothesis. In the light of any scientific theory, it should be possible to generate predictions about the data one expects to observe – this is a hypothesis. Sticking with the aforementioned magnetic theory of friendships, one hypothesis might be as follows: If we measure the personalities of incoming university students who are randomly assigned to live on a given hall in a freshman dorm, we might expect to find that students who have quite discrepant personality profiles are more likely to be friends at the end of the semester than those who had similar personality profiles.

Because it is possible to find evidence that would not support this hypothesis, we can say that this theory is "testable." However, we cannot stop at hypothesizing. To say anything meaningful, we have to complete the research process and actually go out and do the work, set up the study, gather the participants, and carefully collect the data. This leads us to the final step.

Observation. The gathering of scientific observations is done by careful and systematic measurements of events occurring in the world by using our five senses, often with the aid of various scientific tools and instruments. In our example, we would want to measure meticulously our incoming freshman's personalities as well as the nature of the friendships on the hall at the end of the semester. These observations, then, would be organized and interpreted. Ultimately, what is concluded would reflect back upon the theory. Observations will either support the theory, fail to support it, or, perhaps, partially support it. The circle is complete as we relate our findings to our original theoretical proposition.

In our particular example, we should not be too confident that supporting data will be found – previous research suggests we will probably be disappointed (e.g. Buss, 1985). And that is an important point – if supporting data is not found, so much the worse for the theory. We may need to think differently about why some friendships begin and endure while others do not. As would be expected, accurate theories will be supported by our observations. Supporting observations can both affirm a theory and lead to clearer and more refined articulations of that theory. More precise theories, in turn, lead to new hypotheses, and the cycle starts over again. The process is circular and recursive, with each cycle ideally spiraling toward a more accurate understanding of the topic under investigation.

The specific role of statistical analysis is found in the interpretation of our numerically represented observations. What do the numbers mean? What do they *not* mean? For whom do they have meaning? Furthermore, how certain are we that our conclusions are accurate? On what do we base our sense of certainty? These are often not easy determinations to make. The central purpose of

this text is to dissect and explain how this part of the research process works. The remainder of this introductory chapter will lay out an overview of the research enterprise.

1.2 The Goals of the Researcher

Scientific researchers set out with earnest intention to study carefully, logically, and objectively a particular topic of interest. Depending upon what is already known about the topic, what one *wants* to learn about the topic, and what one realistically *can* learn about the topic, researchers adopt different "goals" for their projects. Often, the initial goal a researcher has when first addressing a topic of interest is that of **description**. Scientific description is the process of defining, identifying, classifying, categorizing, and organizing the topic of interest. Explicit delineation of the boundaries of the topic is crucial. What exactly constitutes the topic and what clearly does not constitute the topic? How many forms can it take? How frequently are these various forms found?

For example, if we were interested in studying the various ways in which people take vacations, we would first have to define what a vacation *is* and what it *is not*. Is an afternoon day trip to a community park a vacation? What about an extra day tacked onto a work-related business trip? It is *not* a requirement for all researchers to agree on the same definition of what "is" and "is not" a vacation, in order for vacations to be studied. However, it is absolutely imperative that the readers know explicitly what we, the researchers, mean when we say that we are counting days spent on vacation. In other words, concepts must be operationally defined. An **operational definition** is a precise verbal description of the concrete measurement of that concept, as it will be used in a given research project.

Another issue would be to decide how many different ways "vacation" can take place. For example, someone might suggest that there are fundamentally two different *kinds* of vacations: one kind that is designed around relaxation and focuses on bodily rest and another kind that is designed around engaging in new and exciting experiences. Another researcher may come along and suggest that there is actually a third kind of vacationing – one that combines the two and incorporates both time dedicated to bodily rest *and* time dedicated to having new experiences (e.g. traveling the country in a motor home). Widespread agreement regarding the particulars of the concept "vacation" is not required, of course, for it to be studied. The crucial point to be made is that researchers who are dealing with a topic at this level are going to gather statistics that reflect the relative frequencies and averages pertaining to the categories of the topic under investigation, as *they* understand them to exist. For example, one researcher might find that only 20% of vacations are of the relaxation

variety, while another researcher, using a different operational definition, might find a quite different percentage. Statistical statements, then, can only be properly interpreted once the larger research context is correctly understood. Finally, it should be noted that the statistical needs associated with meeting this initial goal of "description" are usually not too sophisticated.

Another goal of the researcher would be one of **correlation** (or **prediction** or **association** – these are all analogous terms). Correlation involves a description of the degree of relationship between the topic of interest and other variables. For example, in our study of vacations, we might be interested to see if there were a relationship between the age of the vacationer and the type of vacation chosen. Here, we would be measuring two variables (the "age of the vacationer" and the "type of vacation chosen") and determining if there was a relationship between them. As a rule, it requires more sophisticated mathematical work to establish correlations. It is critical to realize that research designed to show correlations does not allow us to draw causal conclusions. For example, if we find that older individuals, more so than younger ones, prefer to take vacations centered on rest, we could not justifiably conclude that age *causes* people to want to take vacations that are more restful. It could very well be, for example, that older people simply grew up in a time when vacations were generally understood to be more restful in nature. As a result, they formed their vacationing expectations and habits accordingly. On the other hand, another possibility might be that there simply are not as many exciting and new vacationing experiences geared toward an older audience, when compared with those available to the younger crowd. If the set of vacation options were different, then perhaps the numbers of older vacationers choosing active vacations would increase. Understand this clearly: One of the most frequently observed critical thinking errors is the tendency to impose a causal interpretation on data that is correlational in nature. The interpretation of data from studies with a correlational goal must refrain from causal language. For example, in this case, it would be correct to conclude only that age and vacationing style are correlated, such that older individuals are more likely to spend their vacation time being restful, when compared with those who are younger.

The most ambitious goal of the researcher would be one of scientific **understanding**. This permits one justifiably to draw a cause-and-effect relationship between the topic of interest and some other variables. There are a myriad of subtle and yet important issues related to making "cause-and-effect" statements. As such, it is simply impossible to treat adequately the subject of sufficient causation in this text. At a minimum, one must realize that even when causal statements are warranted by the research process, their explanatory power is rather limited. A further exploration of these limitations is presented in Section 1.6 near the end of this chapter.

In order to gain even this limited understanding of causality, we will need to engage in a specific form of research investigation termed an **experiment**.

An experiment is a precise term reserved for a research project in which (i) a high degree of control over the presumed *causal* variable, (ii) careful measurement of the presumed *effected* variable, and (iii) complete control over all other variables are all scrupulously maintained. For example, using the "vacationing" study, if we think the amount of disposable income was influencing the type of vacation people chose to experience, then we might (using a generous research grant!) gift some randomly selected participants $1000 and ask them to choose between two different but equally expensive vacation packages: one centered around resting and the other centered around doing. We could then grant other randomly selected participants $5000 and pose the same question. If different rates of vacation preferences emerged between the two groups of participants, the researchers would then be justified in claiming that their findings suggested the following: The amount of disposable income influenced (or, in some sense, *caused*) the type of vacation people chose to experience.

Before going any further into the world of experimentation, we need to step back and survey the different ways scientific researchers think about variables.

1.3 Types of Variables

In a general sense, a variable is anything that can assume different values – anything that can *vary*. In the research process, variables are employed in different types of roles. The following paragraphs will introduce some of the major roles played by variables, as well as some of the related terminology.

Independent and Dependent Variables

As previously noted, when designing an experiment, the researcher attempts to examine how one variable influences another variable. The **independent variable** designates the variable that is thought to play the *causative* role in a cause–effect relationship. It is selected and manipulated by the experimenter in order to observe its effect on another variable. By **manipulation**, we mean the controlled presentation of that variable. The nature of this manipulation can be described as being either quantitative or qualitative. A **quantitative independent variable** means that participants are exposed to different *amounts* of the independent variable. Suppose a researcher hypothesizes that the time it takes rats to learn a maze depends on the magnitude of reward provided at the finish line. Three groups of rats are used, with each group receiving increased amounts of reward: 5 food pellets, 10 food pellets, and 15 food pellets. In this example, there are three levels of the independent variable, and they are distinguished from each other quantitatively.

A **qualitative independent variable** establishes levels by contrasting different *kinds* of treatments or by the *presence or absence* of the independent variable. Using the maze-learning example, suppose one group of rats received one type of food pellet (e.g. salty) at the end of the maze and the other group received another type (e.g. sweet). The groups do not differ in the amount of the independent variable, but rather by the *kind* of independent variable used. Another example of a qualitative independent variable using the maze-learning paradigm would be for one group of rats to receive a food reward at the end of the maze while rats from the other group receive nothing. The groups differ by the *presence or absence* of the independent variable. (Although here it is possible to think of this difference in quantitative terms, it is most readily thought of in qualitative terms, like a toggle switch that is either "on" or "off.")

Note that when a condition is marked by the absence of the independent variable, it is oftentimes referred to as the **control group**. Conversely, when a condition is marked by the presence of the independent variable, it is oftentimes referred to as the **experimental group**. Many research efforts can be described in brief as "control-versus-experimental" studies. Of course, we could have more than a single experimental condition being applied. A three-condition study might have a control condition as well as an experimental I condition and an experimental II condition and so on.

Finally, note that the term **treatment** is often used instead of the term "independent variable" when the procedure introduced might have an effect on the participant's behavior. For example, if our independent variable in the maze-learning study was not food at the end of the run, but rather a prescribed style in which the experimenters handled the rats (e.g. "rough" handling vs. "gentle" handling), then we might say that the rats are getting different *treatments*.

The **dependent variable** designates the variable that is thought to play the *effected* role in a cause–effect relationship. Sticking with our maze-learning example, either the time it took to finish the maze or the number of errant turns made by the rat during the completion of the maze (or both) could be assigned as the dependent variable.

While the independent variable may in reality influence all sorts of other variables, the dependent variables are only those that are being carefully observed and measured. For example, suppose a researcher is interested in teaching people self-control techniques to improve pain tolerance. First, "pain tolerance" needs to be operationally defined. Let us say "pain tolerance" is the duration of time participants can keep their hands submerged in a bucket of ice water. As such, "pain tolerance" serves as the dependent variable in this experiment, and "self-control techniques" serve as the independent variable (or treatment). Participants learn one of two different mental imaging techniques designed to control their reaction to pain (this would be a qualitative independent variable, by the way) – and then they are measured to see how long they can keep one arm submerged in the ice water. While the participants maintain their hands in the

ice water, an array of other behavioral responses may also occur, such as anxiety, memories of past experiences with cold water, fidgeting, looking around the room, increased breathing rate, and so on. Are these other behaviors dependent variables? No, because the researcher is not systematically observing and measuring these responses. If the researcher decided to measure both the duration of time participants keep their hands submerged in ice water and the number of breaths per minute, then the study would contain two dependent variables.

■ **Question.** *A psychologist is interested in determining which of two sales techniques is more effective in influencing participants to purchase a more expensive car. In the "positive condition," participants are told about the many positive attributes of Car A compared with Car B, even though Car A costs $1000 more. In the "negative condition," another group of participants are told about all the undesirable attributes of Car B compared with Car A (the pricing remains the same in both conditions). After the lecture, each participant is asked to state which of the two cars they would most likely purchase. Identify the independent and dependent variables.*

Solution. The independent variable can be identified by asking, "What variable is being manipulated by the experimenter and is presumed to play the causative role in the cause–effect relationship under exploration?" It should be clear that the "sales technique" is what is being manipulated. To determine the dependent variable, we can ask ourselves, "What is the particular behavior of the participants that is being measured?" or "What variable is being affected in the cause–effect relationship under consideration?" The dependent variable is clearly specified at the end of the example: "Each participant is asked to state which of the two cars they would most likely purchase." Simply stated, one might call this dependent variable something like "car preference." ■

As stated earlier, two defining features of an experiment are the careful control (manipulation) of the presumed causal (also known as "independent") variable and the careful measurement of the effected (also known as "dependent") variable. There exists, however, a third defining feature in any experiment: the control of all other variables in the study. With any experiment, there will be several other variables in play (variables other than independent and dependent variables) that can be identified and may very well be influential on the outcome of the experiment. For example, in the aforementioned mental imagery and pain tolerance study, one could imagine several other variables of interest – like the ages, ethnicity, and various health indicators of the participants (e.g. resting heart rate) as well as differences in the experimental environment (e.g. the room temperature, the presence of other people who might be encouraging the participant, how much fat deposition or muscular development the arm exhibits, and so on). Any variable other than independent or dependent variables found in an experiment is referred to as an **extraneous variable.** These extraneous variables must be *controlled* in order to draw accurate conclusions from the results of the study.

1.4 Controlling Extraneous Variables

Controlling Extraneous Variables by Holding Constant

Several techniques have been developed to impose controls on extraneous variables; the following paragraphs will look at a few of the most common ones. When the experimental situation is not controlled, there is the possibility that some variable other than the independent variable will be (at least partially) responsible for changes in the dependent variable. This potential interference by an uncontrolled variable creates confusion, then, regarding what actually caused the change. An uncontrolled extraneous variable is called a **confounding variable**. The presence of confounds in an experiment threatens the researcher's ability to draw meaningful conclusions from the study. Furthermore, there are no statistical techniques or "fixes" that can be used to salvage a poorly designed study containing confounds. Statistical analysis rests fully and firmly on sound research methodology. If the methodology is poor, statistical analyses are helpless to repair the situation. The study is simply a "bust"; the results are invalid, and no real confidence can be attached to them, for how could one ever be certain that the uncontrolled confounding variables were not what was really responsible for the measured outcomes?

Holding constant an extraneous variable involves treating the variable like a constant, simply not letting it vary. Consider the previous research example in which different sales techniques were compared. Presumably, one factor that influences people's decisions to purchase a car has to do with the visual presentation of the vehicle. How the car is displayed, then, is an extraneous variable in our example. If this were uncontrolled, it could clearly confound the study. For example, what if Car A was displayed prominently on a ramp for the participants in the "positive" condition but moved, due to incoming new inventory, and placed blandly amid a row of other cars the following day when participants were exposed to the "negative" condition? If it were discovered that people in the "positive" condition opted for Car A more so than those in the "negative" condition, is there any clear way of knowing that the preference for Car A was due exclusively to the different sales technique utilized (the actual independent variable) and not simply as a result of the differing presentations (the confounding variable)? The potential causal role of the sales technique could be confounded by the variable "presentation style." Imagine, however, if "presentation" were controlled by holding it constant. Simply doing this would effectively remove it as a competing explanation. In this way, the confounding variable is neutralized by making the presentation of both Car A and Car B identical in both conditions.

Consider another example of a confounding variable (see Table 1.1). A simple psychotherapy study is aimed at contrasting two therapeutic approaches to depression. One group of depressed participants receives psychoanalysis

Table 1.1 Diagram A depicts "therapist" confounded with "treatment." Diagram B depicts a redesign in which therapist effects are spread equally across treatment conditions. Diagram C depicts a redesign in which the therapist is held constant.

Diagram A			
Behavioral therapy		Psychoanalysis	
Therapist 1		*Therapist 2*	
30 depressed clients		30 depressed clients	
Diagram B			
Behavioral therapy		Psychoanalysis	
Therapist 1	*Therapist 2*	*Therapist 1*	*Therapist 2*
15 depressed clients	15 depressed clients	15 depressed clients	15 depressed clients
Diagram C			
Behavioral therapy		Psychoanalysis	
Therapist 1		*Therapist 1*	
30 depressed clients		30 depressed clients	

conducted by an experienced psychoanalyst, while a second group of depressed participants receives behavioral therapy from an equally experienced behavioral therapist (Diagram A, Table 1.1). Suppose at the end of the study the participants who received behavioral therapy were, on average, less depressed than the participants who were treated by psychoanalysis.

What is, perhaps, the most obvious confound? It is possible that the behavioral therapist had some personality characteristics (e.g. warmth, understanding) that were actually responsible for the better results instead of the behavioral therapy itself. Had the two therapists switched roles, perhaps the psychoanalytic approach would have appeared superior. In this example, "therapist" is confounded with "treatment type." Any difference, then, between the two groups of participants in terms of their recovery from depression cannot be unambiguously attributed to the relative effectiveness of the therapy styles. How could we control this extraneous variable of therapist personality? One solution, described above (Diagram B, Table 1.1), would be to have half of the participants in both treatment style conditions treated by each therapist. This technique used to control extraneous variables is very similar to "holding constant." It is called **balancing**, and it works by representing two forms of an extraneous variable equally in both conditions. The effect of each therapist would now be equally represented in the two treatment conditions, thereby removing any bias toward one form of therapy.[1] Since the effect of different

1 It would be important to ensure that each therapist was equally proficient in conducting both behavioral therapy and psychoanalysis.

therapists would be balanced across the conditions, any difference in the improvement of depression could be attributed to the treatment rather than the therapist. Another solution (Diagram C, Table 1.1) would be to hold the troublesome variable constant by using the same therapist for both conditions.[2]

One downside to controlling extraneous variables by holding them constant is that the scope of the experiment becomes limited with regard to that variable. For example, if we decided to control the variable "biological sex" by using only females in the experimentation, the findings could only justifiably be applied to females, for how could one know that the same results would extend to males when they were never studied? Suppose only biological females were used in the imagery and pain tolerance study; if one imagery technique was found to result in higher pain tolerance than the other, could it be fairly said that this finding also pertained to biological males? When a variable is controlled by holding it constant, the interpretation of the findings regarding that variable is limited, and any extrapolation to other forms of that variable is subject to debate. We will have more to say about generalizing the findings of a study when we introduce the concept of "external validity" presented later in this chapter.

Controlling Extraneous Variables by Randomization

Another method for controlling extraneous variables is by using a technique known as randomization. A host of potentially confounding variables associated with (i) differences between the various participants in a study, (ii) differences between the experimental settings, and (iii) even differences in the stimuli used can be controlled by careful randomization. It is important to analyze each of these potential problem areas one at a time, for each illustrates a way in which the technique of randomization can potentially be applied to control the effect of extraneous variables. First, consider the numerous differences that naturally occur between different participants in a study.

Participant variables are characteristics of the participant that are fixed before the experiment is even begun. This term "fixed" applies to their a priori existence and not to their inability to be changed (though sometimes, they truly cannot be altered). One way to categorize participant variables is to consider them in three distinct groups: physical attributes, demographics, and psychological traits. Table 1.2 lists many common participant variables in these respective categories. Obviously, an exhaustive listing of participant variables is impossible.

Participant variables will create difficulties in an experiment when they are not adequately controlled. Most problematically, they can become confounding variables when they are proportionately linked to differing levels of the

2 It would be important to ensure ahead of time that the therapist chosen for the study was equally proficient in conducting both behavioral therapy and psychoanalysis.

Table 1.2 Some common participant variables.

Physical attributes	Demographics	Psychological
Height	Income	IQ
Weight	Family size	Need for approval
Handedness	Ethnicity	Type A personality
Running speed	Occupation	Trait anxiety
Strength	Education	Introversion
Sex	Religion	Dominance

independent variable. This thorny dilemma, however, can be elegantly overcome by using the tool of **random assignment**: the designation of participants into various conditions within a study such that each participant is equally likely to be assigned to a given condition. Random assignment is an extremely powerful tool that can control innumerable participant variables all at the same time. The following examples indirectly illustrate this power through their failure to use random assignment, thereby potentially confounding the experiments.

► **Example 1.1** An educational psychologist is interested in evaluating the effects of a psychological technique that is expected to help college students become more efficient in completing a series of arithmetic problems. Efficiency, the dependent variable, is defined as the time it takes students to complete one page of simple math problems. The technique, or independent variable, is the use of self-motivating statements. A self-motivating statement is a phrase the student learns to repeat in the event that they become distracted from solving the arithmetic problems. The experimental group is trained to repeat these self-motivating phrases, while a second group of students, the control group, is not.

The experiment is conducted in a typical classroom setting. To save time, the experimenter assigns the first 20 students who arrive for class to the experimental group. All later arriving students are asked to wait in the hall until the self-motivational training has been completed. These students in the hallway are assigned to the control group and brought in later when it is time for everyone simultaneously to work the math problems. The results of the study showed that students in the group trained to use self-motivational techniques, on average, worked much faster than students in the control group.

While it is impossible to know, it is certainly plausible to consider that a participant variable (e.g. some psychological trait or demographic trait) was responsible for the difference in performance. For example, suppose it is true

that mentally disciplined people tend to be both more punctual (thus overrepresented in the training group) and faster at working math problems than mentally undisciplined people. This psychological variable exists prior to the study. It is also an extraneous variable that might be confounding our study since it was not controlled. While one could control it by holding it constant (e.g. only use mentally disciplined participants), it could also be controlled by randomization – by arbitrarily assigning participants to the two conditions. In this example, a better methodology for placing students into experimental and control groups would have been to flip a coin to assign each participant to a condition. Employing a random group assignment strategy like coin tossing would eliminate the prospects of overrepresenting any specific participant variable in either group. Clearly, there are hosts of other participant variables simultaneously controlled by random assignment as well. Potentially important confounds like age, IQ, motivation, biological sex, and so on would all be neutralized. Notice how all participant variables can be controlled with one well-executed randomizing technique. ◄

It is critical to realize that a variable is controlled even if it is not perfectly balanced between the two conditions. For example, suppose one group contained a couple more biological male students than the other; this would not mean that random assignment had failed to control for biological sex. If an imbalance occurs through a truly random process (e.g. flipping a coin), then the resulting disparity cannot confound the study, rather it is just introducing what is called *error*. This is not as worrisome as it seems, for it is not the sort of error that invalidates a study. In reality, it is this kind of error that statistical methods are well suited to overcome. This will be addressed in more detail later in the text, but a brief introduction right now facilitates the discussion of another point that pertains to the control of extraneous variables by randomization. This is the concept of group size. Because randomization mechanisms often do not perfectly balance a particular participant variable between the conditions, it is important to have groups that are large in number to help correct the problem. A thought experiment will help to explain why. Imagine someone flipping a coin 10 times and noting how many heads are found. Now imagine them doing these sets of 10 flips repetitively. On average, they will get 5 heads for each series of 10 flips, but sometimes they will get a few more or a few less. Periodically they may even get as few as 1 or 2 heads or as many as 8 or 9. This is a big difference, particularly if we consider heads and tails as representative of a binary participant variable like biological sex. Now, imagine the investigator flipping a coin 1000 consecutive times. Further, imagine this being done repetitively in sets of 1000 (thankfully, we are just imagining this). On average about 500 heads will be found for each set of 1000 flips. A few more or a few less would likely be found each time, but the rough percentage of both outcomes would hover around 50%. Very rarely would less than 450 or more than 550 be recorded.

Therefore, it would be exceedingly infrequent to find the sort of large percentage differences between heads and tails that was seen in the 10-coin-flip scenario. This illustrates how larger sample sizes minimize the error that accompanies a randomization process. This is one of the profound advantages that large group size affords. The probabilistic benefits attached to large samples will be revisited in more depth later in the text.

Consider another example – this one based on a well-known study in the area of psychosomatic medicine.

► **Example 1.2** One of the earliest experimental demonstrations of the development of ulcers was conducted by Brady and others (Brady, 1958; Porter et al., 1958). Pairs of monkeys were exposed to electric shocks, which could only be avoided if one of the monkeys pressed a lever. The same monkey was always responsible for working the lever and was labeled the "executive" monkey. If the executive monkey failed to respond in time, not only they but also the "control" monkey would receive the electric shock. Therefore, while the numbers of shocks received by the executive and control monkeys were the same, the executive monkey was always responsible for managing the onset of the electric shock. Within two months, all of the executive monkeys either had died or had become too incapacitated to continue in the study. Autopsies performed on the executives showed extensive gastric lesions. In contrast, the stomach linings of the monkeys who were not given responsibility over shock delivery were discovered to be free of any significant lesions. Brady suggested that the stress associated with having responsibility and control led to the development of fatal ulcers in the executive monkeys.

Other researchers had difficulty replicating these findings. Ultimately, it was discovered that a participant variable was responsible for the ease with which Brady's executive monkeys developed ulcers. When designating which of the two monkeys would be assigned the executive role, Brady selected the monkey who learned how to press a lever to avoid shock the fastest. Subsequent research has revealed that animals can be selectively bred who are susceptible to the development of ulcers (Sines, 1959). One characteristic of ulcer-susceptible animals is that they are more emotional and, as a result, learn avoidance responses more quickly (Sines, Cleeland, & Adkins, 1963). Without realization, Brady had inadvertently assigned the more emotional monkey to the executive position, thereby confounding a participant variable with the treatment condition and rendering the study uninterpretable. ◄

In the previous two research examples, the investigators could have controlled for participant variables by applying a randomization procedure. If the participants had been arbitrarily assigned to the conditions, participant variables would have been eliminated as confounds. However, in both cases the

researchers applied a rule to help create the two groups. (Example 1.1: First students to arrive. Example 1.2: First monkeys to learn how to press a lever to avoid a shock.) This is always dangerous because it introduces a systematic approach to creating the groups (see Box 1.1 for more on this important topic).

As noted earlier, the technique of randomization can also be used to control differences related to the setting in which an experiment takes place. For example, suppose a drug manufacturer wanted to see if their newest product to control blood pressure worked better than the current leading medication. When testing this, the manufacturer would be wise to not only control participant variables by random assignment but also control setting differences by a randomization procedure. For example, if the pharmaceutical company chose to only use their new drug on patients being treated in small private hospitals and then compared their numbers with outcomes of patients from large public hospitals who were being treated with the leading medication, might a critic of the research properly suggest that the level of care may be different in these two types of hospitals, thereby confounding the study? Randomly assigning not only the participants but also the study setting will solve this problem.

Random assignment can also be used to control for different stimuli used in a study. For example, in the first illustration, students solved math problems and were timed for speed. We could design the study to hold the specific math problems constant and have students in both groups working to solve the same ones. However, suppose there was a problem with that arrangement. Imagine there was not enough privacy in the classroom and students could see the problems and answers on other students' papers if they glanced around. To neutralize this confounding possibility, the selection of the problems themselves could be controlled by randomly assigning math problems from a pool of acceptable problems to the individual pages being handed out to the students, thereby making each page of problems unique and different from the others. We might be concerned that some problems are simply easier to solve than others, but even if this were the case, it must be remembered that the ratio of easy to hard problems will tend to even out across the pages, as the list of problems to be solved grows and as the number of students solving problems grows. Any aggregate differences in difficulty between the pages of problems to be solved would not constitute confounding error, but rather the kind of error that statistical techniques are designed to handle.

There are techniques other than "holding constant," "balancing," and "random assignment" that can be used to control extraneous variables, but these are the ones most often employed. To learn more about controlling extraneous variables, a textbook on research methodology can be consulted (e.g. Marczyk, DeMatteo, & Festinger, 2005). Failure to control all extraneous variables coupled with a failure to recognize or acknowledge this lack of control can lead to the spread of misinformation. The box below discusses how this very topic plays a vital part in the current conversation surrounding the erosion of public trust in science.

Box 1.1 Is the Scientific Method Broken? The Wallpaper Effect

The public has recently been informed about a troubling discovery in the world of social and medical science investigation – the nonreproducibility of many scientific findings. Titles like "Scientific Regress" (Wilson, 2016), "Does Social Science Have a Replication Problem?" (Tucker, 2016), and "Over Half of Psychology Studies Fail Reproducibility Test" (Baker, 2015) seem to be popping up all over the place. The titles are unnerving and the issues that are raised are both real and serious. Briefly stated, an alarming amount of published research does not produce, when attempts at replication of the study are performed, the same findings that were stated in the original publication. Upon learning this, it is quite natural for people to wonder if there is something fundamentally wrong with the scientific method or the conduction of scientific research.

There are many aspects to this problem. This box attempts to analyze one such aspect. (Several others will be featured in boxes found throughout the rest of the textbook.) Many suspect that a leading reason for replication failure is the inability of the replicating researchers to reproduce accurately, fully, and completely the precise conditions under which the original study was performed. Even though the replicators faithfully follow every aspect of the methodological situation as recorded by the original researchers, it is proposed that there are other features of the original study that perhaps went unnoticed or were simply not deemed adequately important to note in the write-up of the initial experiment, but nonetheless may have partially determined the recorded outcomes. Perhaps an extraneous variable thought to be negligible or to have been controlled was, in actuality, not controlled and influential. These differences are colloquially referred to as "wallpaper effects" – a tongue-in-cheek way of suggesting that the original experimental situation must have been influenced by the color of the wallpaper in the room. If this were the case, then the original study findings and the replicated study findings may not be in genuine conflict with each other; both may be accurate reflections of reality. The different outcomes, then, would be explained by the effect of some, as of yet, unidentified other variable responsible for causing the difference. Unfortunately, because this subtle influential variable has not been identified, neither finding can be interpreted with any confidence. The interpretations of these studies, then, are best left as incomplete – to be explained at some later time when further light can be shed on the topic.

Look for other boxes in this series (Is the Scientific Method Broken?) peppered throughout the textbook. Each addresses a different aspect of the nonreproducibility problem.

1.5 Validity Issues

Internal Validity

Much of this chapter has focused on the importance of and the means by which definitive conclusions can be made regarding the effect of the independent variable on the values yielded by the dependent variable. As we might imagine, some studies do a better job than others of making this connection clear and unambiguous. We use the term **internal validity** to capture this idea. That is, to what degree can one unambiguously attribute changes in the dependent variable exclusively to the action of the independent variable? As our certainty grows, internal validity goes up. As we realize more and more competing explanations for changes in the dependent variable, our internal validity goes down (Shavelson, 1988). We might want to think of the internal validity of a study as a measure of how logically "tight" and "tidy" are the inner workings of the research effort. Various kinds of reasons can account for a lack of internal validity. For example, real-world constraints may impose inescapable limitations on the experimenter's ability to fully control all extraneous variables, properly manipulate the independent variable, or carefully measure the dependent variable. All causes of diminished internal validity are not however inevitable. There are also many avoidable causes that are actually just instances of poor methodological thinking or poor methodological execution.

Thankfully, we can often learn from our mishaps. In fact, frequently the critique of an initial study can reveal an unconsidered rival explanation for the findings, and this, in turn, will open up the door to a new and fruitful study. What was found to be a confounding variable in the initial experiment can then be used as an independent or dependent variable in a follow-up study. For example, let us revisit our research on the relationship between mental imagery and pain tolerance, but now we will introduce the variable "anxiety." Evidence from past research suggests that highly anxious people tolerate pain poorly (Barber & Hahn, 1962). Suppose that in our study we unintentionally varied anxiety systematically with the imagery conditions. For example, perhaps close analysis of the experiment would reveal that we had participants in the second imagery condition waiting longer in the reception room before their pain tolerance was measured, compared with those in the initial imagery condition. This delay in time could have allowed more anxiety to build. When a reviewer alerts us to this potential confound, not only could we rerun the study correcting this confound, but we might also decide to add anxiety as a new independent or dependent variable to explore.

As previously noted, a tightly controlled study, one having high internal validity, can provide the researcher with evidence that a cause–effect relationship exists between the independent and dependent variables. Conversely, studies that are not tightly controlled and have low internal validity will not justifiably merit the claim of causality. Box 1.2 presents a published study in which a

Box 1.2 Feeling Good and Helping Others: A Study with a Confound

The topic of generosity and helpfulness has long been a popular topic in psychology. Discovering the circumstances that encourage and discourage helpfulness increases our understanding of this important behavioral phenomenon and may suggest ways in which we can facilitate prosocial behavior for the betterment of society (Kanfer & Grimm, 1980). Some of the factors that influence helpfulness include the observation of a charitable model (Rosenhan & White, 1967), the relationship between the helper and the recipient (Goranson & Berkowitz, 1966), a predisposition to value the welfare of a person in need (Batson, Eklund, Chermok, Hoyt, & Ortiz, 2007), and past help received by the would-be helper (Berkowitz & Daniels, 1964). Based on an intuitive formulation (i.e. a hunch), Isen (1970) predicted that the positive feelings one experiences after success (the warm glow of success) would promote generosity. Participants were randomly assigned to experimental conditions in which half of them received success feedback after completing perceptual–motor tasks and the other half were told that they had failed at the tasks. After the experimental manipulation (success or failure feedback), a confederate entered the room and casually placed a canister on a nearby table for donations toward a school project. (A confederate is someone who, unbeknown to the participant, is really part of the experiment and has a prescribed role to play in the study.) The dependent variable was the amount of money participants donated. Consistent with the hypothesis, those participants experiencing the positive feelings of success donated almost twice the amount of money as did those who had been told they had failed at the task. What is the confound?

The researcher asserted a causal connection between a positive emotional state and generosity. The question to ask is: "Did the experimental manipulation (success/failure feedback) *only* alter the emotional states of the participants?" The answer is likely "no." Participants' self-perceptions of competence were probably also altered by the success or failure feedback. "Success" participants were not only in a better mood than "failure" participants but may have also seen themselves as more competent than those who failed. And so, was it the participants' emotional state that determined generosity (as the researcher intended to show), or was it their perceptions of competence? The variables *competence* and *emotional state* were confounded, each varying systematically with the other. This created a problem when it came to interpreting the finding. Thankfully, in a subsequent experiment, the researcher was able to induce a positive mood in participants through an experimental procedure that did not affect self-perceptions of competence, thereby isolating "mood" as a causal variable. The results of this second study showed that a positive affective state did enhance helpfulness (Isen & Levin, 1972).

confounding variable reduces the internal validity of the study and presents a rival explanation for its results.

When real-world constraints do not allow for full control of all variables, a researcher may still choose to move forward with a study, even though the internal validity may be compromised. For example, a professor who wanted to see if a new teaching tool (e.g. a new type of review game) was effective in the classroom might expose one section of the class to this new teaching tool but leave another section unchanged and then compare test performances over the relevant material. Please note that students were not randomly assigned to the classes rather they signed up for the class they wanted or that best fit into their schedule. This opens up the possibility that one class may represent a different type of student population than the other. Imagine that perhaps one class meets at 8:00 in the morning and the other at 2:00 in the afternoon. These classes, although they are both composed of students from the same institution and may have many other similarities, may represent quite different subgroups of students, namely, those who prefer morning classes and those who prefer afternoon classes. Furthermore, during the course of the experiment, each class may experience other unique events that only occur in that class (e.g. a fire drill, problems with the classroom technology). Since these other events are not the independent variable, they now compete with the new teaching tool to explain any difference in the outcomes between the classes. These are not ideal experimental situations; however, oftentimes they are the only viable method available to a researcher attempting to establish a causal relationship. Such designs are called **quasi-experiments** because they share many of the same characteristics with true experiments and yet the participants are not randomly assigned to conditions.

External Validity

A researcher must also analyze the extent to which the experimental findings can be justifiably generalized, thereby reaching beyond the limited context of the study. After all, no one would be interested in learning about the relationship between mental imagery and pain tolerance if it is believed that the findings only pertain to those participants involved in the study and to only the specific relationship between a particular form of mental imagery and the act of holding ones' hand in ice-cold water. Obviously, the value of the study only emerges when we consider the finding in a more generalized context – people in general and pain tolerance in general. The degree to which our study legitimately applies to these broader external categories is of critical importance. Stated formally, "**External validity** asks the question of generalizability: To what populations, settings, treatment variables, and measurement variables can this effect be generalized" (Campbell & Stanley, 1963, p. 5). The external validity of a study can be very difficult to judge and is often subject to intense professional debate.

Strictly speaking, there is no way, without running numerous other studies, to determine whether the results of a research study would be replicated if the experiment were conducted with different participants, in a different location, using a slightly different independent variable or measuring a slightly different dependent variable. Thankfully, there are ways to think about these applicability issues that do not require the running of an endless series of near-identical studies.

Let us look at the problem of different participants first. One's confidence in generalizing to others *not* involved in the study can be increased by the method used to select participants for the study. **Random sampling** is the "gold standard" for sampling participants. It occurs when the means of selecting participants for a study is such that each participant in the population has an equal chance of being included in the sample. (Note that this is different from "random assignment." Random assignment is a technique used to assign already selected research participants to the various conditions of a study. Failure to assign randomly affects the *internal validity* of the study, not the external validity.) A **biased sample** may occur whenever each member of a population does *not* have an equal chance of being included in the sample.

Before moving on, let us clarify a few new terms that have just been introduced: populations, sampling, and samples. A **population** can be simply defined as "every member of a given group." For example, imagine we want to study the effectiveness of a new teaching technique designed to help adults learn how to read English. We can label our population, then, as *adults who are engaged in learning how to read English*. Of course, we cannot include every single person in this entire population in our study; rather, we will select a subset of individuals from that population to investigate. The process of selecting participants is called **sampling**; and the group of individuals, once selected, is called a **sample**.

Let us return to the issue of participant variables. Within any given population, there will be a host of participant variables distinguishing one participant from another. Just as *random assignment* creates groups within a study that are roughly similar across all of these participant variables, *random sampling* creates a group of participants that roughly captures the blend of participant variables found in that population. This is important because it allows the researcher justifiably to claim that findings obtained from their study should also pertain to the population as a whole. This gives us good reason to presume our findings are externally valid and applicable to other participants in our population. Conversely, if a sample is not randomly gathered, certain participant variables may be overrepresented, and other participant variables may be underrepresented. When this is the case, confidence regarding the external validity of the findings falters. For example, if we tested our new teaching technique only on new immigrants coming from countries using the Roman alphabet (the same letters we use in writing English), we should be cautious about concluding that

our findings will pertain to immigrants coming from places where alphabets employing different characters (i.e. Chinese characters) were used. Our sample does not "represent" adequately this larger population. Logically, if this were the case, we should realize that the population we have actually sampled from is not *adults learning how to read English*, but rather *adults learning how to read English who came from countries using the Roman alphabet*. Clearly, this is a much smaller population, and the external validity of our findings necessarily becomes more suspect as we extrapolate to people outside the population from which we sampled.

It is important to realize that even though random sampling is the gold standard for achieving representative samples, rarely is it employed. Take the example we used above: How could one perfectly select a truly random sample of *adults learning how to read English*? While the population can be easily described in the abstract, it is functionally impossible to gain access to it for the purpose of random sampling. Fortunately, many other less-than-ideal-but-more-or-less-adequate sampling procedures can be employed when doing social science research. The specifics of these techniques are usually treated with greater detail in methodology textbooks. In a similar vein, there is the issue of sample size. What is the minimally acceptable number of participants necessary for a sample to be considered representative of its parent population? This can be a complex determination and is customarily discussed with greater precision in textbooks focused on research methodology.

Let us now return to the concept of external validity regarding slightly different locations, independent variables, and dependent variables. Determining the external validity of findings in these situations is much more a matter of reasoned argumentation than one of calculating probabilities and likelihoods. For example, sometimes the findings generated by a study will naturally prompt researchers to think of other similar settings and variables for which these findings might apply. A good example of this is a study by Baddeley and Longman (1978) that compared mass practice with distributed practice for learning the new skill of typing. The basic question was this: How is practice time best spent if one is learning to use a typewriter with a new arrangement of keys? One group practiced in mass (i.e. long practice sessions within a short window of time), while another group distributed their practice (i.e. shorter practice sessions spread out over a longer stretch of time). Both groups actually spent the same total amount of time typing on the new keyboards. The findings showed quite convincingly that those who had a more distributed set of practice sessions ended up typing more quickly and with fewer errors. Many theorists correctly suspected that Baddeley and Longman demonstrated a more general principle regarding the relationship between mass and distributed practice (for example, see Box 1.3). The external validity for Baddeley and Longman's findings appears to be high.

The external validity of a particular finding is best judged by examining the body of existing research to see if a similar finding has occurred with

Box 1.3 A Strategy for Studying Statistics: Distributed over Mass Practice

Remember the research on mass versus distributed practice by Baddeley and Longman (1978) that was mentioned earlier in the chapter? The finding that "distributed practice is more effective than mass practice" is actually a very old one, first uncovered by one of the pioneers of psychology, Hermann Ebbinghaus (1885), when conducting his famous studies on memory. It is also one of the more *robust* findings of psychology shown to apply to the acquisition of a variety of both neuromuscular skills like archery (Lashley, 1915) and cognitive abilities like face recognition (Mammarella, Russo, & Avons, 2002). Karl Lashley (1915) concisely states the core finding when writing, "[a] close correspondence exists between the distribution of practice and the amount of improvement appeared, a given amount of practice being more efficient when distributed through many short periods than when concentrated in a few long ones" (p. 127). Given the demonstrated *external validity* of this finding, the implications for studying material in this text should be clear. It is far better to spend a small but meaningful amount of time each day studying statistics than it is to set aside a few large blocks of time to cram just before a test. The challenge to the student is to discipline themselves to find time every day, or at least every other day, to review recently covered material, run through a few practice problems, and read a little bit ahead in the textbook. Research suggests it will be the best way to spend the time one allocates to learn the material in this textbook.

participants from different populations using different experimental procedures and different measures of the dependent variable. Findings that hold up under a wide range of circumstances (like the "distributed over mass practice" finding) are termed *robust*. Oftentimes a researcher, who has uncovered an interesting finding, will go on to conduct a series of related studies in which they systematically alter populations, settings, and related variables in order to establish the robustness of their finding. This is oftentimes referred to as conducting a *line of research*.

Internal and external validity are only two among many different kinds of validities. They are all important to the research process, and much fuller treatment of these concepts can be found in various methodology textbooks.

1.6 Causality and Correlation

Earlier (in Section 1.2), the limited way in which social scientists think about causality was mentioned. Let us develop this a bit more here. To say that "*X*" *causes* "*Y*" in the behavioral sciences is *not* to suggest that *X* is the necessary

and sufficient reason for *Y* coming about. After all, there may be many other reasons why *Y* has occurred or can change in amount. When we say "*X* causes *Y*," we are merely saying that if we place *X* in this particular situation, we will get more *Y* than if we had not placed *X* there. This understanding of causality is indeed helpful, but it is clearly limited. Saying that the utilization of a mental imagery technique can increase one's pain tolerance is not saying that it is needed for any amount of pain toleration, nor it is to say that this is the only means of changing pain toleration. Rather, this causal statement is merely saying that if one uses a certain form of mental imagery (*X*), one can expect to experience more pain tolerance (*Y*) than if one had not. Finally, causality statements in the behavioral sciences do not necessarily imply that anything whatsoever is known about the much more profound question as to *why* reality is such that if we place *X* in this situation, we will get more *Y*.

Nonetheless, the term *cause*, if understood modestly, can be appropriately used in the behavioral and social sciences. However, other modest phrases are also employed to designate the causal influence of one variable upon another. For example, some causal relationships might be described in phrases like "*X* increases the probability that *Y* will occur" or "*X* tends to bring about a change in the occurrence of *Y*." In fact, many different action verbs can be found in the social and behavioral science literature.

Scientists highly value discovering causal relations because they help to bring about a deeper sense of understanding. Recall that this is the highest goal of the researcher (see Section 1.2). It is critical to see, however, that *causality* is not the only way in which a relationship between two variables can be described. If two different variables tend to change reliably in association with each other, they are said to *covary*. Variables that covary are *correlated* variables. This more limited description of relationship meets the second goal of the researcher: correlation (or prediction, association). Height and foot size, for example, tend to covary. They are correlated: taller people usually have larger feet, and smaller feet tend to be attached to shorter people. When data are gathered nonexperimentally (i.e. without manipulation), two variables may be observed to covary, and this covariation can be quantified. And yet, this covariation does not imply that a causal relationship exists between the variables, let alone the exact nature of a causal relationship (for instance, might *X* cause *Y*, or *Y* cause *X*, or might some other variable, *Z*, cause both *X* and *Y*?). *The manner in which the data are collected will determine the type of interpretation allowed*. Causal relations can only be established in an experimental setting when a manipulated variable is observed to influence another variable. Methods of gathering data without the use of manipulation are called **correlational designs**. (These designs as well as experiments, and quasi-experiments, comprise the three most frequently used designs that employ statistical analysis.) A prototypical example of a correlational design would be the survey. There is no independent variable, nothing is being manipulated, and no causal statements can be made. Data is gathered

simply as it presents itself to the researcher. In this way, it is correct to say that data gathered from "correlational designs do not imply causation." (This topic will be covered in much greater detail in Chapter 15.)

Research on clinical depression demonstrates how variables can be found to covary without knowing the precise causal relation between them. Depression has many characteristics, and the reasons why people become depressed are many: The phenomenon is not exhaustively understood. One view maintains that people feel depressed because of negative thinking. They are pessimistic, are self-critical, and do not praise themselves when they do something well. This perspective strongly implies that these cognitions have some causative role in depression. However, it is also quite possible that when people become depressed, they are more likely to think in a negative fashion. In Figure 1.1, question marks reflecting this interpretive problem are drawn above the arrows between negative thinking and depression.

However, even though negative thinking and depression may correlate, it is possible that neither variable causes the other. They may covary because some third variable, like "loss of control" for example, causes both negative thinking *and* depression. The question marks over the arrows in Figure 1.1 pointing from loss of control to negative thinking and depression reflect this possibility.

Unfortunately, some important research questions are, for ethical, logical, or logistical reasons, not amenable to experimentation. The relationship between negative thinking and depression is a good example. Even if we discovered a means by which researchers could manipulate "depression," it would seem to be unethical to do so. Likewise, it is hard to imagine logistically how one could manipulate, for a sustained period of time, the degree of negative thinking a participant experiences. Some questions, it appears, are restricted to merely a correlational analysis.

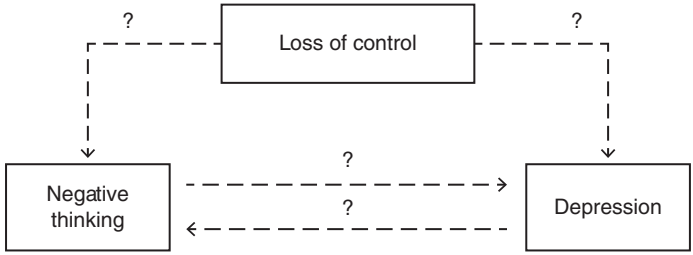


Figure 1.1 "Negative thinking" and "depression" are correlated, but which one causes the other? It is also possible that a third variable, "loss of control," causes both "negative thinking" and "depression."

1.7 The Role of Statistical Analysis and the Organization of the Textbook

The role for statistical analysis may not yet be very evident from this brief overview of the basic concepts of research. Jones (2015), in fact, suggests there are 9 discrete steps to the research process, only one of which involves statistics. According to Jones the sequential steps are (1) topic selection, (2) literature review, (3) theoretical/conceptual framework development, (4) research question/hypothesis clarification, (5) setting the research design, (6) data collection, (7) data analysis, (8) drawing conclusions, and (9) disseminating results.

It is true that the research process is, in fact, much bigger than a statistical analysis; and this first chapter has helped, hopefully, to underscore this point. That being said, the statistical process is nonetheless crucially important. There is simply no way to go from step 6 (data collection) to step 8 (drawing conclusions) without proper and careful statistical analysis. Despite the limited role for number-crunching analyses, they are absolutely necessary to make sense of any study based on data. Additionally, although the actual statistical work is restricted to step 7, statistical knowledge will help the researcher in steps 5 and 6. In terms of research design (step 5), just as a woodworker needs to know what their tools can and cannot do before they design a table, so the researcher needs to know what their statistical "tools" can and cannot do before they design a study. Similarly, when gathering data, awareness of the capacities of various statistical tools will influence which variables are measured and how. Knowledge of research methodology and statistical analysis will always walk hand in hand.

Once data has been carefully and properly gathered and the research situation is fully understood, the researcher will use statistical analysis to help make sense of the observations. **Descriptive statistics** are techniques designed to describe and summarize data in an abbreviated form. Frequency counts, distribution shapes, measures of centrality and variability, and standardized scores are all statistical tools that can be used to help describe the basic features of that which is being measured. Part 2 of the text is set aside to address these statistical concepts. Additionally, analyses that are more sophisticated allow the researcher to test hypotheses, quantify covariation between variables, and determine causal relationships by analyzing samples of data. This branch of statistical analysis is referred to as **inferential statistics** since it grants the researcher the ability to draw valid conclusions about the characteristics of populations from withdrawn samples.

The concepts found in Part 2 regarding descriptive statistics are foundational and must be mastered before inferential statistics can be addressed. Part 3 is composed of two chapters designed to introduce some important theoretical underpinnings for inferential statistical analysis, namely, elementary probability

theory and the basics of hypothesis testing and sampling distributions. Parts 4–7 cover four different groupings of inferential statistical analyses. Part 4 deals with the basic set of *z* and *t* tests as well as the important concepts of *decision errors* and *statistical power*. Part 5 examines three basic "analysis of variance" tests including the very important interpretative concept of *interactions*. Part 6 introduces bivariate data, as well as the fundamentals of some statistical techniques developed to analyze them: the Pearson correlation coefficient and linear regression. Part 7, the last part, addresses a common set of nonparametric tests including two different chi-square analyses. It is possible to rearrange the order of progression starting with Chapter 8 (Part 4); however some concepts needed for full comprehension of material in Parts 5, 6, and 7 are covered in Part 4.

Each chapter will conclude with a summary, a collection of all of the relevant formulas and keywords introduced in that chapter, and a series of questions and exercises to be used by students to practice their skills and for self-assessment. Each part (starting with Part 4) will conclude with additional work problems designed to challenge the student's ability to determine which of the previously covered statistical tests apply to a given research analysis situation. The appropriate analyses may be a test found in the present part or any of the preceding parts. This feature is included to address a fundamental problem with textbook learning, the tendency for end-of-chapter exercises to be solved by making use of only concepts found within that given chapter. By incorporating all the previously introduced concepts and tests, the exercises at the end of each part will better simulate and assess the student's ability to make sense of real-world problems through the application of an increasingly wider assortment of concepts and statistical tools.

The textbook addendums contain three appendices: The first is a collection of tables to help determine critical values for the various statistical tests, the second houses the answers to all of the work problems found at the end of each chapter and part, and the third holds the general instructions for how to use two software programs that are capable of running various statistical analyses. Finally, the text contains a glossary with definitions for all key terms, a list of references, and an index of keywords, terms, and concepts.

Summary

This chapter presented an overview of the basic concepts in research. The scientific method is a three-step cyclical process where theories lead to hypotheses that lead to observations that lead back to theories. Statistical analyses are used to interpret the observations in light of the theory being tested.

Researchers develop methods of studying topics of interest that are designed to answer a particular type of question. Some studies are designed to create a

better explanation of the basic features of the target of study. These are called descriptive studies. Other research methods, termed correlational studies, explore the strength of relationship between a variable of interest and other variables. Another subset of scientific investigations, marked by the controlled manipulation of one variable, allow researchers to gain a better understanding of what *causes* another variable to present itself or change. Studies that feature this controlled manipulation are called experiments.

Basic research terminology involves familiarity with different types of variables. Independent variables are the hypothesized causal variables within experiments; they are the variable manipulated by the experimenter. Dependent variables measure the hypothesized effect of the controlled presentation of the independent variable. All other variables are extraneous and must be controlled so as not to confound the study. Holding constant, balancing, and randomization are some of the mechanisms used to control extraneous variables.

The degree to which a study controls all extraneous variables is a measure of internal validity. Sometimes internal validity is sacrificed in order to study a variable that is not easily manipulated or controlled. These designs are described as quasi-experimental. Quite separately, external validity describes to what extent the findings of a study are generalizable to other populations, settings, and similar variables.

When researchers are interested in measuring the degree of relationship between two variables, they use correlational designs. These designs do not involve manipulating an independent variable, but rather the careful measurement of two variables as they present themselves to the researcher.

Although the role of statistical analysis is limited in the grand sequence of the research process, it is vitally important such that its absence would render most empirical studies meaningless. The remainder of the text will examine various descriptive and inferential statistical concepts and tools designed to help the researcher interpret data.

Key Terms

Theory
Hypothesis
Observation
Description
Operational definition
Correlation (or prediction or association)
Understanding
Experiment
Independent variable
Manipulation
Quantitative independent variable
Qualitative independent variable
Control group
Experimental group
Treatment
Dependent variable
Inferential statistics
Extraneous variable
Confounding variable

Holding constant
Balancing
Participant variable
Random assignment
Internal validity
Quasi-experiment
External validity
Random sampling
Biased sample
Population
Sampling
Sample
Correlational design
Descriptive statistics

Questions and Exercises

1 Identify each of the following phrases as a theory, hypothesis, or observation.
   - **a** 20 out of 25 basketball players improved their free-throw shooting percentage.
   - **b** More positive reinforcement by the professor will improve test scores.
   - **c** Human interactions are best thought of in economic terms; our actions seek to maximize gains and minimize costs.
   - **d** "Chickens" come before "eggs."
   - **e** More people choose Car A over Car B.
   - **f** When giving names in a circle, people will not recall ones given just prior to theirs.

2 Which of the following descriptions is the clearest operational definition for the concept "vacation?"
   - **a** An extended period of recreation, especially when spent away from home.
   - **b** Release from obligation, business, or ordinary activity.
   - **c** A stretch of time set aside to relax or travel for pleasure.
   - **d** The number of days in a year a person spends not working and away from home.

3 A researcher has just begun to start to study the "tiny house" trend in home construction. Think of two researchable questions that line up with each of the three different goals of the researcher: description, correlation, and understanding.

4 Identify the independent and dependent variables in the following four studies.
   - **a** A psychologist is interested in the effects of vitamin E on physical endurance. One group of participants receives 20 units of vitamin E, another 60 units, and a third gets a placebo. Endurance is assessed by the length of time participants can ride a stationary bicycle. (Also, how many levels are there of the independent variable?)

b A teacher evaluates the effectiveness of different educational programs on reading speed and comprehension.

c A social psychologist hypothesizes that attitude change will be greatest when people do not have sufficient justification to explain their counter-attitudinal behavior compared with when they do have sufficient justification. (Imagine a person opposed to watching a scary movie but who is compelled to do so by social pressure. In one condition, they are paid $1 as compensation and in the other $50.)

d An industrial psychologist hypothesizes that the amount of natural light in the work setting will increase productivity. For 15 days of a month, the blinds are drawn, and indoor lighting is the only source of light. For the other 15 days of the month, the shades are left open. Productivity is measured by the number of widgets made.

5 Imagine an experiment investigating the effectiveness of different rewards used by parents to "potty train" their children. Identify several different types of quantitative and qualitative levels for the independent variable. Also, describe how to use the technique of "holding constant" to control at least one potent extraneous variable.

6 Identify a confound affecting internal validity in the following four studies. (Hint: Some studies may not be confounded.)

a An independent marketing company has been hired to assess people's preference for A&W root beer versus Stewart's root beer. To prevent bias, all of the test cans are covered with paper, with the letter *A* placed on the A&W cans and *B* placed on the cans of Stewart's. The order is counterbalanced such that half of the participants experience A&W first and then Stewart's and the other half experience the other order. The results show that the participants prefer A&W over Stewart's by a 2:1 ratio.

b An experimental psychologist claims to have discovered an important cause of bizarre behavior. Laboratory mice are taught to discriminate between two geometric designs. The mouse is required to jump from a ledge through a trap door, which has one of the designs painted on it. If the mouse leaps through the door with the correct design, it lands on a table with food. If the wrong door is chosen, the mouse falls 3 ft onto a net. (Falling 3 ft onto a net may be fun for kids, but it is rather unnerving for a mouse.) Eventually all the mice in the study pick the door that has the correct design painted on it. A discrimination has been formed. To test the intelligence of each mouse, the researcher changes some aspects of the geometric designs so that they look much more alike than they did originally. Now the mice hesitate and many refuse to jump. To observe which door the mice will choose, the psychologist forces them to jump by

blasting a loud noise. Faced with such a difficult discrimination, the mice begin to exhibit unusual behavior. They run in circles, jump up and down, and fall into a catatonic state. It is concluded that the stress produced by having to choose between two very similar stimuli when the consequences of the choice are extremely important leads to abnormal behavior.

c A particular stretch of highway is noted for an excessive number of traffic fatalities. The city council decides to reduce the allowable speed limit, since evidence from national statistics clearly shows that traffic fatalities are correlated with speed limits. To make sure that the proper speed limit is observed, radar units are positioned every 5 miles along the highway. Not only did accidents significantly decrease, but also because of the increased surveillance, more motorists were following the speed limit. Obviously changing the speed limit has led to a decrease in accidents.

d Evidence shows that our reactions to pain are, in part, due to psychological factors. A dentist offers headphones with the patient's choice of music to listen to during procedures involving moderate discomfort. Since some patients may prefer the novocaine, that option is also made available. Patients are not allowed to listen to music *and* use the anesthetic. Patients are free to choose which method they want. At the end of the study, the dentist finds that those patients using the headphones reported less anxiety and less pain than those patients who opted for the novocaine. To address a potential confound, the dentist went back and checked records to see that the type of dental procedures was, on average, similar between the groups.

7 A popular theory of emotion asserts that we label our emotional states based on the perception of our own physiological arousal and the situation within which we find ourselves. However, one could question whether the presence of physiological arousal is really necessary. Maybe all that is required is the *belief* that we are aroused. A study is conducted with biological male undergraduates where 20 slides of different biological females are sequentially presented for 30 seconds each. The participant wears earphones and hears what is *believed* to be their own heart rate; but in fact, it is a recording. The participant hears an increase in heart rate for some slides and a decrease in heart rate for other slides. The assignment of heart rates to pictures is random and different for each of the numerous biological male participants involved in the study. The dependent variable is the participants' ratings of attractiveness made after each slide. The psychologist finds that the females observed when the tape-recorded heart rate was high were perceived as more attractive than the females viewed during decreased heart rates. Therefore, belief in arousal influences perceptions of emotion. Does this experiment have a threat to internal validity?

8 "Controlling extraneous variables" is to "generalizability" as:
- **a** Independent variable is to dependent variable.
- **b** Internal validity is to external validity.
- **c** Correlational design is to experiment.
- **d** Random sampling is to random assignment.
- **e** Population is to sample.

9 Professors may exhibit a good deal of subjectivity when grading papers. For instance, some prefer title pages, while others do not. Suppose we conduct a study in which we obtain the grades for a paper submitted by everyone in class. Since this professor does not specify whether the assigned paper requires a title page, we are able to find a similar number of papers that do and do not have them. Our results show that the average grade for the "title-page" papers is higher than for papers not containing them. Should we conclude that our professor prefers "title-page" papers? What other interpretation of the results can we make?

10 Which of the following means of assigning participants to two experimental conditions in a psychological study represent "random assignment?"
- **a** Flipping a coin right before each participant is due to arrive at the lab.
- **b** Looking at a students' ID card and putting numbers that end in an odd number in one group and those that end in an even number in the other.
- **c** Assigning students who signed up in class to the experimental condition and those that signed up online to the control.
- **d** Asking each student to choose which color they prefer – those that choose "red" go to one condition and those that choose "blue" go to the other.
- **e** Asking participants their age. Even ages go to one group, and odd to the other.

11 Which of the following selection procedures represent the "random sampling" of students at a university campus?
- **a** Standing outside of the cafeteria and asking every third person who walks in if they will participate in a study.
- **b** Asking the registrar to give us the name of every twentieth person going alphabetically through the university enrollment list.
- **c** Sending out an email to the entire campus asking for volunteers.
- **d** Getting permission from our professors to stand up in front of class to ask our fellow classmates.
- **e** Cutting out each name from the university directory, putting them into a basket, and then blindly drawing names.

12 For each of the following research questions, design a study that addresses the question with a controlled experiment. Redesign the study using a correlational methodology. (Hint: It is possible that one or more questions cannot be investigated by controlled experimentation.)

a Is there a relation between pain and anxiety? (For the experimental design, pain is the independent variable.)

b Is there a relation between how often a person exercises and resting heart rate?

c Is there a relation between need for achievement and hours worked per week?

d Do children who attend preschool day care show better social skills in first grade?

Part 2

Descriptive Statistics

2

Scales of Measurement and Data Display

2.1 Scales of Measurement

The scientific method requires the same variables that are referenced by theories and operationalized in hypotheses be carefully measured when they are observed. This is how the process of science comes full circle; the carefully measured observations shed light on the validity of the theory being examined. **Measurement**, simply stated, is the assignment of numbers to attributes, objects, or events according to predetermined rules. A proper understanding of the different sets of rules, or scales of measurement, is required to make sense out of what particular numbers mean within a given context. Four different measurement scales needed for statistical analysis will be presented in order of the amount of quantitative information they convey.

Nominal Scales

Our first scale conveys no quantitative information. A **nominal scale** uses numbers merely to distinguish one type of thing from another type of thing or one event from another event. For instance, the numbers assigned to the members of a soccer team do not carry any quantitative value. They merely distinguish one player from the others on the team. Using a "1" for biological males and a "2" for biological females on a spreadsheet is not meant to suggest that females are somehow or in some way *more* than males. Or, think about dividing people into three groups by having them line up and count off: one, two, three, one, two three, and so on. The numbers used in these examples do not represent quantitative differences between groups, but rather merely qualitative ones. Since there is no quantitative information being communicated, we are free to exchange one number for any other currently unused number. For instance, a midfielder who does not like the number 8 can exchange it for 6, provided

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

it is not currently being used. Since the numbers on a nominal scale carry no quantitative value, it makes no sense to find the average or range of all the jersey numbers on a sports team. Yes, we could find an average or range, but the resulting value would not be related to anything meaningful about the team.

Ordinal Scales

Similar to the nominal scale, ordinal measures also categorize things; however, ordinal numbers additionally reflect a quantitative relationship between the various categories. Stated more succinctly, an **ordinal scale** consists of a set of categories organized quantitatively. The degree of quantitative information communicated, however, is very limited – merely the *relative position* of one event compared to others. For instance, observing that "Jim is less trustworthy than Pam," "Corvettes are faster than Accords," and "Muhammad Ali was the greatest of all time" are all measurements of relative position.

It may also help to think of the commonly used notion of *ranking*. Rankings reflect more or less of something, but not *how much* more or less of something. The difference between the winner and runner-up in a pie-baking contest may not be the same amount as the difference between the fourth- and fifth-place finishers. In other words, the quantitative intervals between adjacent ranks are not held constant over the entire range of the scale; in fact, the various interval quantities may not even be known. Consider the house numbers on a neighborhood street. The numbers mark merely the relative distance a house is from a given point, say, the center of town. If we are standing at a house marked 105 Maple, we know that 123 Maple is closer to us than 157 Maple, but we do not necessarily know how much closer. Furthermore, the house numbering system could be altered by the city planners if they so wished. The newly assigned numbers will simply need to increase as the houses sit farther from the center of town. Relative position information is all that an ordinal scale conveys.

Interval Scales

The next step-up in quantitative information is the interval scale. An **interval scale** consists of a set of quantitatively ordered categories but for which all of the intervals between the categories are held constant (or "conserved"). A good example is the Fahrenheit temperature scale. The amount of heat needed to add to a room to move it from 80 to 85 °F would be the same amount of heat needed to move it from 90 to 95 °F. At each point along the scale, a degree is a degree. However, interval scales do not possess a *true* zero point. That is, the value "zero" is just an arbitrary point on the scale and not the absence of quantity. A room that is measured at 0 °F is not devoid of heat; it could be made colder. It is also important to realize that any given interval scale does not have

a monopoly on how the variable is to be measured. The degree of heat in a room, for instance, can be measured using a Celsius scale. Even though the Celsius scale possesses larger intervals and has a different zero point, both have conserved intervals across the length of the scale.

The distinction between ordinal and interval scales, however, is not as straightforward as it may appear, especially in the behavioral and social sciences. In these areas of study, many scales have been constructed to measure various psychological concepts – like intelligence, pain, extroversion, authoritarianism, and so on. The problem with these scales is the difficulty in determining if the intervals are conserved. For instance, is the intelligence difference between individuals with IQ's of 100 and 105 the same as the intelligence difference between individuals with IQ's of 45 and 50? The numerical distance between the scores is the same, but is the difference in the *amount of intelligence* between the two sets of scores the same? This is not an easy question to answer. Yet, it is critical because many statistical tests require data to be measured on a scale possessing equal intervals across the entire continuum. This is a necessary scale feature for the important statistical concept of an "average" to be meaningful. This thorny issue will be discussed more below in Spotlight 2.1 and in Chapter 3 (see Box 3.1).

Spotlight 2.1 Rensis Likert

Use the scale below to respond to the following statement: I enjoy studying statistics.

○	○	○	○	○
1	2	3	4	5
Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly agree

If we have ever had to respond to a question in this manner, we can thank the social scientist, Rensis Likert. Likert, born in 1903 in Cheyenne, Wyoming, first began his undergraduate studies in 1922 in civil engineering, but then soon discovered that he preferred to study people instead of inanimate objects (Faculty History Project, 2011). Over the next ten years, he studied a variety of disciplines including sociology at the University of Michigan and theology at Union Theological Seminary, culminating with a PhD in psychology from Columbia University. In his dissertation, he presented a measurement tool he developed to assess how people felt about various topics related to international affairs. This scale, which asked the respondent to place their attitude on a 5-point scale of favor to disfavor with a neutral midpoint, was a much simpler procedure

compared with the currently preferred but cumbersome method developed by another important psychologist named Leon Thurstone (Croasmun & Ostrom, 2011). Since Likert's scale yielded similar results to Thurstone's and yet was much easier to use, Likert's procedure soon became the method of choice for measuring personal attitudes on almost any topic across the behavioral and social sciences. However, there was a price to be paid for the scales' simplicity; it was unclear if these sequential categories from 1 to 5 were best understood to be characteristic of an ordinal scale or an interval scale. The debate still goes on today and will be discussed more in Box 3.1.

In 1939, now working for the US Department of Agriculture (USDA), Likert began to use his scale to measure farmers' reactions and attitudes toward a variety of President Roosevelt's New Deal programs that were sponsored by the USDA (Kish, 1990). Gaining information about people's perceptions is extremely important in a democratic system where those who are governing are beholden to the people they govern. Likert realized this, and over time his measurement efforts gathered a lot of interest and attracted many professionals from a variety of other disciplines who were, likewise, interested in assessing the personal attitudes of people regarding all sorts of policies, products, and perspectives. As the United States entered World War II, his research and measurement team became involved in the construction of several national surveys designed to measure American's attitudes toward many war-related government efforts like the selling of war bonds, the instituting of price controls, and rationing.

After World War II, Likert and several colleagues, eager to apply their measurement methods to new problems in a postwar world, started what is now called the Institute for Social Research at the University of Michigan (see http://home.isr.umich.edu/). Despite the never-ending challenge to secure funds for research support, Likert's unfailing optimism that evaluation research would become increasing recognized as a needed commodity by academic, commercial, and government organizations was soon shown to be accurate. His institute grew rapidly, in just a few years becoming the largest university organization of its kind.

In addition to Likert's measurement insights and administrative and organizational talents, he was also a theorist. For instance, he long held an interest in the study of management styles. While at the Institute for Social Research, Likert proposed a theory of management-style development arguing that it culturally evolved through a series of four stages, from "exploitable authoritarianism" through "benevolent authoritarianism" and "consultative management," finally arriving at "participative management" (Hall, 1972). He was a very prolific writer and thinker authoring several books and over 100 published articles. Looking over the totality of his life, it is easy to conclude that his influences on the field of attitude measurement as well as his writings on business, industry, and management have been unquestionably significant, and yet perhaps ironically, a bit hard to measure.

Ratio Scales

A **ratio scale** possesses all of the properties of an interval scale, with the addition of an absolute zero point. For instance, one could argue that the Kelvin scale of heat measurement is a ratio scale based on a theoretical state of no energy: 0 °K. A measure of length is a ratio scale since there is an absolute zero point (i.e. a point of no length). With an absolute zero point, any number on a ratio scale can be used to establish its standing relative to any other number on the scale; in addition, a given number also represents an absolute amount of something. Could "intelligence" be measured on a ratio scale? No, because an IQ of 0 is meaningless. Furthermore, it would require us to claim that someone with an IQ of 100 is twice as smart as someone with an IQ of 50. However, because the underlying dimension of height has an absolute zero point, a person whose height is 80 in. is indeed twice as tall as someone whose height is 40 in. Time is another variable that is often measured with regard to an absolute zero point. A participant's reaction time of two seconds is twice as slow as another participant's reaction time of one second. Difference scores using interval scales are also ratio measurements. For instance, the increase of heat in a room from 50 to 55 °F is half the size of an increase from 70 to 80 °F. Even though the foundation scale is interval, comparisons of change or difference numbers from that scale are ratio.

In the behavioral and social sciences, many of the concepts researchers measure use either ordinal or interval scales. For instance, there are no measures of achievement, aptitude, personality traits, or psychopathology that have a meaningful absolute zero point. On the other hand, studies that investigate performance often use a ratio scale – the number (or percentage) of correct answers, the number (or percentage) of errors, and the amount of time needed to complete a task are all variables measured on scales with constant intervals and with a "zero" marking the absence of quantity.

2.2 Discrete Variables, Continuous Variables, and the Real Limits of Numbers

Discrete Variables

Another important feature of measuring variables concerns how many different values can be assigned. A **discontinuous** (or **discrete**) **variable** can take on only a finite number of values. No meaningful values exist between any two adjacent values. For instance, an undergraduate student is a freshman, sophomore, junior, or senior; an adult is single, married, divorced, or widowed; and a roll of a typical die yields a one, two, three, four, five, or six; there are no "in-between" possibilities. One cannot claim to be halfway married or roll a die and hope to

get a 4.5. It is permissible, however, to find statistical features of sets of discrete data, even if the number produced is not itself an acceptable value. For instance, it may be true that the average American family size is 2.58 persons, even though a 0.58[1] person is not possible. We just need to keep in mind a correct interpretation of this statistic. In this instance, it means that for every 100 American families, there are, on average, 258 people.

Continuous Variables

A **continuous variable** can theoretically have an infinite number of points between any two numbers. Unlike discrete variables, continuous variables do not have gaps between adjacent numbers. Although 7 and 8 cm may be adjacent options on a ruler, there are an infinite number of values between them. Even if the scale is only marked to the millimeter, there are still an infinite number of values between 7.3 and 7.4 cm. When the underlying dimension of a scale is continuous, any number on the scale is an approximation. Even though we could measure someone's reaction time down to the tenth of a second, this measurement could still be refined with a more precise instrument. Therefore, one can always theoretically increase the precision of measurements for continuous variables. This is not the case when measuring discrete variables. Greater measurement precision will not alter the fact that, for instance, family members exist in whole numbers.

Detecting the continuousness of a variable is not as simple as looking at how it is reported. Age, for instance, is often reported in whole-number years, but the underlying dimension is clearly continuous. The same point can be made with respect to psychological measures. Suppose a psychologist administers an anxiety questionnaire in which scores can range from 16 to 30. Although the measuring tool may only allow the assignment of whole numbers, the underlying concept is arguably continuous.

The Midpoint of an Interval and Real Limits of a Number

If a variable is continuous, any assigned number is an approximation. When someone weighs 195 lb, it does not mean that the person is exactly that weight. A person who weighs 195.1 lb and another who weighs 194.8 lb might both be

1 The use of a "0" in front of a fractional value that is less than one will be the standard practice for this text. This increases reading clarity. However, there will be occasions when this is not the case, in particular in situations where the professional reporting of values is being used (as well as in the tables of Appendix A).

listed as weighing 195 lb. The number 195 lb is located at the **midpoint** of an interval of weights, that is, the balance point of an interval of weights. The upper and lower boundaries of the interval are called the **real limits**. The upper real limit of the number is one-half the unit of measurement above the number, and the lower real limit is one-half the unit of measurement below the number. If the unit of measurement is 1, the real limits for the number 13 are 12.5 and 13.5. If the unit of measurement is 0.1, then the real limits for 13 are 12.95 and 13.05. Figure 2.1 graphically illustrates the concept of upper and lower limits for numbers with different units of measurement.

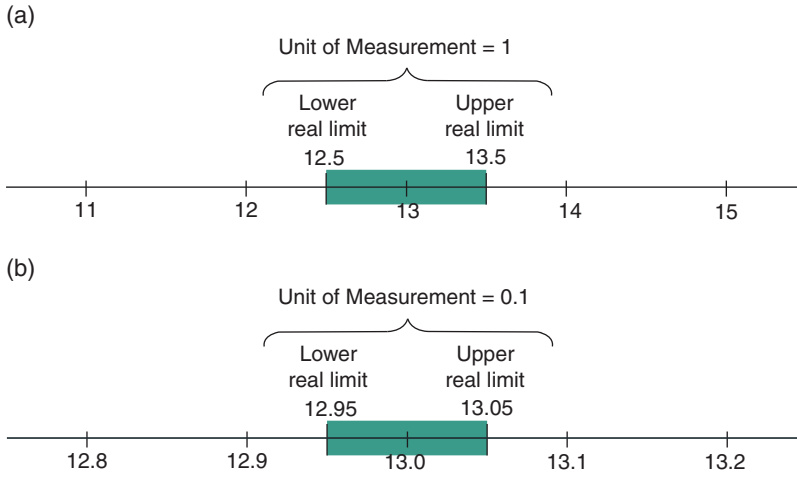


Figure 2.1 The upper and lower limits of a score of 13. (a) The unit of measurement is 1. The real limits are 13 + 0.5 = 13.5 and 13 − 0.5 = 12.5. (b) The unit of measurement is 0.1. The real limits are 13 + 0.05 = 13.05 and 13 − 0.05 = 12.95.

■ **Question** *The following data set contains the average temperature and amount of rainfall of several cities for the month of March. For each number, specify the upper and lower real limit.*

Place	Temperature (°F)	Rainfall (in.)
Acapulco	88	0.1
Chicago	43	2.6
Honolulu	77	3.1
Orlando	76	3.4

Solution The unit of measurement for temperature appears to be 1 °F. The unit of measurement for rainfall appears to be 0.1 in. Since the boundaries of a number are one-half the unit of measurement, the upper and lower real limits of 1° C are 0.5° C above and below the number used to report temperature. Therefore, the upper and lower real limits for the temperature in Chicago are 43.5 and 42.5.

Since the unit of measurement for rainfall is 0.1 in., the upper and lower real limits are specified as 0.05 in. When establishing the upper and lower real limits, it helps to think of the decimal place that is one notch greater in precision. If the scale uses whole numbers, the limits will be stated using the tenth decimal place. If the scale of measurement uses one decimal place (e.g. rainfall in this example), the upper and lower limits will be reported using the second decimal place. For example, the upper and lower real limits for the March monthly rainfall of Honolulu are 3.15 and 3.05. The number 3.1 is the midpoint between 3.05 and 3.15. Table 2.1 presents the upper and lower real limits for the temperature and rainfall data of this problem. ■

In most research situations, a list (or *distribution*) of numbers, called **raw** (or **original**) **scores**, will be obtained. Unorganized data, though, is hard to interpret. However, if the distribution is presented in a tabular or graphical form, summaries and important features of the data set can be communicated to others. The remainder of the chapter presents numerous ways in which data can be presented in tables and on graphs.

Table 2.1 The midpoint, upper, and lower real limits for average temperatures and amount of rainfall for several cities in the month of March.

Temperature			Rainfall		
Lower limit	**Midpoint**	**Upper limit**	**Lower limit**	**Midpoint**	**Upper limit**
Acapulco					
87.5	88	88.5	0.05	0.1	0.15
Chicago					
42.5	43	43.5	2.55	2.6	2.65
Honolulu					
76.5	77	77.5	3.05	3.1	3.15
Orlando					
75.5	76	76.5	3.35	3.4	3.45

2.3 Using Tables to Organize Data

Simple Frequency Distributions

Table 2.2 presents scores from 90 participants who completed a questionnaire measuring their "need for achievement." A quick glance tells us very little about the scores. Indeed, even a longer look only tells us that the values seem to be between 1 and about 30. It would be nice at least to know how many participants received each score.

A **simple frequency distribution** can accomplish this by systematically listing all of the *possible* scores as well as the frequency with which each score appears. Table 2.3 shows a simple frequency distribution for the scores found in Table 2.2. Note that all possible scores are listed, typically in descending order, under the heading X. (The letter X is typically used to represent the concept "scores.") The number of participants that received each score is placed correspondingly and under the heading f (for "frequency"). Adding up all the scores under f will tell us the total number of participants in the study. If the variable being measured is continuous, recall that the X values are actually midpoints of intervals.

Grouped Frequency Distributions

Some sets of data cover a wide range of possible scores, making the resulting frequency distributions long and cumbersome. In these situations, researchers are often willing to exchange the loss of some information to create a table that is easy to understand. A **grouped frequency distribution** indicates the number of scores that fall into each of several ranges of scores (see Table 2.4).

Table 2.2 Unorganized raw data.

15	8	20	16	12	18	14	22	17	5
19	15	18	29	6	13	16	19	10	24
15	3	26	30	13	17	7	16	23	25
1	15	18	14	5	27	16	20	14	6
24	14	20	25	21	15	17	8	23	21
17	14	10	13	18	16	21	9	11	22
15	12	9	16	20	11	13	22	17	13
9	22	16	12	19	17	14	10	19	18
11	16	12	18	13	17	15	14	15	28

Each number is a score from a need for achievement questionnaire.

Box 2.1 Some Notes on the History of Statistics

Although ancient civilizations like the Egyptians and Chinese used tabulation and other simple statistics to keep track of tax collections, government expenditures, and the availability of soldiers, the modern use of statistics arguably began with the Englishman John Graunt (1620–1674). Graunt tabulated information on death rates in his hometown of London and noted that the frequency of certain diseases, suicides, and accidents occurred with remarkable regularity from year to year. This realization, by the way, helped to develop the establishment of insurance companies. Graunt also found the occurrence of greater biological male than biological female births. However, due to the greater male mortality rate (occupational accidents and wars), the number of men and women at the marriageable age was about equal. Graunt believed that this arrangement was nature's way of assuring monogamy (Campbell, 2001).

Most early uses of statistics revolved around simple descriptions of data, but starting around the seventeenth century advances in statistics began to take place, mostly springing from mathematicians' interest in the "laws of chance" as they apply to gambling. The French mathematician Blaise Pascal (1623–1662) was asked the following question by Chevalier de Méré, a professional gambler: "In what proportion should two players of equal skill divide the stakes remaining on the gambling table if they are forced to stop playing the game?" Pascal and Pierre Fermat (1602–1665), another French mathematician, arrived at the same answer, although they offered different proofs. It was their correspondences in the year 1654 that established modern probability theory (Hald, 2003).

The work of Pascal and Fermat was actually anticipated a century earlier by the Italian mathematician and gambler Girolamo Cardano (1501–1576). His volume, *The Book on Games of Chance*, published posthumously in 1663, contains many tips on how to cheat when gambling and established some of the origins of probability theory. Cardano also practiced astrology. Indeed, by using astrological charts he even predicted the year of his death. Upon arriving at that year and finding himself in perfect health, he decided to drink poison to ensure the accuracy of his prediction (Gliozzi, 2008)!

Yet more advances in the field of statistics occurred in the nineteenth and early twentieth centuries. Many of the chapters of this, and every other statistics textbook, are based on the statistical advances of the period between 1850 and 1930. Sir Francis Galton (1822–1911), among other accomplishments, formalized a method for making predictions of one variable with knowledge of a second, related variable (regression analysis) (see also Spotlight 16.1). William Gosset (1876–1937) ushered in the era of modern experimental statistics by

developing analyses that could allow a researcher to make generalizations based on only a small number of observations (the *t* test) (see also Spotlight 9.1). Sir Ronald Fisher (1890–1962) made extensive contributions to the field of research design and developed statistical analyses that can be used to compare the relative influence of several different treatment variables on a dependent variable (the *F* test) (see also Spotlight 12.1). Contemporary statisticians are continuing to make advances in statistics, each advance allowing researchers to ask increasingly complex questions about the mysteries of human behavior.

Table 2.3 The simple frequency distribution constructed from the unorganized data of Table 2.2.

X	*f*	*X*	*f*
30	1	14	7
29	1	13	6
28	1	12	4
27	1	11	3
26	1	10	3
25	2	9	3
24	2	8	2
23	2	7	1
22	4	6	2
21	3	5	2
20	4	4	0
19	4	3	1
18	6	2	0
17	7	1	1
16	8	0	0
15	8		

Table 2.4 A grouped frequency distribution based on the raw data from Table 2.2.

Lower limit	Class interval	Upper limit	Midpoint	*f*
29.5	30–32	32.2	31	1
26.5	27–29	29.5	28	3
23.5	24–26	26.5	25	5
20.5	21–23	23.5	22	9
17.5	18–20	20.5	19	14
14.5	15–17	17.5	16	23
11.5	12–14	14.5	13	17
8.5	9–11	11.5	10	9
5.5	6–8	8.5	7	5
2.5	3–5	5.5	4	3
–0.5	0–2	2.5	1	1

Instead of displaying a frequency count for each score, the viewer learns how many participants obtained scores within a given range. **Class intervals** are groups of equal-sized ranges, determined by the researcher and based on how much information loss one is willing to sacrifice in exchange for

simplicity. The class intervals, typically organized in descending order, cover the full range of scores with no gaps and no overlaps. Each particular score belongs to exactly one interval. The table on display in this chapter features class intervals of 3 units.

Class intervals have midpoints, and when depicting continuous variables, they also have upper and lower real limits. An interval of, say, 20–25, would have a midpoint of 23, a lower limit of 19.5, and an upper limit of 25.5. In Table 2.4 the midpoints, lower limits, and upper limits for each interval from the "need for achievement" data are represented. Rarely are the midpoints and real limits presented in published research. They are included here for educational purposes.

Conventional Rules for Establishing Class Intervals

A grouped frequency distribution sacrifices some information by collapsing numbers into a set of intervals, but it is assumed that this information loss is inconsequential and perhaps even beneficial. Being able to examine the pattern of scores over the range of potential scores is often more useful than knowing the frequency of occurrence for each individual score. Table 2.4 uses 11 intervals. As we view the frequency column of the table, we can now easily see that just a few people received scores in the extreme ends of the distribution. Most of the scores are in the middle of the distribution, with the greatest number of scores in the interval 15–17. (This realization is not as easily seen in a simply frequency distribution.) If too few or too many intervals are used, it can be difficult to see how the numbers are concentrated. The use of about 10 class intervals is customary; however, the needs of the researcher vary from situation to situation. The proper number of intervals to use should be determined by what best illustrates a meaningful pattern or distribution of the scores.

Common interval sizes, symbolized by i, are $i = 3$, $i = 5$, $i = 10$, or $i =$ some multiple of 10. There are no fixed rules for constructing a grouped frequency distribution. However, the following additional guidelines will be helpful:

1) Select an interval size that is suitable. As stated earlier, an interval size that leads to about 10 class intervals is usually ideal for interpretation.
2) Some graphs of continuous measures require the use of the interval midpoint. A midpoint that is a whole number makes a graph easier to read. Try to combine the interval width and the number of intervals in such a way that the midpoint is a whole number. Using an i that is an odd number will accomplish this.
3) The first number of the interval should be a multiple of i. If the interval width is 10, then the first number of the interval should be a multiple of 10. If the

interval width is 2, then the first number of the interval should be a multiple of 2. This guideline is sometimes violated when the interval width is 5. For instance, instead of using an interval of 25–29, with a midpoint of 27, one may decide to use an interval of 23–27 so that the *midpoint* is a multiple of 5 – in this case, 25.

Cumulative Frequency Distributions

A **cumulative frequency distribution** has an additional column that keeps a running tally of all scores up through each given interval. Table 2.5 presents the grouped frequency distribution data found in Table 2.4. The third column of Table 2.5 lists the cumulative frequencies, abbreviated *Cum f*. The arrows in the table show the additive procedure used to find the cumulative frequency at each interval. It is customary to start accumulating the scores from the bottom of the frequency distribution. For instance, for interval 15–17, the cumulative frequency is 58. That is the sum of frequencies found at that interval plus all preceding intervals (1 + 3 + 5 + 9 + 17 + 23 = 58). Note that the total number of scores in the distribution is the top number of the *Cum f* column.

Table 2.5 A cumulative frequency distribution based on the grouped frequency distribution in Table 2.4.

Class interval	*f*	*Cum f*
30–32	1	90
27–29	3	89
24–26	5	86
21–23	9	81
18–20	14	72
15–17	23	58
12–14	17	35
9–11	9	18
6–8	5	9
3–5	3	4
0–2	1	1

2.4 Using Graphs to Display Data

One of the best ways to display data trends is to summarize them in the form of a graphic. In this section, we will learn about some of the graphic displays commonly used in the behavioral and social sciences and their construction. We will also learn how to view graphs with a healthy degree of skepticism. Statistical information in graphical form can be presented in a way that may be technically correct and yet entirely misleading.

The Axes of a Graph

The typical graph has two axes. The horizontal axis is called the *X* axis, or **abscissa**. The vertical axis is called the *Y* axis, or **ordinate**. Larger numbers are to the right on the abscissa and upward on the ordinate. Smaller numbers progress to the left on the *X* axis and downward on the *Y* axis. Figure 2.2a shows the *X* and *Y* axes of a graph with no negative numbers. Figure 2.2b presents a larger perspective, including negative values. Typically the *X* axis reflects the possible values (*X*'s), either quantitative or qualitative, and the *Y* axis reflects the frequency.

The Frequency Polygon

A **frequency polygon** plots the number of scores in each of the intervals of a frequency distribution. The interval width may be 1, as in a simple frequency distribution, or greater than 1, as in a grouped frequency distribution.

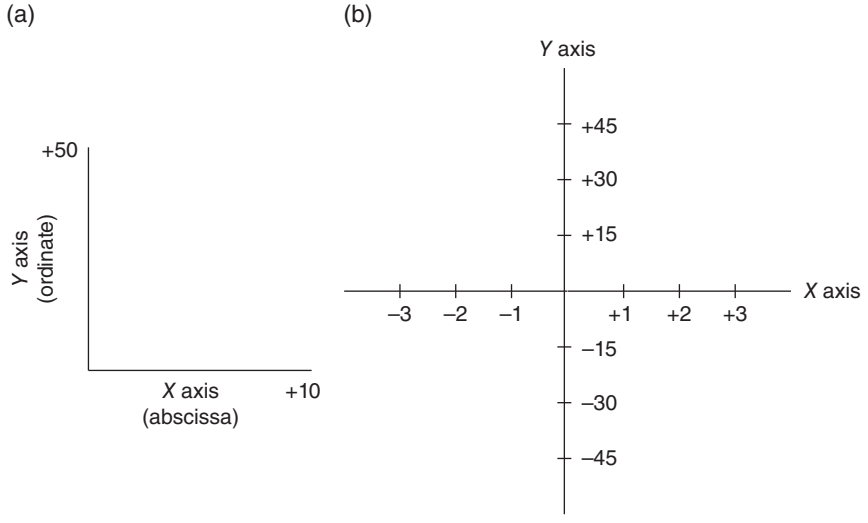


Figure 2.2 (a) The axes of a group without negative numbers. (b) A graph where the axes intersect at their midpoints, which allows for the inclusion of negative numbers on each axis.

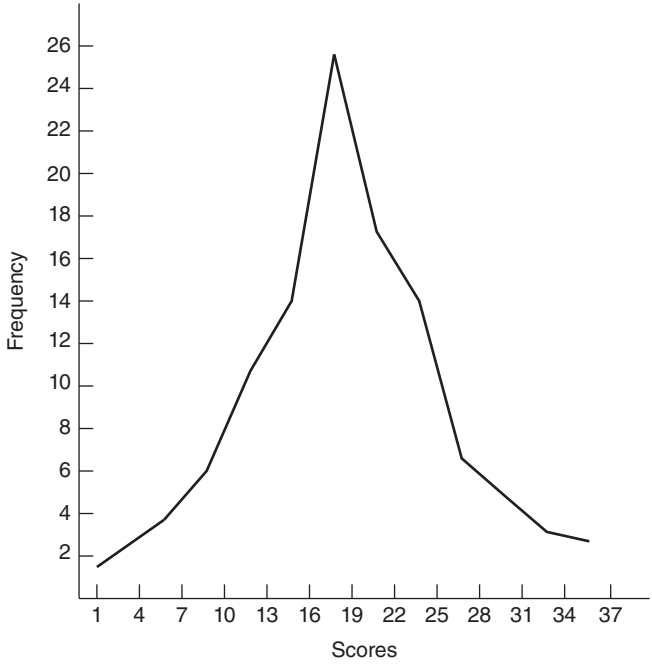


Figure 2.3 A frequency polygon of the data in Table 2.4. Points are plotted above each interval's midpoint.

Figure 2.3 is a frequency polygon drawn from the grouped frequency distribution in Table 2.4. We will note that in Figure 2.3 (as well as Figure 2.4) the first and last points of the graph do not meet the horizontal axis. Whether or not to draw the graph so that the end points meet the *X* axis is a matter of personal preference.

As we view Figure 2.3, note that the *X* axis marks the midpoints of the class intervals. The *Y* axis is labeled "Frequency" and presents equally spaced numbers that specify the frequency of scores. A *single* point on the graph indicates the midpoint of a class interval represented on the *X* axis, and the number of scores found in the interval is indicated on the *Y* axis.

The intersection of the *X* and *Y* axes usually represents the 0 point for each of the variables. However, sometimes the first number of a class interval is some distance from 0, or the first frequency count of an interval is much greater than 0. Should this situation arise, the *X* and/or *Y* axes can be truncated (i.e. shortened where not needed) with broken lines. Figure 2.4 shows a frequency polygon in which the first midpoint of the lowest class interval is 20 and the first frequency count is 100. Note the truncation marks at the base of axes. The truncation technique will be further discussed later in the chapter.

The frequency polygon is a useful graphic for depicting the overall concentration of numbers. It is easy to construct and it is possible to compare two or more

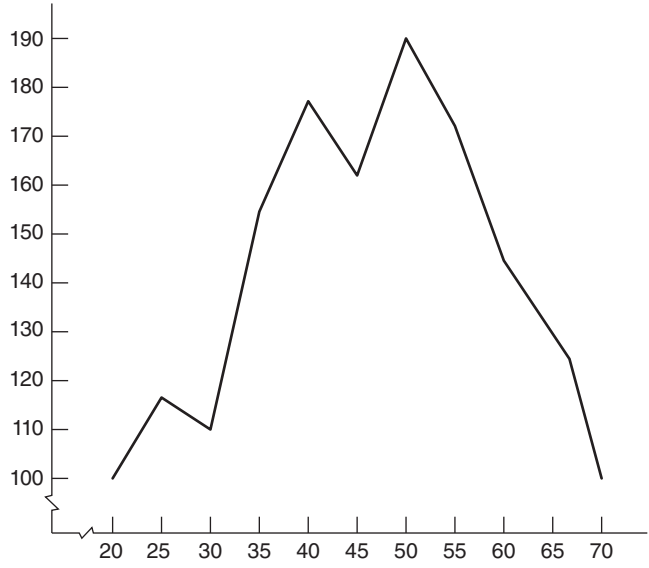


Figure 2.4 The *X* and *Y* axes are broken between 0 and the lowest scores of each axis.

distributions on the same graph (see Box 2.2). However, many suggest frequencies are easier to read when using a different type of graphic display – a histogram.

The Histogram

The **histogram** is a graph of vertical bars with shared borders in which the height of each bar corresponds to the frequency of scores for a given class interval (see Figure 2.5). The width of the bar spans the width of the class interval, including the real limits. This is why there are no spaces between the bars. The bars of a histogram are typically colored in to contrast with the background.

The frequency polygon and the histogram are related. If we were to place a point at the midpoint of the top of each bar of the histogram, erase the bars, and connect the data points, we would have a frequency polygon. A frequency polygon has been superimposed on the histogram depicted in Figure 2.5 so that we can directly compare these two ways to display data graphically.

The Bar Graph

A **bar graph** is used to represent the frequency of scores associated with categories. A bar graph looks like a histogram except the bars do not share a common border. Since the categories represented on the *X* axis are discrete in nature, they do not have real limits. Gaps between the bars clearly communicate this. For example, in Figure 2.6, the scale used on the *X* axis is nominal.

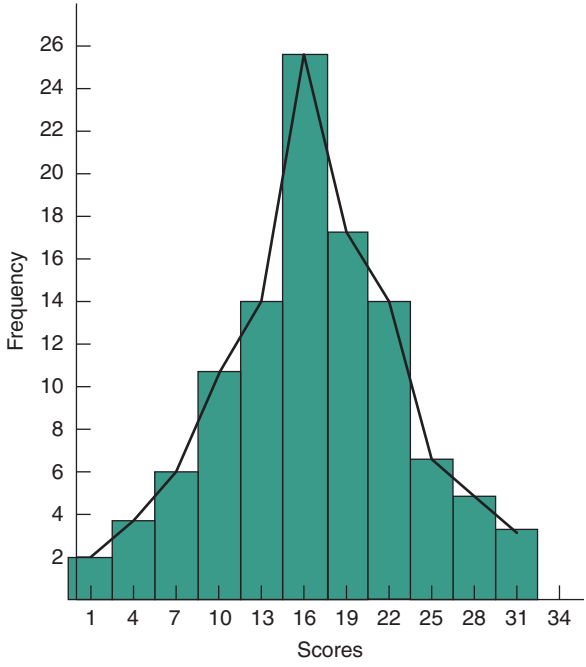


Figure 2.5 A frequency polygon superimposed onto a histogram based on the data in Table 2.4.

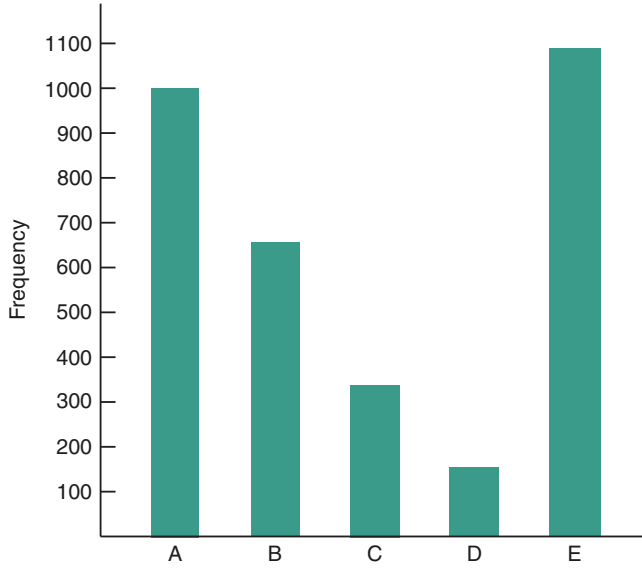


Figure 2.6 The number of undergraduates majoring in psychology (A), sociology (B), history (C), biology (D), and business (E).

Psychology, history, and so on are names of different majors. Figure 2.6 shows hypothetical data depicting the number of undergraduate majors in each of several university programs.

Box 2.2 Using a Graph to Provide a Visual Display of Data

Over the past several years, social scientists have been asking Americans how much confidence they have in specific public institutions. Some interesting trends have been noted. The recent results of a few of these surveys are summarized in the table below (Confidence in Institutions: Trends in Americans' Attitudes toward Government, Media, and Business, 2016).

The table below is a useful summary of confidence measurements for three public institutions. However, to determine if there is a change in confidence over a recent span of years for any one institution requires careful examination of each row of the table by scanning back and forth between the columns. Nonetheless, one can see that the public has much more confidence in medicine than education and more confidence in both of those institutions compared with Congress. Representing these findings on a graph, however, provides a visual display that allows one to observe more quickly these differences across time.

Percent of the Public Expressing a Great Deal of Confidence in Three Public Institutions: Medicine, Education, and Congress

	2006	2007	2008	2009	2010	2011	2012	2013	2014
Medicine	40	40	39	40	41	40	39	39	38
Education	28	28	29	29	29	28	27	26	26
Congress	13	11	10	10	10	10	8	7	6

The abscissa of Figure 2.7 presents the years. This can be understood as a discrete variable representing consecutive categories of time. As a result, a bar graph could have been used. However, by using the line graphs for each institution, trends in the data can be more easily observed.

Examine the relative heights of the lines to compare the differences among institutions in terms of the percentage of people expressing confidence. Finally, do not forget to examine the span of percentages along the *Y* axis. In this case, the relative heights of the lines reveal meaningful differences among the three institutions. However, if there were just trivial differences between the three institutions, it would be possible to adjust the scale on the *Y* axis by truncating it to highlight these minor differences. Whether or not this maneuver would lead to a misrepresentation of the findings is debatable.

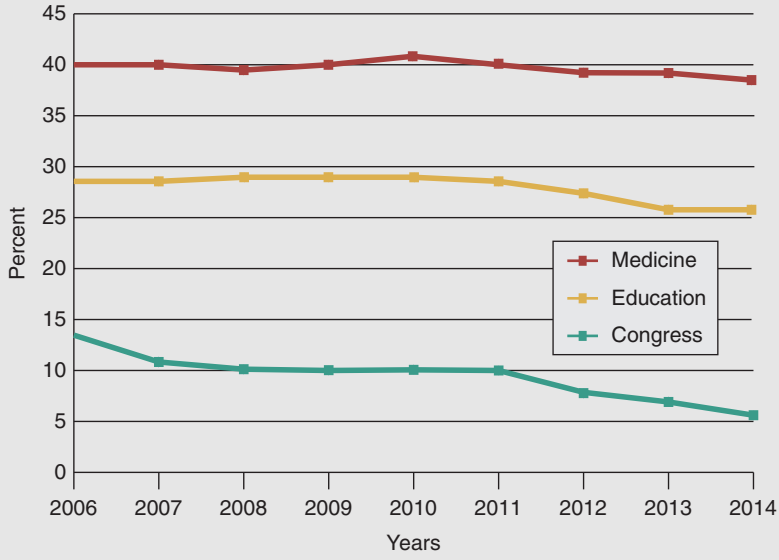


Figure 2.7 Data from "Confidence in Institutions: Trends in Americans' Attitudes toward Government, Media, and Business, 2016" presented in graphical form.

Graphs Can Be Misleading

Suppose a researcher compares two different methods for enhancing learning. As it turns out, Method *B* produces a relative gain of three points, whereas Method *A* does not have any effect on learning. However, let us suppose that the difference between no change and a three-point change is actually very modest. Figure 2.8 labels pretest and posttest scores on the *X* axis. The data points above the pretest indicate the average number of correct responses for each method *before* the participants are administered any training. The data points above the posttest represent the average number of correct responses for each method *after* training. Note that the line for Method *A* is parallel with the *X* axis, indicating no change in performance as a result of training. The line for Method *B* rises *slightly*, reflecting the modest increase in performance. Since the lines show little divergence, it appears that the two methods are very similar in their effects on performance.

Now examine Figure 2.9. The same data are graphed, but now it looks like Method *B* is vastly superior to Method *A*. Why? Notice the *Y* axis. The scale of measurement has been altered so that an increase of three points spans a much greater distance along the *Y* axis. The two graphs are both *technically* accurate. However, the second graph is very misleading. In this example, the *Y* axis has been truncated, but without including the broken axis line. Furthermore, the highest value on the *Y* axis is now 14 instead of 30. This leaves the

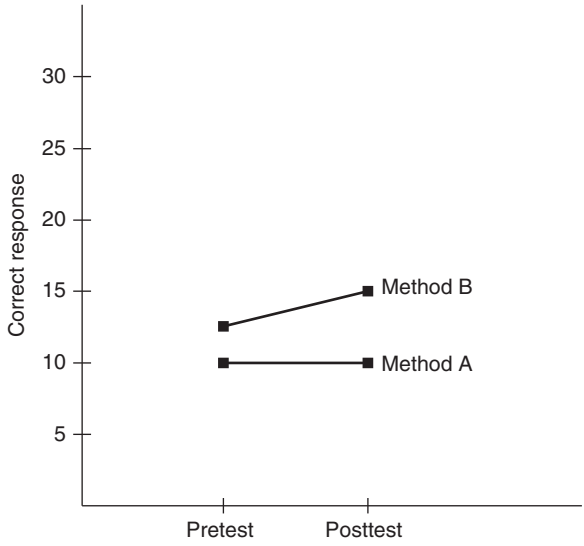


Figure 2.8 A graph that shows the relative effects of two training methods on performance. The lines on the graph indicate that there is little difference between the two methods. Note the scaling of the *Y* axis; it appears to start at zero.

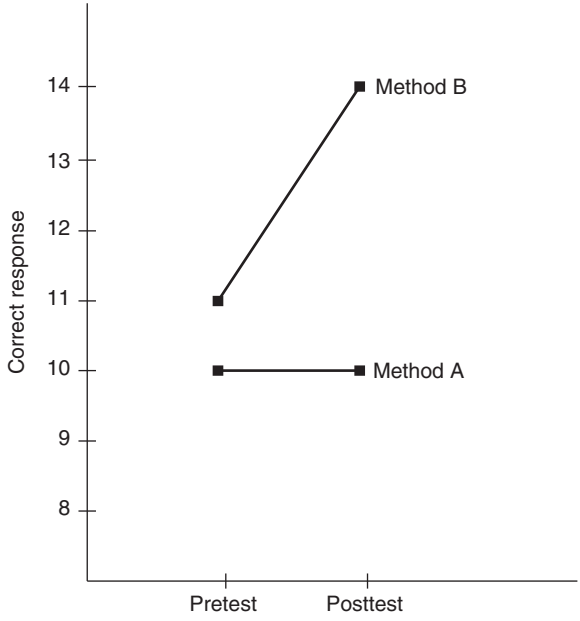


Figure 2.9 The data points of Figure 2.7 are redrawn to create the impression of a vast difference between the two training methods. Altering the numbers on the *Y* axis, especially without signifying that it has been truncated by including a broken line, can give a misleading picture of the results of the study.

impression on the viewer that participants using Method *B* virtually topped out in terms of performance. Together, these techniques serve to amplify the differences between the data lines. However, this amplification seems to misrepresent the actual degree of difference between the two methods. Viewers should always pay close attention to not only the data lines but also the scaling of the axes.

Box 2.3 Is the Scientific Method Broken? The Misrepresentation of Data/Findings

In Box 1.1 we started a series asking whether the scientific method is broken. Public polling suggests most Americans do not possess a "great deal of confidence" in the scientific community (Confidence in Institutions: Trends in Americans' Attitudes toward Government, Media, and Business, 2016). Part of the problem might be the misrepresentation of scientific data and findings.

Data misrepresentation can occur in a number of different ways. One way concerns how science writers interpret scientific findings for the general public. Since most people get their scientific information from the media, those who interpret scientific findings for the general public bear a tremendous responsibility to convey accurately the findings of scientific investigators. However, many writers of science are not sufficiently familiar with the scientific process or the subtleties of doing and interpreting research. Furthermore, there is no getting around the fact that there is a financial incentive behind eye-catching headlines. This situation can often lead to oversimplified descriptions of findings to the general public. A recent example concerns a team of psychologists who, in 2013, reported no cognitive improvement for preschoolers briefly exposed to a music enrichment experience (Mehr, Schachner, Katz, & Spelke, 2013). It was a limited study designed only to see if effects could be found in young children with just an initial transient exposure to music. Great lengths were taken by the authors to clarify the limits of the study. Nonetheless, headlines soon appeared like this one from the Times of London, "Academic benefits of music a myth" (Devlin, 2013), clearly overstating the study's modest conclusions, not to mention bucking most people's strong intuitions to the contrary. Indeed, other research performed just a year later suggests children from disadvantaged backgrounds show improved neuroplasticity and language development with exposure to community music classes (Kraus, Hornickel, Strait, Slater & Thompson, 2014). Some of the public's distrust of science results from the careless way in which many popular interpreters of science report findings – "findings" oftentimes shown to have been stated in far too simplistic terms.

Another form of data misrepresentation concerns the researchers themselves, either through data collection or interpretation. Assuming, for the

moment, the purest of motives, researchers can unintentionally bias participant responses through the ordering of questions (which question comes first, then second, and so on), the limited number of response options available, or even the specific wording of the questions. For example, a 2005 Pew Research survey (Pew Research Center, n.d.) found that the 51% of respondents who favored "making it legal for doctors to give terminally ill patients the means to end their lives" dropped to 44% when asked if they favored "making it legal for doctors to assist terminally ill patients in committing suicide." Phrases that may seem identical to the researcher may be interpreted differently by respondents. In addition, there are hard-to-answer questions regarding how to treat data that does not fit and seems like it may have been gathered incorrectly – so called "outliers." (Should it be discarded? What if it really is good data?) Some researchers also selectively report findings, only publishing relationships that are standout even though numerous relationships were compared. Sometimes a proper understanding of a finding can only be found when placed in a broader context – a context some researchers choose to leave out of their report. For instance, would we be impressed by someone if they said they have such mastery over coin flipping that they can control which side of a coin comes up? What if they said they once got a coin to end up on "heads" nine times in a row? Seems impressive, does it not? However, our amazement might be dulled a bit if we found out their reference to a string of nine heads-in-a-row was dug out of the middle of a series of 4000 coin flips. Context matters. (This topic will be explored more in Box 8.1.) Unfortunately, several scientific articles, many of which misrepresented findings unintentionally, are retracted by academic journals every year. Retractionwatch.com is an example of one website that monitors these retractions.

Finally, there is the issue of academic fraud (e.g. Carey, 2016). Science, we must remember, is not practiced by purely objective robots or angels, but rather by people – people possessing the frailties, temptations, and pressures common to us all. Science is also a cultural enterprise, with its own hierarchy of authority, internal rewarding structure, and value system – a value system that places a premium on new findings, new ideas, and numerous publications. Researchers that do not make original discoveries, propose interesting innovative theories, or generate numerous publications often find themselves out of a job. Given this reality, we should not be surprised to learn that just as the enterprise of professional sports, financial investment, politics, and virtually all other human communities deal with different cheating scandals, this practice can and does take place within the world of scientific investigation. Thankfully, just as in these other professions, there are correcting mechanisms in science – mechanisms designed to ferret out falsehoods and eventually get to the truth. Nonetheless, when the public finds out that a headline may be incorrect, a journal article needs to be retracted, the journal itself is fake, or a scientist is found to be fraudulent, we should not be surprise to learn that to some people it feels as if "science" is broken.

2.5 The Shape of Things to Come

Graphs allow us to look at the "forest" of a distribution set instead of the "trees" of individual data points. As data sets grow in size, the ability to describe efficiently, yet accurately, the features of a distribution becomes more important. It is interesting to learn that most data sets, whatever the size, can be well described by identifying just three characteristics: the shape of the distribution, a measure of centrality, and a measure of variability. The second and third characteristics will be the subject matter for the next two chapters, but for now let us discuss several terms and concepts related to the shape of distributions.

Bell-shaped Distributions

If we were to keep adding numbers to our set of "need for achievement" values (see Table 2.2) and then graph them, using either a histogram or a frequency distribution, we would likely see a bell-shaped curve start to emerge. In fact, with the 90 scores that are given, this shape is already beginning to form (see Figure 2.5). Many naturally occurring phenomena, when measured and graphed, begin to cluster in the middle and approximate what is called a normal distribution. A **normal distribution** (or **normal curve**) is a symmetrical, bell-shaped curved line escalating gradually at first and then more sharply, inflecting at some point and then tapering to a peak (see Figure 2.10 for one example). We will learn much more about normal curves in later chapters. Of course, frequency distributions and histograms, even ones possessing large numbers, do not perfectly take on this flowing curve, but a smooth-lined representation of the data is certainly invited by the viewer. It makes sense, theorists argue, to begin to think of many data sets in terms of being normally distributed. This shape will become exceptionally important to us as we learn more about descriptive and inferential statistical analyses.

Skewed Distributions

Not all data sets, however, approximate a normal curve. One non-normal curve type finds most of the scores near one end of a distribution. This is a called a **skewed distribution**. Figure 2.11 shows a positively skewed distribution and Figure 2.12 depicts a negatively skewed distribution. The type of skewedness can be determined by looking at the direction in which the elongated tail is pointing. The elongated tail points toward the larger positive numbers in a **positively skewed distribution** and toward the smaller or negative numbers in a **negatively skewed distribution**.

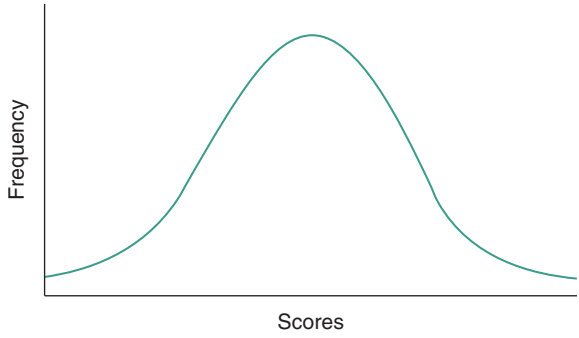


Figure 2.10 One type of bell-shaped curve is the normal curve. Note that it is symmetrical and most of the scores are found in the middle of the distribution.

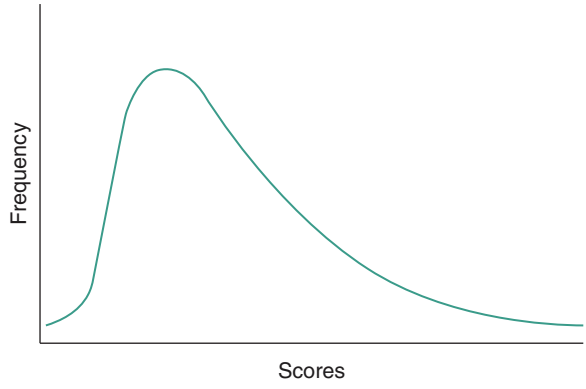


Figure 2.11 A positively skewed distribution. Here, most of the scores are found in the lower end of the distribution, and the elongated tail is pointing toward larger positive numbers.

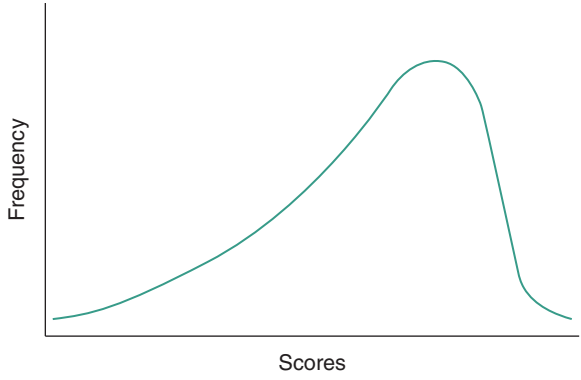


Figure 2.12 A negatively skewed distribution. Here, most of the scores are found in the upper end of the distribution, and the elongated tail is pointing toward smaller or negative numbers.

Kurtosis

Kurtosis refers to the quality of the peak of the curve. The sharpness of the peak reflects the relative concentration of the scores. In Figure 2.13, most of the scores are found very close to the middle of the distribution. When the shape of the curve is relatively narrow and possessing an accentuated peak, the distribution is labeled **leptokurtic**. In Figure 2.14, most of the scores are spread out and widely dispersed. When the shape of the curve is relatively broad and possessing a muted peak, the distribution is labeled **platykurtic**.

Figure 2.13 A leptokurtic distribution. Scores are concentrated heavily around the middle of the distribution with little dispersion among the scores.

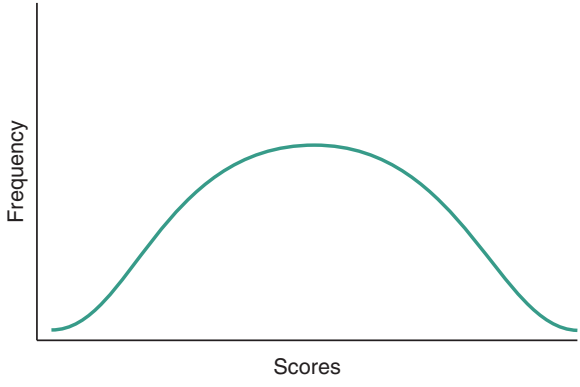


Figure 2.14 A platykurtic distribution. Scores are dispersed widely, and the shape of the curve is relatively broad.

The curves drawn to illustrate the concepts of skewness and kurtosis do not exhaust the myriad of ways in which scores can be patterned. A distribution can be shaped like a *J*, a box, a *U*, or even an *M*; in fact, distributions can assume practically any shape. And these shapes matter. In Chapters 3 and 4 we will learn that different statistical concepts developed to describe data sets should be used for different shaped distributions.

2.6 Introduction to Microsoft® Excel and SPSS®

Starting with Chapter 2, information will be presented at the end of most chapters designed to help students connect concepts explored in the chapter with the capabilities of two statistical application programs, Microsoft Excel and SPSS. Excel was chosen because of its accessibility; it is ideal for students who do not have access to college or university software programs. SPSS was chosen because it is commonly available for students who do have access to college and university computer systems. Appendix C contains general instructions for using Excel and SPSS, including the inputting of data.

Tables and graphs can be made using Excel. Tables are easily constructed by creating column headings and then inputting the corresponding data below the column headings. There are actually three different features within Excel that can be used to create graphs. The Pivot Table feature is recommended when working with discrete or nominal categories (e.g. like ethnicity, religion, academic major, etc.) for the *X* axis. Either the Histogram Analysis Tool or the Frequency function is recommended when working with continuous variables (e.g. time, age, etc.) for the *X* axis. See Figure 2.15 for an example of a table and frequency polygon created using Excel. This data set reflects the frequency and academic classification of students taking a statistics class in a given year. Excel is a very sophisticated program possessing numerous options available to the user for generating customized graphical displays. For specific help using Excel, students are recommended to either purchase an Excel tutorial manual for statistics or use available instructional videos easily found on the Internet.

SPSS is an extremely sophisticated but fairly easy-to-use statistical software analysis tool. For creating tables, use the *Custom Tables* option. This feature is not standard on all forms of SPSS. If not available, users will need to export data into Excel to create tables. For creating graphs, SPSS has the *Chart Builder* feature. See Figure 2.16 for an example of a histogram created in SPSS using the Chart Builder feature. For specific help using SPSS, students are recommended to either purchase a SPSS tutorial manual or use available instructional videos easily found on the Internet.

Students need to be aware that different publication forms have specific formatting standards. Most written work in the behavioral and social sciences requires use of the American Psychological Association (APA) writing style

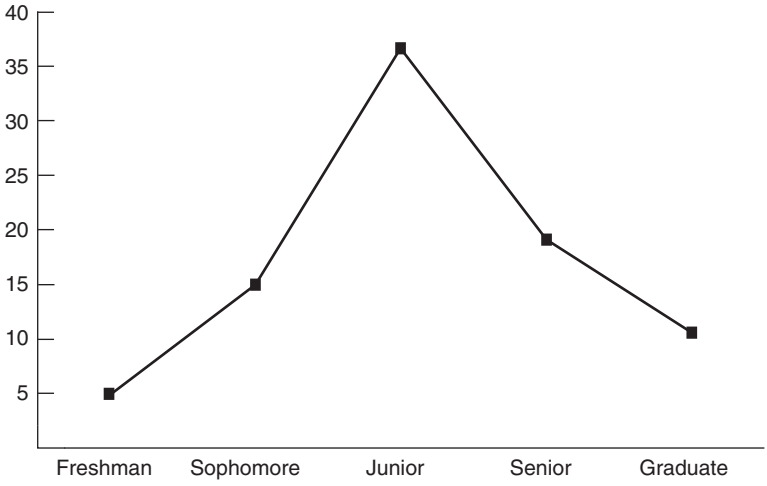


Figure 2.15 A table and graph generated using Microsoft Excel.

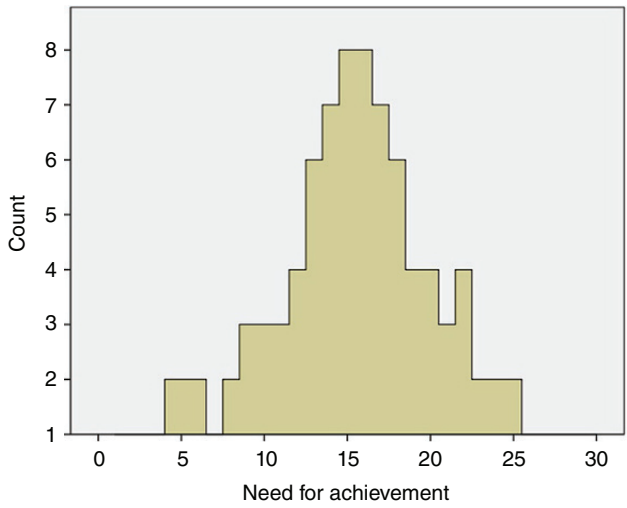


Figure 2.16 A histogram generated using the Chart Builder function in SPSS. The data comes from Table 2.2.

and format. Information about the specific formatting expectations of the APA, including the use of graphs and tables, can be found in their publication manual and on various Internet sites.

Summary

Careful measurement is a necessity for conducting scientific research. This chapter introduces us to the 4 most useful scales of measurement for social and behavioral scientists. A nominal scale merely distinguishes one attribute, event, or thing from another. There is no quantity reflected in nominal numbers, and any number can be assigned to any event as long as it is currently unassigned. Variables measured on a nominal scale are qualitative.

Ordinal measures also categorize things, but in quantitative relation to one another. An ordinal scale is used to identify the relative position of an attribute, event, or object in comparison with others in terms of more or less. The concept of "ranking" is helpful when considering the ordinal scale.

More specific descriptions of quantity are found in the interval scale. Here all intervals between units on the scale are held constant. The amount of quantity needed to go from a five to a six on an interval scale is the same amount needed for a one-unit movement at any other position on the scale. This additional feature is necessary for many statistical techniques since it allows averages to be found.

A ratio scale possesses all of the properties of an interval scale with the addition of an absolute zero point. This allows ratio statements to be made. For example, on a ratio scale, 5 is actually half of 10 and 20 is one-third of 60.

A discontinuous or discrete variable is one that typically increments from one whole number to another whole number. Discrete variables are characterized by gaps between numbers that cannot be filled by any number. A continuous variable, on the other hand, does not have gaps between adjacent numbers. The upper and lower boundaries for a unit interval of a continuous variable are called the real limits. The upper real limit of the number is one-half the unit of measurement above the number, and the lower real limit of a number is one-half the unit of measurement below the number.

Using tables and graphs to organize data allows us to view a summary of the raw scores of the study. A simple frequency distribution lists all the possible scores and the frequency with which each score appears. A grouped frequency distribution indicates the number of scores that fall within each of several intervals. Class intervals are sets of equal-sized ranges used to organize data in grouped frequency distributions. A cumulative frequency distribution includes a column that shows the accumulation of the number of scores for a given interval as well as all of the preceding intervals.

The typical graph has two axes. The horizontal axis is called the X axis or the abscissa. The vertical axis is called the Y axis or ordinate. Larger positive numbers are to the right on the abscissa and upward on the ordinate.

A frequency polygon uses the midpoint to plot the number of scores in each of the intervals of a frequency distribution. A histogram represents the frequency of scores using the real limits of class intervals to create bars with common

borders. A bar graph is used to represent the frequency of scores associated with categories. A bar graph looks like a histogram except that the bars do not share a common border. Viewers of graphs should take note of the scaling on the axes to avoid misunderstanding the presented data.

Statisticians use terms to describe important features of the shape of distributions. A normal curve is a symmetrical, bell-shaped line escalating gradually at first and then more sharply, inflecting at some point and then tapering to a peak. A distribution that has scores that bunch at one end of the distribution is skewed. A positively skewed distribution has scores that bunch at the lower end of the distribution with an elongated tail pointing toward larger positive numbers. A negatively skewed distribution has scores that group around the upper end of the distribution with an elongated tail pointing toward smaller or negative numbers. Kurtosis is a term that refers to the quality of the peak of a distribution. A leptokurtic distribution features a narrow width and accentuated peak, while a platykurtic distribution features a wide width and muted peak.

Microsoft Excel and SPSS are two programs students may have access to when learning about statistics. Tables and graphs can be constructed within both of these programs. While sophisticated, both programs are easy to use with the help of tutorials. Most behavioral and social science manuscripts require APA formatting of figures and tables. The APA Publication Manual is a recommended resource for all behavioral and social science students.

Key Terms

Measurement
Nominal scale
Ordinal scale
Interval scale
Ratio scale
Discontinuous (discrete) variable
Continuous variable
Midpoint
Real limits
Raw (original) scores
Simple frequency distribution
Grouped frequency distribution
Class intervals
Cumulative frequency distribution
Abscissa
Ordinate
Frequency polygon
Histogram
Bar graph
Normal distribution (or normal curve)
Skewed distribution
Positively skewed distribution
Negatively skewed distribution
Kurtosis
Leptokurtic
Platykurtic

Questions and Exercises

1 Indicate whether each of the following scales of measurement are nominal, ordinal, interval, or ratio.
- **a** Amount of change in attitudes.
- **b** Attitudes toward nuclear disarmament (for/against).
- **c** Ratings of popularity.
- **d** Amount of time tolerating a painful stimulus.
- **e** Heart rate under stress.
- **f** A measure of need for approval.
- **g** Amount of weight lifted.
- **h** Numbers assigned based on political affiliation.
- **i** Numbers assigned to different types of diagnostic categories.
- **j** Mood disorder versus no mood disorder.
- **k** A listing of tennis players from best to worst.

2 Think of different ways to measure the following concepts using as many different scales as possible.
- **a** Academic proficiency
- **b** Athletic prowess
- **c** Creativity
- **d** Daily food consumption
- **e** Size of extended family

3 For each class interval, specify the width, the real limits, and the midpoint.
- **a** 1–3
- **b** 5–10
- **c** −4−−8
- **d** −2−+2
- **e** 1.50–3.50
- **f** 25–50

4 The data in the following table are from a midterm examination. Set up frequency distributions with:
- **a** $i = 1$ (simple frequency distribution)
- **b** $i = 3$
- **c** $i = 10$
- **d** $i = 20$
- **e** Include *cum f* columns for each one

Midterm examination scores

40	98	63	90	70	60	45	43	78
67	56	54	78	87	43	90	81	81
77	80	79	80	81	66	75	88	84
49	63	78	79	80	92	89	84	77

5 For each of the frequency distributions in Problem 4, specify the real limits of each class interval.

6 Construct histograms for each of the frequency distributions of Problem 4.

7 Based on the histograms of Problem 6, draw frequency polygons.

8 Think of two variables that may be normally distributed; defend the rationale.

9 Think of two variables that may be negatively skewed; defend the rationale.

10 Think of two variables that may be positively skewed; defend the rationale.

11 The amount of sugar per serving in breakfast cereal might be misrepresented on the side of a cereal box. Draw two different bar graphs, both using the data in the table below: one graph faithfully representing the relationship between the cereals and the other misrepresenting the relationship in such a way as to suggest Cereal A is far superior to these other brands in terms of sugar content.

Cereal type	A	B	C
Sugar in grams/serving	12	14	15

Computer Work

12 Use a software package to establish simple, grouped, and cumulative frequency distributions for the following numbers. Also, generate a graphic of

the following numbers. If the program allows for various graphics, represent the data in each graphic form (e.g. polygon, histogram, etc.).

15	12	13	14	10	15	30	12	17	15	15	30
16	17	28	19	22	25	10	19	32	11	22	32
14	43	32	20	25	29	19	18	29	10	18	39
30	35	19	29	47	25	25	45	16	75	60	25
74	55	18	70	50	20	40	50	45	60	40	62
62	89	61	72	90	65	85	80	60	45	22	49
35	18	49	25	30	59	50	78	35	60	75	39
60	70	25	53	74	74	43	74	72	70	90	75
75	99	77	75	89	60	67	80	80	64	77	82
68	85	80	63	82	75	48	34	16	17	22	25

3

Measures of Central Tendency

3.1 Describing a Distribution of Scores

When social and behavioral scientists collect data, they gather up the numerically represented observations (or "scores") associated with a given variable. As we learned in the last chapter, viewing a list of scores associated with a variable can be rather uninformative. For example, suppose someone has used a tracking device to record the distance walked each day during an eight-week period. If we were to ask them what they found, we might find it rather cumbersome and not particularly helpful if they responded by reciting a 56-item list of daily totals. Although a lot of information would be communicated, it would be hard to make much sense out of it. In Chapter 2 we learned ways in which a distribution of scores could be *visually* displayed by using various tables and graphs. In addition to these visual displays, however, a variety of **descriptive statistics** have been developed, which are designed to quickly and efficiently *numerically* communicate the basic features of a distribution.

What are the most relevant aspects of a distribution to communicate? In Section 2.5 we learned that there are three key features of any data set: the shape of the distribution in terms of the frequency of occurrence for the scores, a measure of central tendency, and a measure of variability. We introduced some concepts related to distribution shape in Chapter 2, and more information about shape will be presented as we go along. This chapter will focus on measures of centrality and Chapter 4 will focus on measures of variability. **Measures of central tendency** (or **centrality**) are statistical indices designed to communicate what is the "center" or "middle" of a distribution, and measures of variability (or dispersion) are statistical indices designed to communicate the degree to which scores are dispersed around this center or middle point.

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

3.2 Parameters and Statistics

In Chapter 1 we introduced the terms "population" and "sample." Recall that a population is defined as every member of a given group and a sample is a subset of a population. Deciding whether a collection of scores is a population or a sample can easily change based on one's perspective. If a researcher gathers test scores from a class and is only concerned with that class, then those scores constitute a population of scores. However, if the researcher is interested in their students in general, then this set of test scores should be considered a sample – a subset of the larger population of "her students." **Parameters** are numerical values that describe the distribution characteristics of a population. **Statistics** are those numbers used to describe the characteristics of a sample. In one case, an average of test scores equaling a 78 represents a parameter; in the other that same number represents a statistic. Unfortunately, the term "statistic" can still be confusing. It can refer to a general field of study (the subject matter of this book), it can be colloquially referenced as a number ("Bob got divorced, now he is just another statistic"), or it can be a value that summarizes a feature of a sample of scores.

Most studies involve the use of samples since the researcher is interested in generalizing the findings to people not in the study. The distinction between a population and a sample is very important in the field of statistics. To help keep the distinction straight, different symbols are used in formulas to indicate when a data set is considered a population or a sample. Moreover, some formulas, such as those used for measures of variability, are actually slightly different, depending on whether the distribution is considered a sample or a population. Misunderstanding the way in which a data set is being conceived can actually cause one to generate a wrong statistical value.

3.3 The Rounding Rule

Before discussing the three main ways to measure central tendency, a brief digression is in order. Since this is the first of many chapters in which we will be performing calculations, we should reach some agreement on how precise we need to be in our calculations. Given that most of the data is centered on whole numbers, it makes sense to take fractional values out to two decimal places. This entails completing our computations to the *third* place to the right of the decimal point and then rounding the value to *two* places to the right of the decimal place. For example, 27.534 would be rounded to 27.53. If the third decimal place is a 5, always round up. So, 34.785 becomes 34.79. Throughout the text, multistep computations are illustrated, and interim values are typically rounded to two places. If our answers to work problems are slightly different than the answers provided in the text or in Appendix B, it is possible that the difference is due to rounding errors. These minor discrepancies should not concern us.

3.4 The Mean

The **mean**, colloquially referred to as the "average," is the most frequently used measure of central tendency. It is also commonly used in formulas designed to test experimental hypotheses. As a descriptive measure, the mean has some advantages and disadvantages, which will be discussed later. Formula 3.1a shows how to find the mean for a population.

Population mean

$$\mu = \frac{\Sigma X}{N} \text{ (Formula 3.1a)}$$

where

μ = (pronounced "mew") the symbol for the mean of a population
X = a score in the distribution
N = the total number of scores in the population (or population size)
Σ = (pronounced "sigma") a notation that directs one to sum up a set of scores. Thus, $\Sigma X = X_1 + X_2 + X_3 \cdots X_n$

The formula for the *sample* mean is identical to the population formula, with the exception of two different symbols. These different symbols clarify if the data set is considered a sample or a population. As different formulas are presented in the text, we will see that Greek letters are used to represent features of a population, while Romanized letters are used to represent features of samples.

Sample mean

$$M = \frac{\Sigma X}{n} \text{ (Formula 3.1b)}$$

where

M = the symbol for the mean of a sample
n = the total number of scores in the sample (or sample size)

■ **Question** *What is the mean of this population of scores?*

5, 8, 10, 11, 12

Solution $\mu = \frac{46}{5} = 9.20$

Notice that the answer of 9.2 would be the same whether the set of scores is considered a population or a sample. Although the designation is theoretically important, it does not impact the calculation of the statistic. ■

The common practice in many statistics books is to use $\bar{X}$ (pronounced "X bar") to represent the sample mean. This is the more traditional symbol. However, most recent published manuscripts in the social and behavioral sciences report sample means using an M. Since students are much more likely to encounter this symbol in their readings and to use this symbol in their professional writing, M will be the symbol used in this textbook. If we see the $\bar{X}$ symbol elsewhere, keep in mind that it also stands for the sample mean.

There are three measures of central tendency discussed in this chapter: the mean, median, and mode. Each measure is designed to communicate where scores tend to center or group in the distribution. However, each measure approaches the concept of "centeredness" differently. In what way does the mean reflect the center of a distribution? Or stated in other words, what does the mean *mean*?

Each raw score in a distribution can be thought of as being "off" of some middle point or deviating from some middle point by a certain amount, even if that amount is zero. The mean is the value where the sum of those raw score deviations across a data set equal zero. To clarify, let us tackle this a different way. A **deviation score** (sometimes referred to as an **error score**) is the distance a raw score is from the mean ($X - M$), and let us symbolize it as x (pronounced "little x"). Therefore, $x = X - M$. So, if the mean of a distribution is 10, a raw score of 12 has a deviation (or error) score of 2.

In Table 3.1, the deviation score for each raw score is listed in the fourth column. Note that a raw score has a negative deviation score when it falls below the mean and a positive deviation score when it falls above the mean. *The sum of all the deviation scores equals 0*; this is how the mean defines the middle or the center of a distribution. Stated mathematically, $\Sigma(X - M) = \Sigma x = 0$. In Table 3.1, both distributions have identical scores except for Participant 5. A score of 30, instead of 10, is obtained by Participant 5 in Distribution *B*. As a consequence, the M of Distribution *B* (10) is greater than the M of Distribution *A* (6). However, the deviation scores still sum to 0. In a manner of speaking, the mean has

Table 3.1 Deviation scores always sum to zero.

Distribution *A*				Distribution *B*			
Participant	**Score**	***M***	***X* – *M* (*x*)**	**Participant**	**Score**	***M***	***X* – *M* (*x*)**
P_1	2	6	–4	P_1	2	10	–8
P_2	4	6	–2	P_2	4	10	–6
P_3	6	6	0	P_3	6	10	–4
P_4	8	6	+2	P_4	8	10	–2
P_5	10	6	+4	P_5	30	10	+20
			$\Sigma x = 0$				$\Sigma x = 0$

adjusted itself so that the Σx is still = 0. It is in precisely this sense that the mean is the center of a distribution. For every distribution, no matter what its shape or number of raw scores, *the sum of the deviation scores off of the mean always equals 0.*

The Weighted Mean

Imagine the mean SAT Writing and Language scores from three high schools in one school district are 425, 470, and 410. If we wanted to find the mean SAT score for the district, would we be justified in taking the mean of the three high school means? No, not unless each school had the same number of students. For instance, imagine the school with the highest SAT average has twice the number of students compared with the other schools. Failing to take that into account would generate a combined mean that would be too low. We need a system of taking each mean into account based on the number of scores that were used to create it. Formula 3.2 accomplishes this task by computing the **weighted mean** (or **grand mean**).

Weighted mean

$$M = \frac{n_1(M_1) + n_2(M_2)\cdots + \cdots n_n(M_n)}{n_1 + n_2\cdots + \cdots n_n} \text{ (Formula 3.2)}$$

where

$n_1.n_2$ = the number of scores in the first group, the second group, and so forth
n_n = the number of scores in the last group
M_1, M_2 = the mean of the first group, the second group, and so forth
M_n = the mean of the last group

■ **Question** *What would be the weighted mean, assuming the following values?*

School 1	*School 2*	*School 3*
$n_1 = 220$	$n_2 = 178$	$n_3 = 192$
$M_1 = 425$	$M_2 = 470$	$M_3 = 410$

Solution

$$M = \frac{220(425) + 178(470) + 192(410)}{220 + 178 + 192}$$

$$= \frac{255\,880}{590}$$

$$M = 433.69 \blacksquare$$

■ **Question** *The mean blood pressure for three age groups has been recorded. What is the overall mean blood pressure?*

	Age		
	20–39	*40–59*	*60+*
Systolic	118	128	145
Diastolic	70	78	82
n	13	12	16

Solution $M_{systolic} = 131$ and $M_{diastolic} = 77$ ■

The Mean of a Frequency Distribution

Chapter 2 shows how a distribution of scores can be displayed in a table, which allows us to ascertain the frequency with which each score occurs. It is an easy matter to calculate the mean of a distribution displayed in such a fashion, whether it is considered a population or a sample. Simply use the following formula.

Mean of a frequency distribution

$$\mu \text{ or } M = \frac{\Sigma Xf}{\Sigma f} \quad \text{(Formula 3.3)}$$

where

f = frequency with which a score appears

Table 3.2 includes a column of raw scores, a frequency column, and a column of cross products, *Xf*. The sums at the bottom of each column are used to find the mean.

The mean is an attractive measure of centrality not only because it incorporates every value in a data set but also because it includes each score's interval distance away from the center. As we will see, the other measures of centrality cannot do this. However, this ability is a double-edged sword. In Table 3.1, we saw that replacing Participant 5's raw score of 10 with the quite discrepant value of 30 shifted the mean tremendously – from 6 all the way to 10. This highlights a problem with using the mean; it is very sensitive to extreme scores. This problem intensifies as data sets get smaller. An extreme score has a greater influence on the resulting mean as the size of the sample or population shrinks.

To illustrate this, consider Congressman Ezra Windblows. The congressman is elected on the promise to bring prosperity to the district. During the next election Windblows would like to convince the constituents that the promise has been kept. The definition of prosperity Windblows uses is the mean income

Table 3.2 Calculating the mean from a frequency distribution.

X	*f*	*Xf*
7	1	7
6	3	18
5	2	10
4	5	20
3	4	12
2	1	2
1	1	1
	$n = \Sigma f = 17$	$\Sigma Xf = 70$
	$M = \frac{\sum Xf = 70}{\sum f = 17} = 4.12$	

of families living in the exceptionally small district. When the congressman was first elected, the mean income in the district was $50 000. Two years later, one couple moved into the district with a yearly income of $250 000. Everyone else's income remained the same. Look what happens to the average family income when the mean is used as the measure of central tendency.

Family income (beginning of Congressman Windblows' term)	Family income (end of Congressman Windblows' term)
$44 000	$44 000
$48 000	$48 000
$50 000	$50 000
$52 000	$52 000
$56 000	$56 000
μ = $50 000	$250 000
	μ = $83 333

Congressman Windblows could honestly report that the average income per family had dramatically increased during this short term in office. Since extreme scores in small samples can result in a mean that does not appear to represent the middle of a distribution, it is necessary to have an index of central tendency that is not particularly sensitive to extreme scores. Now imagine what would happen if that same family moved into a district with about 50 000 families. Would the mean change much?

Unfortunately, many distributions possess more than one extreme score. Skewed distributions, in fact, can feature a moderate percentage of scores trailing well off to one side. If we are interested in accurately communicating where scores of a distribution are bunched, and the existence of extreme scores would lead to a misleading impression, then a different measure of centrality is needed.

3.5 The Median

We have learned that the mean defines centrality as the point in a distribution where the $\Sigma x = 0$. The **median** defines centrality, however, as the number where 50% of the scores in a distribution fall above it, quantitatively speaking, and 50% of scores fall below it. In other words, the median divides the distribution based on the *frequency* or *number* of scores above and below a given point. The median is not algebraically defined, and so for most distributions there is not an algebraic formula to determine it.

Finding the Median When Given an Odd Number of Scores

■ **Question** *What is the median of this distribution?*

1, 4, 6, 8, 40, 42, 43, 45, 47

Solution 40

The median is 40 because the same number of scores (four) fall above 40 as fall below 40. To find the median, we need to position the scores in ascending (or descending) order and then identify the midpoint of the distribution. ■

Note that in the above example the values of the scores surrounding the median are irrelevant. The median is based strictly on the ranking of scores. We could say that the median is "rank sensitive," whereas the mean is "value sensitive." The median takes care of the problem of extreme scores; they only count as one score, no matter their distance from the middle of the distribution. But the downside comes for numbers measured on an interval or ratio scale. Their exact position carries important quantitative information, and yet it is not factored into finding the median; only relative position matters. In this example, a measure of centrality as high as 40 might seem to misrepresent the concept of "center." A few illustrative problems will further emphasize this point.

■ **Question** *What is the mean and median of this sample distribution?*

2, 4, 7, 9, 12, 15, 17

Solution $M = 9.43$ median = 9 ■

■ **Question** *What is the mean and median of this sample distribution?*

2, 4, 7, 9, 12, 15, 17, 46, 54

Solution M = 18.44 median = 12 ■

Look closely at the two previous distributions. They are identical with the exception of two extreme scores added to the second. This has greatly influenced the mean – nearly doubling it. The median, however, was only shifted one score to the right.

Finding the Median When Given an Even Number of Scores

In the examples provided so far, it was easy to identify the median because there were an odd number of scores in the distribution. But what do we do in these situations?

4, 6, 9, 10, 11, 12

1, 2, 4, 6, 8, 11, 14, 18

The median will fall between two scores anytime there is an even number of scores; it will typically be a value that does not occur in the distribution. The median may even be a number, like a fraction, that does not seem to make sense in terms of what is being measured. For instance, imagine the median number of traffic tickets handed out each month for a given city equals 207.5. (What is the meaning of one-half of a traffic ticket?) Remember that statistical concepts convey a feature of an entire distribution; it is not a requirement that the statistical value itself make sense as a score in that distribution. The following examples will show us how to calculate medians when there is an even number of scores in the distribution.

■ **Question** *What is the median of this distribution?*

3, 9, 15, 16, 19, 22

↑

Median

Solution 15.50

The median of a distribution having an even number of values is the mean of the middle two numbers, provided there are not a string of identical numbers in the middle. ■

Finding the Median When There Are Identical Scores in the Middle of the Distribution

■ **Question** *What is the median of the following distribution?*

7, 7, 7, 8, 8, 8, 9, 9, 10, 10

↑

Median

Solution

If discrete, the median = 8.00.

If continuous, the median = 8.17.

It may not be immediately obvious why the type of variable matters and why one answer would be 8.17. Recall from Chapter 2 that discrete variables can take on only a finite number of values. No meaningful values exist between any two adjacent values. In situations like this, since the same number is found on both sides of the middle count, the resulting median is simply that number. However, for continuous variables, every number of a distribution is considered to be at the midpoint of an interval; remember using real limits to draw histograms? Ok, since there are 10 values, we need to have 5 values on each side. Coming up from the bottom, the three 7's get us to within two values of the middle. So, we need 2/3rds of the three 8's to get us to five on each side. Remember that for a continuous measure the value of 8 is the midpoint of the interval 7.5–8.5. So, we need two of those three values that are centered on 8 to go to the lower side of the middle and one of the values centered on 8 to go to the higher side. But those 8's cannot be separated since they are stacked on top of each other. (Look at Figure 3.1.) Ok, so we will have to split those three boxes identically so that a total of two of them fall to the lower half of the distribution and the remaining parts of the boxes fall to the upper half. If we drew a line down through the three boxes such that 2/3rds of each box was to the left and 1/3

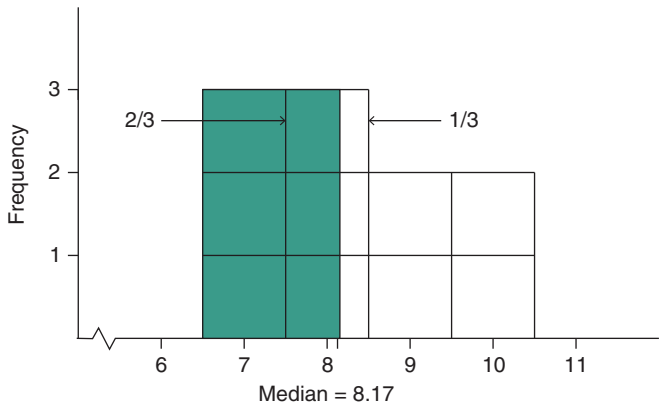


Figure 3.1 A visual representation of how to find a median when there are identical scores in the middle of the distribution.

of each box was to the right, that would "do the trick" (7, 7, 7, 2/3rds of the first 8, 2/3rds of the second 8, and 2/3rds of the third 8 would all be on the left). Well, what is 2/3rds of the way from 7.5 to 8.5? Let us add 0.67 to 7.5. That gives us 8.17. ■

■ **Question** *What is the median of this distribution?*

7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 10, 10

↑

Median

Solution

If discrete, the median = 8.00.

If continuous, the median = 8.10.

Here is another one where, if the measure is continuous, we will need to split up the 8's. Three of them need to go to the lower half of the distribution and two of them to the higher half. The fairest way to get three of the five 8's would be to get 60% or 0.6 of each one. Since the 8's actually start at 7.5, the answer would be (7.5 + 0.60) 8.10. ■

Box 3.1 The Central Tendency of Likert Scales: The Great Debate

In Chapter 2 readers were introduced to a critical difference between ordinal scales and interval or ratio scales – the nature of the relationship between numerical values. Ordinal scales are a quantitatively organized series of categories and, as such, make no assumptions about the quantitative distance between these categories. Interval and ratio scales hold the intervals constant throughout the measure. In this chapter we learned that the concept of a deviation score is necessary to find a mean. A mean is defined as the point where the deviation scores sum to zero. Deviation scores, however, cannot be found for numbers on an ordinal scale; the intervals are not held constant. (It would be like suggesting first- and third-place finishers in a pie bake-off are equidistant from second. We have no reason to presume that.) Readers were also made aware of the ambiguity that surrounds how to interpret numbers generated by a Likert scale – scales that typically offer 5–11 options ranging from strongly disagree to strongly agree (usually with a neutral point in the middle) that are used to measure the amount of agreement people have with a given statement. The great debate is this: Is it appropriate to generate means for Likert-scale data?

It is very important because if we decide the answer is "no," then we eliminate all statistical tests that make use of the concept of the mean. Much of the content in statistics textbooks is not applicable to data measured on an ordinal scale. An answer of "no" resigns us to use what are called nonparametric tests.

(Chapters 17 and 18 in this textbook are devoted to nonparametric tests.) These tests are less powerful (a concept we will explore in Chapter 11) and, therefore, less likely to help us find meaningful differences between groups.

The conservative approach is to argue that Likert scales have no way to determine constancy between values and should therefore be considered ordinal. As a result, data gathered using Likert scales must be analyzed nonparametrically; end of discussion. Others argue that the line between ordinal and interval is rather vague, some even calling it "fuzzy" (e.g. Abelson, 1995). If, they argue, the data from a Likert scale takes on the shape of a normal distribution, and if there are a good number of options for the respondent to choose from, then the data can be considered "normal" or "sufficiently close" to normal and analyzed with more standard statistical techniques.

We do not aim to settle the debate here, but merely raise it as an important issue. Perhaps it will be a good class discussion topic. As we think about this issue, keep in mind some recommendations made by Karen Grace-Martin, a specialist in data analysis. They are paraphrased below (Grace-Martin, 2008):

1) Realize the difference between a Likert-type item and a Likert scale. A Likert scale is actually made up of many items. Collectively, they attempt to provide a measure of the attitude in question. Many people, however, use the term Likert scale to refer to a single item.
2) Proceed with caution. Look at the particulars of our Likert-scale data. Would treating it as interval data influence our conclusions? The fact that everyone else is treating it as interval data is not sufficient justification in and of itself.
3) At the very least, insist (i) that the item have at least nine points, (ii) that the underlying concept be continuous, and (iii) that there be some indication that the intervals between the values are approximately equal. Make sure the other statistical assumptions for the test are met.
4) When we can, run the nonparametric equivalent to our test. If we get the same results, we can be more confident about our conclusions.
5) If we do choose to use Likert data in a parametric procedure, make sure we have particularly strong results before making a claim.
6) Consider the consequences of reporting inaccurate results. Is the analysis going to be published? Will it be used by others to make decisions?

The hope here is less about bringing readers to some desired position and more about helping students develop an appreciation for some of the more subtle and yet important debatable issues related to data analysis. How we understand what the value of our scores mean is critical, requiring us to first figure out the type of scale being used. And if we find that we are using Likert data; well then, welcome to the great debate!

3.6 The Mode

The third and final measure of centrality is the mode. Recall that the mean defines centrality as the point where all of the deviation scores sum to 0 ($\Sigma x = 0$). The median defines centrality as the point where half of the scores of the distribution fall above it and half fall below it. The **mode** defines centrality as the most typical or most frequent score in the distribution. It is the easiest of all three measures to determine. All we need to do is look to see which score occurs most often.

■ **Question** *What is the mode of this distribution?*

100, 101, 105, 105, 107, 108

Solution 105 ■

Some distributions may have two scores that are most typical. Consider the frequency distribution in Table 3.3. Here, the distribution has two modes: 40 and 34. This is known as a **bimodal** distribution ("bi," meaning two). A distribution with a single mode is termed **unimodal** ("uni," meaning one). The graph of a bimodal distribution has two distinct humps. The humps do not have to be exactly the same height to be a bimodal distribution. Consequently, two modes can be reported even if the number of observations associated with each modal score is not identical. In the rare case in which all scores occur with the same frequency, there is no mode.

Table 3.3 A distribution with two modes: 40 and 34.

X	*f*
43	1
42	4
40	6
39	3
37	2
34	6
30	1

3.7 How the Shape of Distributions Affects Measures of Central Tendency

Chapter 2 introduced us to the notion that distributions can assume different shapes. A distribution can take on virtually any shape, and there are names for some of them (platykurtic, leptokurtic, positively skewed, negatively skewed, normal, etc.). The particular shape of the distribution has implications for the relative position of the mean, median, and mode. If the distribution is symmetrical, then all three measures of central tendency will be identical. Figure 3.2 depicts this fact. Note that symmetry is what is important here, not kurtosis (peakedness).

There is one exception to this rule: A symmetrical bimodal distribution, as shown in Figure 3.3, has identical values for the mean and median, but not

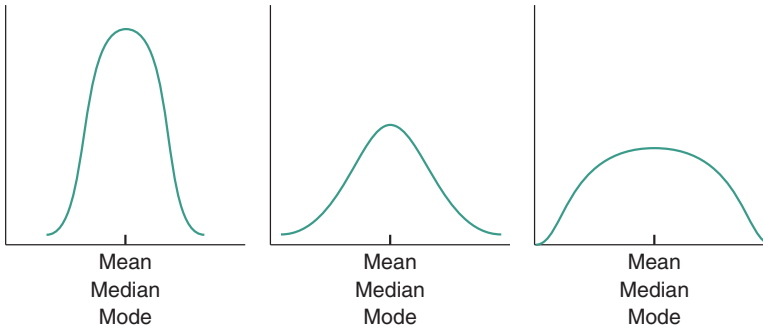


Figure 3.2 Three distributions with varying degrees of kurtosis but with the same mean, median, and mode.

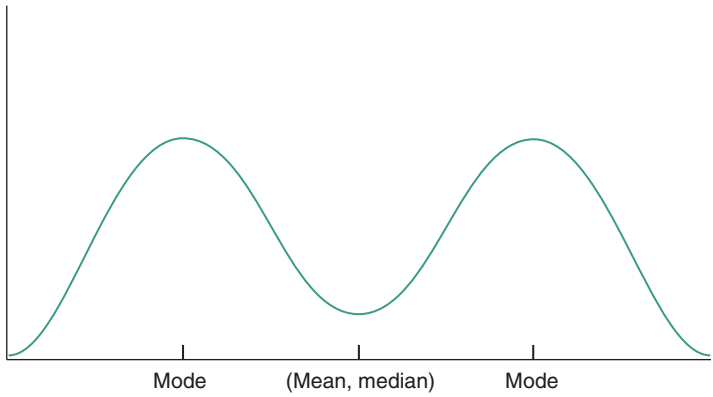


Figure 3.3 A symmetrical, bimodal distribution. The mean and median are the same.

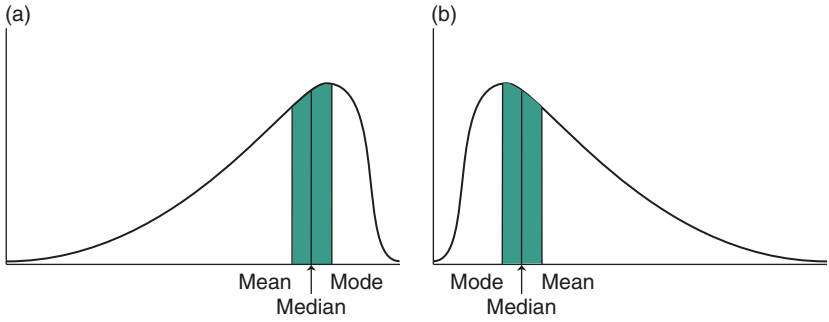


Figure 3.4 (a) The mean is to the left of the median in a negatively skewed distribution. (b) The mean is to the right of the median in a positively skewed distribution.

the mode. Also, note that in a bimodal distribution, the mean and median are not located in the midst of highly frequent scores. In this instance, the mean and median do not reflect where scores tend to bunch.

If a distribution is skewed, then the mean, median, and mode will all be different. Figure 3.4 shows two distributions: (a) negatively skewed and (b) positively skewed. In a negatively skewed distribution, the mean is to the left of the median. The scores in the elongated tail pull both the median and the mean to the left, but they pull the mean more so, because the mean takes into account the actual distance these extreme scores are from the center. In a positively skewed distribution, the mean is to the right of the median. The scores in the elongated tail pull both the median and the mean to the right, but they pull the mean more so, because the mean takes into account the actual distance these extreme scores are from the center.

3.8 When to Use the Mean, Median, and Mode

Using the mean as an index of central tendency has several advantages. First, as mentioned earlier, the mean takes into account all scores in a distribution, including each score's interval distance from the center. For this reason, the mean usually captures a distribution's centeredness well. Second, as the size of a random sample grows, it turns out to be a very stable estimate of the population mean. This will make much more sense once we learn how to create sampling distributions (Chapter 7). Third, because of the preceding two reasons, the mean is used in many statistical formulas.

However, there are two disadvantages to using the mean. First, means do not carry meaningful quantitative information for data gathered from either nominal or ordinal scales. This means that the large numbers of statistical analyses

based on the mean are not available for use with data coming from these scales. Second, as previously mentioned, the mean is sensitive to extreme scores; when present they can produce a mean value that does not seem to be well centered. This disconnect can occur in a couple different ways. Sometimes it only takes one extreme value in a small distribution set to pull the mean far away from what appears to be the middle. But even for larger distribution sets, if they are tremendously skewed (either positively or negatively), the mean may not appear to be near the bulk of the scores in a data set (see Figure 3.3). Medians, though also pulled out of the middle by extreme scores, are not pulled out as far as the mean. For this reason, medians are oftentimes preferred as measures of centrality for skewed distributions.

Another problematic situation for the mean occurs when the scores of a distribution are truncated. For example, consider a study on self-control and pain tolerance by Grimm and Kanfer (1976). In this study the researchers were interested in teaching a self-control technique for tolerating pain induced by immersing a participant's hand in freezing water. As the experiment unfolded, a few participants kept their hands in the water a very long time. As a result, the experimenters decided to establish a cutoff time of 300 seconds. The exact scores of the participants who kept their hands immersed in the water for the full 300 seconds cannot be known – some may have been ready to stop right then, but others might have gone on for much longer. Nonetheless, they all got a score of 300 seconds. Therefore, the mean is likely to be artificially low as a result of the arbitrary cutoff time. The median, however, would be the same value whether the scores have been cut off at 300 or not. Since the median is not influenced by the value of extreme scores, it should be used when the distribution is skewed, truncated, or has inexact upper or lower cutoff scores.

The median is also the central tendency measure of choice for data drawn from an ordinal scale. Neither the ordinal scale nor the concept of the median makes any assumptions about the uniformity of the intervals between values.

The mode is the least used measure of central tendency. The mode of a sample, for instance, is never used to infer the mode of a population. In addition, the mode ignores all the numbers in a distribution except the one score that occurs most often. The mode does have its place, however. The mode is used when one wants to convey the most typical score found in a distribution, such as when students want to know what score on an exam was received by the most people. And the mode is the preferred measure to use when working with nominal data. Recall from Chapter 2 that a nominal scale merely distinguishes one kind of thing from another. Suppose we ask students at our university to name their favorite leisure activity. We report that 12% prefer gaming, 50% report tending to social media accounts, and 38% point to sporting activities. These three activities are categories on a nominal scale. There is no way to compute a mean or median, but the mode can be declared to be tending

to social media accounts. Oftentimes when the data are in the form of *how many* (i.e. a nominal scale), the mode is the appropriate measure of central tendency.

It should now be clear that each measure of central tendency has advantages and disadvantages. Bear in mind that when using descriptive statistics, the goal is to communicate the features of a distribution in as accurate a manner as possible. Further, there is no rule that says we cannot report two or even all three measures of central tendency.

3.9 Experimental Research and the Mean: A Glimpse of Things to Come

In Chapter 1, several fundamental concepts of experimentation were presented. At the most basic level, experimental designs compare the performances of different groups of participants. The typical statistics used to determine whether the independent variable affected the dependent variable are the means of the various groups. Some examples of studies in which group means are used to reach conclusions are presented below.

► **Example 3.1** An educational theorist is interested in comparing the effectiveness of two teaching techniques. Participants assigned to one group are exposed to educational material via an online teaching experience. Participants assigned to another group take part in a traditional classroom experience. The dependent variable is the amount of material learned. The mean of the amount of material learned is computed for both groups. Through the use of statistical analyses (presented in later chapters), the means of the groups are compared to decide if one teaching method is superior to the other.◄

► **Example 3.2** A social psychologist is interested in learning about the relationship between different mood states and charitable giving. In one condition, it is arranged for participants to experience a pleasant interaction with the experimenter. In another condition, participants are treated in a cold, rude manner by the experimenter. Soon after leaving the laboratory, a person approaches the participant and asks for a donation to a homeless shelter. The dependent variable is the amount of money donated. The means are computed for both groups and compared to see if mood states influence generosity.◄

► **Example 3.3** A child psychologist would like to evaluate two treatment techniques for helping children overcome their fear of the dark. Participants assigned to one group are taught to imagine themselves as a superhero on a mission during the night. Participants assigned to a second group are told to repeat over and over, "I'm a big boy/girl." The dependent variable is the amount of time the child is willing to stay in a dark room. Means are calculated for both groups and compared to judge whether one method is more effective than another in helping children tolerate the dark.◄

This chapter has presented several factors that should guide us in deciding which measure of central tendency to use when describing a distribution of scores. As we make our way through the text, we will discover that, when conducting an experiment, the mean is almost always the statistic that serves as the point of comparison between different conditions.

Box 3.2 presents a study that taught participants how to control their heart rate. Means were computed for groups of participants at two points in the study. Statistical techniques discussed in later chapters will show us how to use the means to compare the two groups of participants and interpret the results.

Box 3.2 Learning to Control Our Heart Rate

For several decades, biofeedback was a popular treatment for many stress-related physical ailments. In the 1980s researchers started to investigate its effects. Biofeedback entails the provision of external feedback in the form of a visual display or varying auditory stimulus, which changes as some physiological response changes. Thousands of people have learned how to relax with biofeedback training; there is little doubt that most people can achieve an impressive degree of control over their physiological responses, at least while they are attached to the biofeedback equipment. But therein lies the problem. What good is it to learn how to relax if we can only experience that state when we're hooked up to a machine? Posed as a research question, we might ask, "When participants learn how to control one of their physiological responses, will they be able to transfer learning to control that response during their everyday activities?" It was this question that led Gloria Balague-Dahlberg (1986) to conduct the following study.

Study Method

Eighteen participants who scored high on an anxiety questionnaire participated in the experiment. To assess the participants' heart rate throughout the day, they were asked to wear a Holter monitor (a device that continuously records heart rate). The participants were asked to try to keep their heart rate low while going about their usual daily routine.

Half of the participants were seen individually for five biofeedback sessions, during which time they tried to lower their heart rate as much as possible. Although biofeedback is always conducted in a relaxed, comfortable atmosphere, Balague-Dahlberg reasoned that the transfer of learning to the natural environment would be augmented if participants initially learned to control their heart rate in a setting filled with distractions. So with each successive session, participants attempted to lower their heart rate amid an increasing level of distractions. This was procedurally accomplished by having participants sit in a hard chair while performing a series of mental tasks. As the sessions progressed, a tape of distracting noises was played: people talking, phones ringing, machines running, and other "office noises."

In addition to this experimental group, a control group was included: a group that received the same instructions but did not have experience with the biofeedback equipment. After the training phase of the experiment, all participants' heart rates were once again monitored for a 24-hour period.

Results

The data from this study are presented in the following tables. The baseline score (also called a pretest score) is the mean heart rate for the 24-hour period before training; the posttest score is the mean heart rate during the final 24-hour recording period. A graph (Figure 3.5) is presented so that we can easily see the difference between the groups at each phase of the study. (Yes, truncation was used to highlight this difference. More will be said about this at the end of this box.)

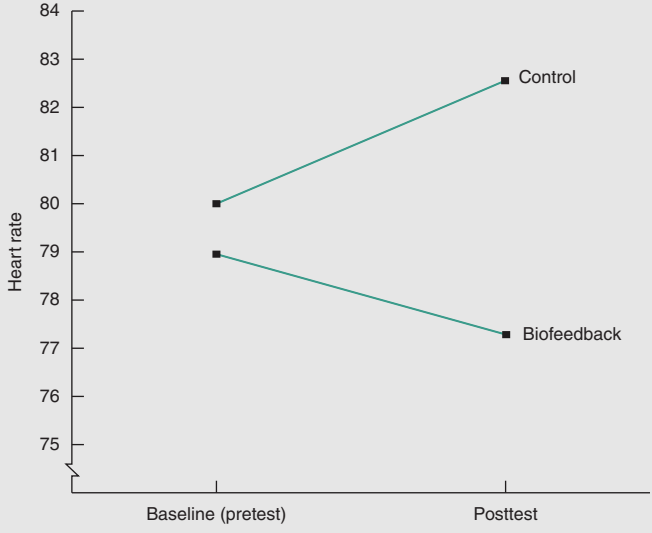


Figure 3.5 Graphical presentation of the results of the Balague-Dahlberg study on heart rate control.

Biofeedback group		
Participant	*Baseline*	*Posttest*
1	92	92
2	64	70
3	93	86
4	70	71
5	67	69
6	93	74
7	63	62
8	86	93
9	84	79
	M_{pre} = 79.11	M_{post} = 77.33
Control group		
Participant	*Baseline*	*Posttest*
10	90	95
11	92	99
12	79	82
13	85	86
14	75	73
15	82	84
16	78	73
17	80	83
18	61	63
	M_{pre} = 80.22	M_{post} = 82.00

The biofeedback and control groups have similar mean heart rates at the pretest baseline measure. This was to be expected because participants were randomly assigned to conditions and had not yet received the different treatments. It is evident from the 24-hour posttesting data that the biofeedback participants appear to have learned from the training and were able to keep their heart rate at a level lower than the control participants. However, one would not want to conclude anything by merely visually comparing the means. At this point, the experimenter would conduct the appropriate statistical test to determine if these differences are unlikely to occur by chance. (We will learn about these tests later in the text.) If these posttest differences are unlikely to occur by chance, then tentative conclusions can be made about the superior effects of biofeedback training. (Furthermore, we can say that the graph does not mislead the viewer. If, however, the analysis suggests that chance factors can explain the posttest difference, then some could argue that the graphic appears to mislead unsuspecting viewers about the effects of biofeedback training.)

Summary

Descriptive statistics are statistical indices that summarize and communicate basic characteristics of a distribution. Values that communicate where scores center in the distribution are called measures of central tendency. Measures that communicate the degree to which scores are spread out around the center of a distribution are called measures of dispersion or variability. Statistical values that describe the distribution characteristics of a population are called parameters; statistical values that describe the distribution characteristics of a sample are called statistics.

The mean is the most important and most often used measure of central tendency. Not only can it be used as a descriptive index of central tendency, but the mean is frequently used in formulas designed to test experimental hypotheses. The degree to which a score deviates from the mean is $X - M$. This deviation amount can be called a deviation score (or error score) and is symbolized as x. Therefore, $x = X - M$. The sum of all the deviation scores equals 0. Therefore,

$$\Sigma(X - M) = \Sigma x = 0.$$

The mean has several advantages. First, it takes into account not only all of the scores in a distribution but also their precise distance from the middle. Second, it is used in many statistical formulas. Third, as the size of the distribution increases, the mean becomes a very stable measure of central tendency. A sample mean is usually a good estimation of the mean of a population. However, since the mean is sensitive to extreme scores, it is oftentimes not seen as a good measure of centeredness when the distribution is skewed. Using the mean as a measure of central tendency can also present a problem when the distribution is truncated, that is, when one or both ends of the distribution have been limited by the nature of the measuring instrument.

The median is the point in the distribution where 50% of the scores fall above and 50% fall below it. Since the median is not affected by the value of extreme scores, it should be used when the distribution is skewed, truncated, or has scale-limited upper or lower cutoff scores. The median is also the appropriate measure for central tendency when the values in the distribution come from an ordinal scale.

The mode is defined as the most typical or most frequent score. It is the least used measure of central tendency. The mode ignores all of the numbers in a distribution except the one value that occurs most often. On the other hand, the mode is the only measure of central tendency to use when evaluating scores measured on a nominal scale.

The particular shape of the distribution has implications for the relative position of the mean, median, and mode. If the distribution is symmetrical, then all three measures of central tendency will be identical. In a positively skewed

distribution, both the mean and median are pulled to the right, although the mean is pulled farther. In a negatively skewed distribution, both the mean and median are pulled to the left, although the mean is pulled farther.

Using Microsoft® Excel and SPSS® to Find Measures of Centrality

Excel

General instructions for data entry into Excel can be found in Appendix C.

Data Analysis

1) Once the data is entered, select **Data Analysis** and then **Descriptive Statistics**. Click **OK**.
2) Highlight only the scores and put those quadrant numbers into the **Input Range**.
3) Select a location for the output. Use the **Output Range** box if needed.
4) Make sure to click **Summary Statistics** before clicking **OK**.

This should generate a table with all three measures of centrality.

SPSS

General instructions for inputting data into SPSS can be found in Appendix C.

Data Analysis

1) Once the data has been entered, click **Analyze** on the tool bar and select **Descriptive Statistics** and then **Frequencies**.
2) Move the column label containing the data we wish to analyze from the left box to the **Variable** box. Also, unclick the **Display Frequency Tables** box in the lower left corner.
3) Before we run it, click on the **Statistic**s box in the upper right corner. Here we will find boxes labeled Mean, Median, and Mode. Click them and then click **Continue**.
4) Now we are ready to click **OK**. The first box in the resulting output will give us these three measures of central tendency.

Key Formulas

Population mean, μ

$$\mu = \frac{\Sigma X}{N} \text{ (Formula 3.1a)}$$

Sample mean, *M*

$$M = \frac{\Sigma X}{n} \text{ (Formula 3.1b)}$$

Weighted mean

$$M = \frac{n_1(M_1) + n_2(M_2)\cdots + \cdots n_n(M_n)}{n_1 + n_2\cdots + \cdots n_n} \text{ (Formula 3.2)}$$

Mean of a frequency distribution

$$\mu \text{ or } M = \frac{\Sigma Xf}{\Sigma f} \text{ (Formula 3.3)}$$

Key Terms

Descriptive statistics
Measures of central tendency (or centrality)
Parameters
Statistics
Mean
Deviation score (error score)
Weighted mean (or grand mean)
Median
Mode
Bimodal
Unimodal

Questions and Exercises

1 Given a distribution in which M = median, what must be true about the shape of the distribution?

2 If a distribution of 25 scores that has a mean of 15 is said to be a population, how will the mean change if the distribution is later claimed to be merely a sample?

3 Find the deviation scores for the following raw scores in a distribution that has a mean of 10.
a 12
b 9
c 0
d 10
e −1
f 9.5

4 Find the deviation scores for the following raw scores from the following sample of scores: 6, 6, 7, 7, 7, 8, 9, 10, 11, 12, 16.
a 7
b 11
c 0
d 8.5
e −20
f 11.5

5 Identify the mean, median, and mode of these six distributions.
a 3, 3, 4, 5, 6, 8, 8, 8, 9
b 2, 4, 4, 4, 6, 7, 7
c 7, 7, 8, 9, 10, 10, 10
d 1, 1, 3, 4, 4, 5, 9
e 1, 4, 6, 7, 8, 8
f 9, 11, 6, 8, 12, 15, 3, 5, 5

6 For distributions (a) and (b) of the previous problem, identify:
a $\Sigma(X - M)$
b $\Sigma(X - \text{Median})$
c $\Sigma(X - \text{Mode})$

7 What is the (a) mean and (b) mode of this frequency distribution?

X	*f*
12	3
10	4
9	6
7	5
4	2

8 What is the (a) mean and (b) mode of this frequency distribution?

X	*f*
23	1
19	3
16	4
15	4
12	2

9 What is the median of this distribution?

4, 5, 7, 7, 7, 7, 9, 10

10 What is the grand mean of these four group means?

M	*n*
156	5
199	10
88	11
145	4

11 What is the grand mean of these five group means?

M	*n*
6.5	2
7.5	4
5.0	6
4.0	4
13.0	1

12 What can be said about the shape of each of these distributions?

a Mean = 24; median = 16; mode = 12
b Mean = 123; median = 143; mode = 150
c Mean = 6; median = 6; mode = 6
d Mean = 19; median = 19; mode = 9 and 29
e Mean = 56; median = 66; mode = 70
f Mean = 48; median = 36; mode = 32

13 Think of two ss of variables that may be distributed in such a way that they might take on a bimodal shape. Defend these choices.

14 A national team of researchers is studying depression among women. Several samples are taken across the country, and the mean score on a depression inventory is computed for each sample. The data are summarized in the following table. What is the mean depression score for all women?

	East	Midwest	West
M	12	19	14
n	46	29	32

15 A school psychologist obtains the following sample of IQ scores from a local high school. What are (a) the mean and (b) the median? (c) Is there a mode?

98	111	101	100	99
99	123	100	134	101
96	102	102	101	105

16 Which measure has the most difficulty with extreme scores? Why?

17 Which measure is the best to use for ordinal data? Why?

18 Which measure is the best to use for nominal data? Why?

19 A population of scores includes 10 numbers ($N = 10$) and has a mean of 100. One of the scores is changed from an 80 to a 90. What is the value of the new mean?

20 A sample of $n = 9$ scores has a mean of 12. If one new score with a value of 5 is added, what is the value of the new mean for the new distribution?

21 A sample of $n = 17$ scores has a mean of 25. After a new score is added to the sample, the new mean is found to be 26. What is the value of the score that was added?

22 A sample of $n = 6$ scores has a mean of 25. If one score with a value of 15 is removed, what is the value for the new mean?

23 Three friends sampled students at their university to see how much time is spent daily on social media. One asked 25 people and got an μ of 45 minutes, another asked 50 people and got an μ of 65 minutes, and a third asked 500 people and got an μ of 52 minutes. All three attempted to randomly sample the student body. What is our best guess of the actual population mean?

Computer Work

24 The distribution in the following list is a hypothetical sample of IQ scores from the incoming freshman class at a university. From the data set, plot a histogram and compute the mean, median, and mode. In what way does the shape of this distribution influence the relative values of the three measures of central tendency?

100	100	102	135	143	94	120	114	111	87	95
109	82	94	142	100	97	100	100	101	99	98
167	176	154	100	85	88	124	180	90	96	92
149	103	102	101	104	92	103	103	105	99	92

25 The following hypothetical data set is all the scores obtained by a statistics class on a final exam. Construct a histogram, compute all the measures of central tendency, and comment on the relative values of the mean, median, and mode in the context of the shape of this population distribution. (Treat "exam score" as a discrete variable.)

35	35	36	50	23	16	22	23	35	35	42	43	47
13	20	9	11	42	23	2	35	40	42	47	22	19
11	8	22	19	8	14	4	28	29	32	41	40	44
2	10	38	33	9	16	22	31	30	35	35	5	20
19	23	35	44	48	34	34	29	33	36	37	37	39
12	35	33	32	33	30	30	29	35	28	39	40	4
11	49	50	35	37	37	38	33	34	35	32	30	28
13	16	17	11	19	18	15	10	35	17	40	41	42

4

Measures of Variability

4.1 The Importance of Measures of Variability

Chapter 3 discussed the three measures of central tendency: the mean, median, and mode. Although conveying central tendency is crucial to the description of a distribution, it is only part of the picture. Measures of central tendency do not provide information about the degree to which scores are spread out in a distribution. If we were asked to imagine two distributions, each with a mean of 100, it would be a mistake to form automatically a mental image of two identically shaped distributions. For example, a platykurtic (mound shaped) and a leptokurtic (pointy shaped) distribution may each have the same mean, but their distributions are very dissimilar. Figure 4.1 shows two very differently shaped distributions, which nonetheless have the same mean, median, and mode.

To complement our measures of centrality, we need to have statistical techniques designed to convey the degree to which scores are spread out and dispersed around a central point. Measures that reflect the amount of variation in the scores of a distribution are called **measures of variability** (or **dispersion**). Several measures of variability along with their advantages and disadvantages will be presented and discussed in this chapter.

4.2 Range

The range is the simplest measure of variability to calculate. The **range** simply reflects the overall span of the scores in a distribution – from the lowest value up to the highest value. The range is calculated by subtracting the lowest score of the distribution from the highest score.

Range

$Range = X_H - X_L$ (Formula 4.1)

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

Figure 4.1 Two distributions with different variabilities yet having the same mean, median, and mode.

where

X_H= highest score in the distribution
X_L= lowest score in the distribution

■ **Question** *What is the range of this distribution?*

17, 44, 50, 23, 42

Solution 50 – 17 = 33 (make sure we organize the data so that the lowest and highest scores can be identified) ■

The next worked problem illustrates one of the main issues that can arise when using the range as a measure of variability.

■ **Question** *What is the range of this distribution?*

2, 4, 5, 7, 34

Solution 32

Do we see how the range can give a misleading impression of dispersion? Most of the numbers are fairly close together, but there is one extreme score (34), which generates a large value for the range; this creates the impression that the distribution of scores is rather spread out. If we are going to use a measure of dispersion that reflects the span of scores, then it would be nice if we could use a measure that is less affected by extreme scores that might lie at either end of the distribution. ■

The Interquartile Range and Semi-Interquartile Range

Every distribution can be divided into four equal sections or quartiles. A **quartile** is one-fourth of a distribution of scores. The bottom 25% of the

values in a distribution make up the first quartile. The second quartile marks the next 25% of scores in the distribution. The *total* percentage of scores below the second quartile is 50%. The median, in fact, is located at the end of the second quartile of a distribution. The third quartile is located at the value that marks the bottom 75% of scores in a distribution. The upper 25% of the scores in a distribution define the fourth quartile.

A **percentile** is a distribution value corresponding to a certain percentage of scores that fall below it. Therefore, the 20th percentile is the value at which 20% of the distribution's scores fall below. The first quartile ends at the 25th percentile, the second quartile ends at the 50th percentile, and so on.

The **interquartile range (IQR)** is the span of scores between the first and third quartiles of the distribution. Stated in terms of percentiles, the IQR is the span of scores between the 25th percentile and the 75th percentile. This measure effectively lops off the upper 25% and lower 25% of the distribution. The two numbers that define the IQR bracket the middle 50% of the distribution (see Figure 4.2). In removing the outermost quartiles, the IQR solves the problems created by extreme scores by only focusing on the middle half of the distribution.

Interquartile range, IQR

$IQR = Q_3 - Q_1$ (Formula 4.2)

where

Q_3= the third quartile (75th percentile)
Q_1= the first quartile (25th percentile)

The **semi-interquartile range (SIQR)** is the IQR divided by 2.

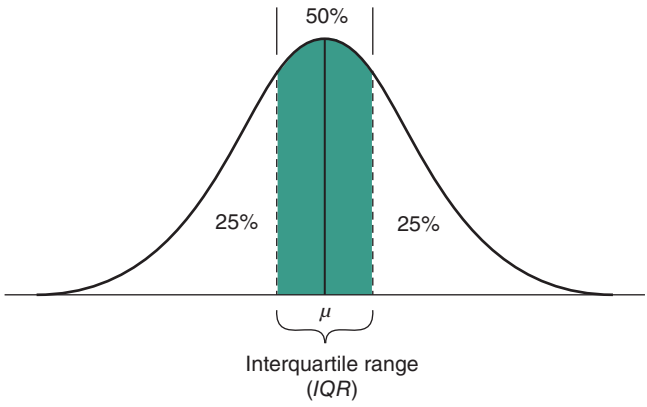


Figure 4.2 The interquartile range (*IQR*) spans the width of the middle 50% of the distribution.

Semi-interquartile range, SIQR

$$SIQR = \frac{Q_3 - Q_1}{2} \quad \text{(Formula 4.3)}$$

Like the *IQR*, the *SIQR* is more stable than the simple range because it is unaffected by an extreme score. Although neither the *IQR* nor the *SIQR* is affected by a single extreme score, they are influenced by distributions with numerous extreme scores, like skewed distributions.

■ **Question** *The hypothetical "Highbridge Community College Aptitude Test" has a median of 100. The score that is the 75th percentile is 130, and the score that is the 25th percentile is 70. What are the IQR and SIQR?*

Solution

$$IQR = Q_3 - Q_1 = 130 - 70 = \mathbf{60}$$

$$SIQR = \frac{Q_3 - Q_1}{2} = \frac{60}{2} = \mathbf{30} \quad ■$$

The family of measures associated with the range can be very helpful measures of variability. For instance, the *IQR* and *SIQR* are commonly presented, along with other descriptive statistics, when conveying the distribution characteristics of standardized psychological or intellectual tests (e.g. the Wechsler Adult Intelligence Scale). (Recall that many psychological concepts, like intelligence, are measured on scales that have some ordinal-like properties and some interval-like properties. In these instances, sometimes researchers choose to use either the *IQR* and/or the *SIQR* to communicate variability.) In these measures, every number is counted, but the specific distance between a number and the mean is not taken into account. A more useful measure for interval or ratio data would be one that takes into consideration the specific distance of every score from the mean. These measures of variability have become the most valuable to researchers interested in statistical analysis.

4.3 Mean Deviation

Each raw score in a distribution of interval or ratio scores sits at some distance from the mean. In Chapter 3 we learned that this distance is called a deviation or error ($X - \mu$ or $X - M = x$). The degree to which scores deviate from the mean is a direct reflection of the variability of a distribution. Consider these two distributions:

Distribution *A*: 11, 12, 13, 14, 15, 16, 17 $\mu = 14$
Distribution *B*: 5, 8, 11, 14, 17, 20, 23 $\mu = 14$

The mean of each distribution is 14. However, Distribution *B* shows more variation than Distribution *A*. In other words, in relation to the mean, there is an overall greater amount of deviation among the scores with respect to the mean. But how can we arrive at a measure that reflects overall deviation? We cannot simply sum the deviation scores of each distribution because Σx always equals 0. This is because the negative deviation scores from below the mean always *balance* the positive deviation scores from above the mean, that is, after all, how the mean defines centrality. However, taking the absolute value of each deviation score will remove the negative signs and free us from this problem. Taking the average of the absolute values of all deviation scores will give us a measure of variability. This arithmetic manipulation is called the **mean deviation**. Formulas 4.4 and 4.5 and the question that follows show how we can calculate this measure of dispersion.

Mean deviation for population, *MD*

$$MD = \frac{\Sigma |X - \mu|}{N} \quad \text{(Formula 4.4)}$$

Mean deviation for sample, *MD*

$$MD = \frac{\Sigma |X - M|}{n} \quad \text{(Formula 4.5)}$$

■ **Question** *What are the mean deviations for the two previously mentioned distributions having the same mean?*

Distribution *A*				Distribution *B*			
Scores	μ	$(X - \mu)$	$\lvert X - \mu \rvert$	***Scores***	μ	$(X - \mu)$	$\lvert X - \mu \rvert$
11	14	−3	3	5	14	− 9	9
12	14	−2	2	8	14	− 6	6
13	14	−1	1	11	14	− 3	3
14	14	0	0	14	14	0	0
15	14	1	1	17	14	3	3
16	14	2	2	20	14	6	6
17	14	3	3	23	14	9	9
$N = 7$			$\sum \lvert X - \mu \rvert = 12$	$N = 7$			$\sum \lvert X - \mu \rvert = 36$

Solution

$$MD_A = \frac{\Sigma |X - \mu|}{N} = \frac{12}{7} = 1.71 \quad MD_B = \frac{\Sigma |X - \mu|}{N} = \frac{36}{7} = 5.14 \quad ■$$

As we can see in the examples above, smaller deviation values reflect tighter distributions, and larger deviation values reflect more dispersed distributions. Any formula that uses deviation scores as a measure of variability has the advantage of using the actual magnitude of the difference between each score and the mean in its calculations, unlike the range, *IQR*, and *SIQR*, which only use the relative position of each score. The mean deviation, however, has some undesirable properties due to the use of absolute values that preclude it from being used in the formulas we will be introducing in subsequent chapters.[1] What is needed is a formula that capitalizes on the conceptual basis of deviation scores but does not have the disadvantages that come with using absolute values. Is there something else we could do with the deviation scores that would keep them from summing to zero?

4.4 The Variance

Another way to remove the negative signs when summing deviation scores is to square them, since a negative number multiplied by another negative number yields a positive number. This minor change defines the difference between the concept of a mean deviation and what is called the variance of a distribution. We can define the **variance** as the average squared deviation score; it is symbolized as σ^2 (pronounced "sigma squared"; σ is the Greek lower case of Σ) for population variances and s^2 for sample variances. Please note that the formula for the variance of a *population* is slightly different from the formula for a *sample* variance.

Population variance, σ^2

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N} \quad \text{(Formula 4.6)}$$

where

σ^2 = the symbol for the population variance
X = a raw score
μ = the population mean
N = the number of scores in the population

1 The main problem with the mean deviation has to do with estimating the variability of a population from a sample of scores. The mean deviation of a sample does not bear a consistent relation to the mean deviation of the population from which the sample was drawn. Since much of the field of statistics involves using the characteristics of samples to infer the characteristics of populations, the mean deviation is rarely used.

Sample variance

$$s^2 = \frac{\Sigma(X-M)^2}{n-1} \quad \text{(Formula 4.7)}$$

where

s^2= symbol for sample variance
M= the sample mean
n = the number of scores in the distribution

The Sample Variance as an Unbiased Estimate of the Population Variance

Recall that a sample is a subset of scores drawn from a population. Researchers are always interested in the characteristics of a population; samples are often used to make inferences about a population. Suppose we want to know the mean of a population, but a sample of scores from that population is all that is available. The best estimate of the mean of the population is the mean of the sample. Usually our estimate will be off; rarely is the sample mean identical to the population mean. Sometimes the sample mean will be larger than the population mean, and sometimes it will be smaller. It is important to note that the sample mean is just as likely to be smaller as it is to be larger than the actual population mean. Since both types of errors are equally likely, the sample mean is said to provide an *unbiased* estimate of the population mean. It would be said to be biased if one type of error was more likely than the other. Also, please note that the degree to which the sample mean is off of the population mean decreases as the size of the sample increases. Larger sample sizes yield more accurate estimates of population parameters. This observation will prove to be very useful later on in the text.

In comparing the formulas for the variance of a population and a sample, note that the denominator of the sample variance is $n - 1$, instead of N. (Compare Formulas 4.6 and 4.7). This difference is a correction factor designed to adjust a bias that occurs when using sample variances to estimate population variances. To show this, suppose we take 100 samples from a population (always replacing the scores from a drawn sample before taking the next sample) and compute the variance of each sample, but using the population formula, without the correction factor in the numerator (Formula 4.6). Assume that we know the true population variance. What we would discover is that, of the 100 computed variances, most would be smaller than the true population variance, and only a few would be larger. If we were to apply the *population* formula for variance to a single sample of scores, and then use that value as an estimate of the population variance, we would most likely *underestimate* the size of the population

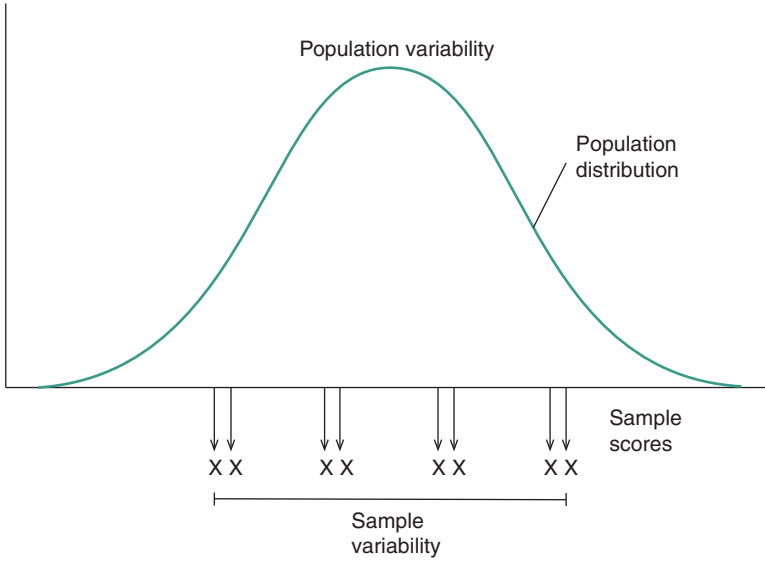


Figure 4.3 The variability of a sample of scores will tend to be less than the variability of the population from which the scores are taken. So that the variance of a sample is an unbiased estimate of the population variance, a correction factor ($n - 1$) is used in the denominator of the formula for the variance of a sample.

variance. Dividing by $n - 1$ provides a correction so that the formula for the sample variance becomes an unbiased estimate of the population variance – just as likely to overestimate as it is to underestimate.

Figure 4.3 gives a visual description of why the $n - 1$ correction factor is necessary when estimating the variance of a population. In this figure, the scores of the population assume a normal distribution. Since most of the population scores are found in the middle of the population, a random sample of, for instance, eight scores would likely come from the middle of the population distribution. As a result, the spread of sample scores is not as spread out as the spread of population scores. For this reason, the variance of a sample will tend to underestimate the variance of a population unless corrected. Placing $n - 1$ in the denominator of the sample variance formula effectively increases the value of the sample variance, providing for a much less biased estimate of the population variance. The sample size also matters. As the size of the sample increases, the sample variance better approximates the population variance. (Recall that this relationship between sample size and statistical accuracy was also true for the mean.)

■ **Question** *What is the variance of this sample of scores?*

3, 4, 6, 8, 9

Solution

X	M	$X - M$	$(X - M)^2$
3	6	–3	9
4	6	–2	4
6	6	0	0
8	6	2	4
9	6	3	9
0	0	0	26

$$s^2 = \frac{\sum (X - M)^2}{n - 1} = \frac{26}{4} = \mathbf{6.5} \blacksquare$$

Equivalent Formulas for the Variance

If we were to take a random sample of introductory-level statistics textbooks and turn to the chapters covering measures of variability, we would be surprised, and maybe confused, by the many different formulas that can be used to compute the variance of a distribution. All of the formulas will give us the same answer (provided we note the distinction between a population and sample of scores). Several formulas for the variance are offered here for two reasons. First, one formula may be easier to use when performing hand calculations with raw data, and a different formula may be helpful in reminding us of the conceptual basis of variance. Second, by presenting a few formulas for the variance, we will be more easily able to make the transition to other textbooks. Formulas 4.6 and 4.7 are the basic formulas for the variances of the population and sample. They are called **definitional formulas** since in reading them, we can be reminded of the concept behind the measurement: the average squared deviation score. In Chapter 3, we learned that a single score minus the mean, $X - \mu$, could be expressed as x, ("little x") a deviation score. Therefore, Formulas 4.6 and 4.7 can be rewritten as Formulas 4.8 and 4.9, respectively. These formulas are called deviation score formulas.

The deviation score formulas

Population variance	*Sample variance*
$\sigma^2 = \frac{\Sigma x^2}{N}$ (Formula 4.8)	$s^2 = \frac{\Sigma x^2}{n - 1}$ (Formula 4.9)

The numerators of both sets of variance formulas direct us to sum all the squared deviation scores. For this reason, the numerator of a variance formula is referred to as the **sum of squares** (or ***SS***). Hence, $SS = \Sigma(X - \mu)^2$ or $\Sigma(X - M)^2 = \Sigma x^2$. Substituting SS in the numerator of the population and sample formulas for the variance defines the SS manner of expression. The SS is a component of numerous statistical formulas.

The sum of squares formulas

Population variance	*Sample variance*
$\sigma^2 = \frac{SS}{N}$ (Formula 4.10)	$s^2 = \frac{SS}{n-1}$ (Formula 4.11)

When working with raw scores, a **computational** (or **raw score**) **formula** eases the calculation task. Formulas 4.12 and 4.13 are used to compute the population and sample variances, respectively. (Yes, they look more involving, but they are actually much easier to use when performing hand calculations, especially as the sample size grows.) When using a computational formula, pay close attention to the difference between ΣX^2 and $(\Sigma X)^2$! The ΣX^2 is found by first squaring each raw score and then summing all squared values. The quantity $(\Sigma X)^2$ requires that we first sum the raw scores and then square the final total. This algebraic distinction is a frequent component in hand calculations of statistical values.

If calculating by hand, it is recommended to simply create two columns, one containing the raw data (labeled X) and the other containing the square of each raw number (labeled X^2). Simply sum up both columns. The sum of the raw score column is (ΣX); by squaring this value we will get $(\Sigma X)^2$. The sum of the squared column is ΣX^2.

The computational formulas

Population variance

$$\sigma^2 = \frac{\Sigma X^2 - \left[(\Sigma X)^2/N\right]}{N} \quad \text{(Formula 4.12)}$$

Sample variance

$$s^2 = \frac{\Sigma X^2 - \left[(\Sigma X)^2/n\right]}{n-1} \quad \text{(Formula 4.13)}$$

Keep in mind that all sample formulas lead to the same answer, with any discrepancies accounted for by rounding errors. Of course, all population formulas also yield the same answer. Table 4.1 presents all of the formulas for the variance.

■ **Question** *Use the computational formulas to determine the variance of this distribution when it is a sample of scores and when it is a population of scores.*

X	X^2
2	4
4	16
5	25
7	49
9	81
$\Sigma X = 27$	$X^2 = 175$

Table 4.1 Several equivalent expressions of the population and sample variances.

Variance formulas	
Population variance	**Sample variance**
Definitional formulas	
$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$ (Formula 4.6)	$s^2 = \frac{\sum(X-M)^2}{n-1}$ (Formula 4.7)
Deviation score formulas	
$\sigma^2 = \frac{\Sigma x^2}{N}$ (Formula 4.8)	$s^2 = \frac{\Sigma x^2}{n-1}$ (Formula 4.9)
Sum of squares formulas	
$\sigma^2 = \frac{SS}{N}$ (Formula 4.10)	$s^2 = \frac{SS}{n-1}$ (Formula 4.11)
Computational formulas[a]	
$\sigma^2 = \frac{\Sigma X^2 - \left[(\Sigma X)^2/N\right]}{N}$ (Formula 4.12)	$s^2 = \frac{\Sigma X^2 - \left[(\Sigma X)^2/n\right]}{n-1}$ (Formula 4.13)

[a] Use these two formulas when working from raw data and calculating by hand.

Solution

Sample Formula

$$s^2 = \frac{\Sigma X^2 - \left[(\Sigma X)^2/n\right]}{n-1}$$

$$s^2 = \frac{175 - \left[(27)^2/5\right]}{5-1}$$

$$s^2 = \frac{175 - (729/5)}{5-1}$$

$$s^2 = \frac{175 - 145.80}{4}$$

$$s^2 = \frac{29.20}{4}$$

$$s^2 = \mathbf{7.30}$$

If the distribution were a sample of scores, the variance would be 7.30. If the scores were a population, we would use the following formula.

Population Formula

$$\sigma^2 = \frac{\Sigma X^2 - \left[(\Sigma X)^2/N\right]}{N}$$

$$\sigma^2 = \frac{175 - \left[(729)/5\right]}{5}$$

$$\sigma^2 = \frac{175 - 145.80}{5}$$

$$\sigma^2 = \frac{29.20}{5}$$

$$\sigma^2 = \mathbf{5.84}$$

Viewing the scores as a population, the variance is 5.84. ■

Sometimes an investigator will learn something important about a phenomenon when the dispersion of scores is examined. Box 4.1 presents a finding in which the variability of scores reflects an interesting aspect of aging.

Box 4.1 The Substantive Importance of the Variance

Measures of variation are essential indices for describing the degree of dispersion among scores of a distribution. The variance and its square root, the standard deviation, can both be used as descriptive measures of dispersion; however, the standard deviation is the more useful measure because it is stated in the original units of the measured variable. Yet the variance is still used in many statistical formulas designed to answer research questions.

In experimental research, comparisons are typically made between the means of two conditions. Evaluating two methods for improving communication skills would entail a comparison between the group *means* of some measure of communication. Discovering ways to help children overcome their shyness would involve comparing *mean ratings* of shyness after different treatments. In other words, in the experimental context, investigators examine group means to determine the effect of the independent variable on the dependent variable. However, sometimes between-group differences in *variability* are important as well. They reveal an important facet of the phenomenon under investigation. An example in which variability has substantive importance comes from the literature on aging. Chronological age is intrinsically a poor predictor of almost any measure of psychological functioning (Woods & Rusin, 1988). However, as an investigator compares different age groups, they would find that the within-group variability, on a number of cognitive and physiological measures, increases with age (Krauss, 1980). In other words, older individuals are more unlike each other than are younger individuals; their distributions are more spread out compared with the distributions of younger people. As a result, researchers investigating questions related to the aged need to pay careful attention to individual differences. A treatment, for instance, that seems ineffective for some older people may prove highly beneficial to other older people.

4.5 The Standard Deviation

The variance measure of dispersion is especially important because it is used in many statistical formulas. However, it is not the best measure when we want to communicate variability to others. This is because the variance is a squared value; it is not stated in the original units of the measured variable. The value seems inflated when compared with the raw scores or other descriptive statistics like the mean. For example, if someone told us that a sample mean of IQ's was found to be 100 with a variance was 225, we might wonder if there was a small, medium, or large amount of variation among the scores. To resolve this, we can take the square root of the variance; it is called the **standard deviation**. The definitional, deviation score, SS, and computational formulas for the standard deviation are identical to the formulas for the variance, *with the exception that the formulas for the standard deviation are under square root signs*. These two important measures of variability are very similar. If we have one, we are one mathematical step away from the other. The symbols for the population and sample standard deviation are σ and s, respectively; this makes sense given that the symbols for the population and sample variance are σ^2 and s^2, respectively. Using the distribution in the preceding worked example, the sample and population standard deviations would be

$$s = \sqrt{7.30} = \mathbf{2.70} \quad \textit{and} \quad \sigma = \sqrt{5.84} = \mathbf{2.42}$$

Table 4.2 presents the same formulas found in Table 4.1. Now, however, each formula is under a square root sign, thereby making them standard deviation formulas.

The Standard Deviation and the Normal Curve

In Chapter 5, much more will be said about the characteristics and importance of the normal distribution. For now, recall that a normal distribution is symmetrical and bell shaped. When depicted on a graph, a normal distribution is called a normal curve. The standard deviation has a very attractive property when applied to normal curves. This property is one reason the standard deviation is so useful in describing the variability of a distribution. In a normal curve, approximately 68% of the scores will fall between one standard deviation below and one standard deviation above the mean. For instance, if a set of IQ scores has a mean of 100 and a standard deviation of 15, then approximately 68% of the scores in that distribution fall between 85 (100 – 15) and 115 (100 + 15). Furthermore, approximately 95% of the scores will fall between plus and minus two standard deviations from the mean. Finally, nearly all of the scores in a distribution (approximately 99.7%) will fall within plus and minus three standard deviations of the mean. See Figure 4.4 for a visual depiction of what has been

Table 4.2 Several equivalent expressions of the population and sample standard deviations.

Standard deviation formulas	
Population standard deviation	**Sample standard deviation**
Definitional formulas	
$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X-M)^2}{n-1}}$
Deviation score formulas	
$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$	$s = \sqrt{\frac{\Sigma x^2}{n-1}}$
Sum of squares formula	
$\sigma = \sqrt{\frac{SS}{N}}$	$s = \sqrt{\frac{SS}{n-1}}$
Computational formulas[a]	
$\sigma = \sqrt{\frac{\Sigma X^2 - \left[(\Sigma X)^2/N\right]}{N}}$	$s = \sqrt{\frac{\Sigma X^2 - \left[(\Sigma X)^2/n\right]}{n-1}}$

[a] Use these two formulas when working from raw data and calculating by hand.

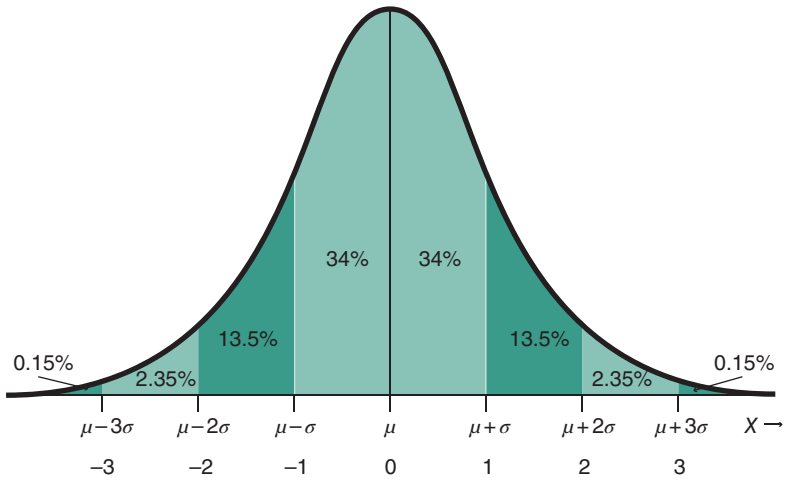


Figure 4.4 The 68-95-99.7 rule describes approximate areas of a normal curve that are respectively within plus and minus one, two, and three standard deviations of the mean.

termed the **68-95-99.7 rule**. This figure displays how the percentages of a normal curve are allocated to areas under the curve in terms of the standard deviation. For instance, the percentage of a normal curve that is between the mean and a score that is two standard deviations below the mean is approximately (13.5% + 34%) 47.5%. Chapter 5 will further develop the idea of a normal curve as a probability distribution.

Please note the standard deviation, or for that matter, any measure of variability, can never be a negative number. Variation is always based on distance, whether it is the span of scores or the average distance scores are from the mean; and there is no such thing as negative distance.

Box 4.2 informs us of the origins of the standard deviation concept.

Box 4.2 The Origins of the Standard Deviation

Karl Pearson (featured in Spotlight 15.1) proposed the standard deviation as a measure of dispersion in 1894. Before Pearson, statisticians used a closely related index of variability: the *probable error* (*pe*). Approximately 68% of the scores of a distribution fall between plus and minus one standard deviation of the mean. The formula for the *pe* is

$$pe = 0.6745\sqrt{\frac{\sum(X-\mu)^2}{N}}$$

The mean, plus and minus one *pe*, includes 50% of the scores of the distribution. (This makes the *pe* concept very similar to the interquartile range.) Pearson believed that multiplying by 0.6475 was an unnecessary step in the calculation. Moreover, as a measure of variability, there was no compelling advantage to the *pe*; both the standard deviation and the *pe* reflect the degree to which a distribution of scores is spread out. Pearson dropped the multiplier, named the new measure of variability the *standard deviation*, and symbolized it as σ. The *pe* is still used, although very infrequently.

4.6 Simple Transformations and Their Effect on the Mean and Variance

As recently as 50 years ago, researchers almost always analyzed data by hand. A single, advanced statistical analysis could literally take an entire week to complete. That same analysis is now performed in a few seconds using modern computer programs. Sometimes the scores of a distribution are very small, for instance, 0.026 or 0.001, or very large, like 10 054 or 11 123. To make the

numbers easier to work with, researchers sometimes transform them by adding, subtracting, multiplying, or diving all scores in a data set by the same value. Performing simple arithmetic operations on a distribution of numbers has predictable effects on the mean and variance:

1) Adding a constant to every score of a distribution will increase the mean of the distribution by the value of the constant. For example, if the mean is 0.01, adding 10 to every score changes the mean to 10.01.
2) Subtracting a constant from every score of the distribution reduces the mean by the value of the constant. If the mean is 1003, subtracting 1000 from every score will change the mean to 3.
3) Adding or subtracting a constant to every score in a distribution *will not* have an effect on the variance. Adding or subtracting a constant does not alter the relative positions among the numbers. The variance is based on the relation each number has to the mean of the distribution. That relation is not altered when adding or subtracting a constant to every score in a distribution.
4) When each score in a distribution is multiplied or divided by a constant, the mean will change by the value of the constant. For example, if the mean is 5, multiplying every score by 2 will render a new mean of (5×2) 10. Dividing every score by 2 would render a new mean of (5/2) 2.5.
5) Multiplying or dividing each score by a constant *will* change the variance exponentially. Multiplication and division will change the relative spacing among the numbers. If multiplying by a constant, the distribution will spread out; the new variance will be the old variance multiplied by the constant squared. If dividing by a constant, the distribution will tighten up; the new variance will be the old variance divided by the constant squared. For example, if a distribution has a variance of 20, multiplying every number by 2 will give us a new variance of (20×2^2) 80, and dividing every number by 2 will give us a new variance of $(20/2^2)$ 5.

Table 4.3 shows the effect on the means and variances when each score of a distribution is transformed by the four basic arithmetic operations.

Table 4.3 The effect on the mean and variance when the scores of the distribution are transformed by the four basic arithmetic operations.

Original scores		+100	−100	×100	÷100
μ	105.5	205.50	5.50	10 550	1.055
σ^2	9.67	9.67	9.67	96 700	0.000 967

4.7 Deciding Which Measure of Variability to Use

In Chapter 3 we discussed the relative strengths and weaknesses of various measures of central tendency to help us select the right measure for a given situation. Here we will look at some of the issues that influence the selection of variability measures.

Extreme Scores

The presence of extreme scores in a distribution, depending on the degree of extremity and the percentage of scores considered extreme, can affect most of the measures of variability. Researchers often look at extreme scores with some degree of suspicion, wondering whether they are an accurate measurement. If they are not accurate, then any measure of variability that is influenced by extreme scores will convey a misleading impression of the actual dispersion of scores in the distribution. The range is clearly the statistic that is most vulnerable to extreme scores. Since only the highest and lowest scores of a distribution are used to compute the range, one extreme erroneous score can lead to a very inaccurate view of dispersion.

The *IQR* and *SIQR* are not much influenced by a *small* number of extreme scores, thereby offering a more reasonable statement of variability when extreme scores are found in a distribution. The variance and standard deviation are also affected by extreme scores. Since both measures use squared deviations, an extreme score that is a great distance from the mean will have a disproportionate effect on the variance, especially for small data sets. We must exercise caution when using the variance and standard deviation as measures of variability when there are extreme scores.

Sometimes researchers, who suspect an extreme score is erroneous, consider discarding that score in an effort to generate a more accurate measure of variability. However, what if there are several extreme scores, as in a skewed distribution? There is no justification for discarding several scores. In these instances, the variability of a distribution is best described by the IQR or SIQR statistic. Furthermore, if the scale of measurement does not allow for the calculation of a mean (e.g. nominal or ordinal scale), then deviation scores cannot be calculated. This eliminates the mean deviation, variance, and standard deviation from consideration.

An Arbitrary End Point to the Distribution

Recall the study discussed in Chapter 3 about self-control and pain tolerance (Grimm & Kanfer, 1976). Participants were asked to place their hands in ice water and were told that they could remove them whenever they wanted. The number of seconds that the participants kept their hands in the water

was the dependent variable. The researchers found that some of the participants did not remove their hands from the ice water and would have continued to keep their hands in it for an unknown length of time. The researchers decided to terminate the task at 300 seconds – an arbitrary end point for the high side of the distribution.

Another example of an arbitrary end point to a distribution occurs when participants are asked to complete a problem-solving task. What score should be assigned to participants who cannot figure out the answer? At some point, the researcher has to stop them and assign a score, which is supposedly the time it took them to complete the problem-solving task. These situations present a problem when we would like to describe the variability of the distribution. Since, in these two examples, the highest score is arbitrary, any measure of variability that relies on these scores will be under-representative of the actual variability and therefore unreliable as a true measure of dispersion. The *IQR* and *SIQR*, however, are relatively impervious to arbitrary cutoffs at the tail end of a distribution, provided there are not too many arbitrary scores.

Common Practice

In common practice, it is rather rare to hear a researcher in the social or behavioral sciences report the *IQR* or *SIQR* when describing a distribution. Researchers in these academic disciplines simply do not have an "intuitive feel" for these measures. Telling a colleague that the *SIQR* of our data is 6 will likely produce a blank stare. The range, despite all its vulnerabilities, is much more likely to be identified than the *IQR* or *SIQR*. However, in many instances where the range is presented, it is only indirectly stated; the highest and lowest scores in the distribution are identified. The variance, despite its essential role in statistical formulas, is also rarely stated among researchers. There is no "intuitive feel" to the variance, being that it is measured in squared units.

If the data is normally distributed, by far the most commonly reported measure of variability is the standard deviation, the square root of the variance. Therefore, if someone says, "The mean was 50," we can bet that the first question asked by another researcher will be, "What was the standard deviation?" Moreover, articles in scientific journals often include a table of means and standard deviations. We will rarely see a table of means and variances or other measures of variability.

Although the standard deviation is the most popular variability measure, we should not ignore all the other measures. Indeed, researchers may err in relying too much on the standard deviation as the measure of dispersion when describing a distribution. It is the responsibility of the researcher who has the most knowledge of the characteristics of the data to choose the most appropriate measure of variability. Finally, there is no rule in the social or behavioral sciences restricting us to report only one measure. If we believe it would be helpful, we should feel free to report more than one measure of variability or central tendency.

Box 4.3 Is the Scientific Method Broken? Demand Characteristics and Shrinking Variation

Throughout the text a series of several boxes are asking whether the scientific method is broken in light of the nonreproducibility problem currently plaguing the social, behavioral, and medical sciences. In Box 1.1 we looked at the "wallpaper effect" and the difficulty in identifying and controlling all extraneous variables. In Box 2.3 we looked at, among other things, different ways the collection of data may be biased through wording effects and order effects. In this box, let us explore some of the problems that occur in the data gathering process.

Sometimes researchers, even those with the best of intentions and who take precautions to remove their own biases, can influence how others respond merely by their involvement in the study. In general, these are referred to as "demand characteristics" or "experimenter effects." Sometimes the unintentional involvement of the researcher can take a response that might ordinarily be quite varied in a population and shrink it to almost nothing. Perhaps the most famous historical example of demand characteristics features "Clever Hans," the horse that could do mathematics (Pfungst, 1911). Hans was owned by Wilhelm Von Osten, a German schoolteacher who had claimed to teach his horse to add, subtract, multiple, divide, and work with fractions. Hans would answer a question by tapping out numbers with his hoof; his accuracy, though not perfect, was remarkable. It was so remarkable that the initial 1904 investigation by a team of academics concluded that it was not a trick. Three years later, however, a local psychologist, Oskar Pfungst, concluded that Hans was indeed clever, but was not doing mathematics. Instead, Hans had learned a "start" cue and a "stop" cue. The "start" cue was the act of being addressed by Von Osten who then looked down at his hooves. As Hans approached the proper answer, Von Osten, believing Hans was about to stop stomping, would make subtle straightening movements of the body and head. This served as the "stop" cue for the clever horse; his reward of food would be forthcoming. Hans usually landed on the right answer or very close to it. The normal variation of a horse periodically stomping its hooves was now being unintentionally managed by cues from the handler such that the variance of stomps less than or greater than the "right" number shrunk dramatically.

The same problem can take place when researchers are gathering data from other people. Numerous studies suggest that the hopes, expectations, and fears of data gatherers can subtly influence people's responses (e.g. Nichols & Maner, 2008; Rosenthal & Fode, 1963; Rubin, Paolini, & Crisp, 2010), taking what might ordinarily be a very diffuse set of responses and pulling them tightly together around the researchers desired or expected response.

Demand characteristics and the role of the experimenter in the data gathering process are potential problems for many issues that social and behavioral scientists investigate. Even under the best of conditions, responses that may naturally be quite varied can begin to narrow around a certain response and mislead us about the true nature of reality. Perhaps those of us in the social and behavioral sciences need to take the human element into more careful consideration when gathering and interpreting data.

Summary

Measures that reflect the amount of variation in the scores of a distribution are called measures of dispersion or measures of variability. The range defines the overall span of scores. It is calculated by subtracting the lowest score of the distribution from the highest score. Unfortunately, the range is extremely sensitive to extreme scores. The *IQR* avoids this by measuring the span of scores between the first and third quartile (the 25th and 75th percentile). The *SIQR* is the IQR divided by 2.

Another family of dispersion measures takes into account the magnitude of difference between each raw score and the mean – the deviation scores. For example, taking the average of the absolute values of all deviation scores is called the mean deviation. Taking the average of all the squared deviation scores is called the variance. The population variance formula is used when the scores represent a population. If our intent is to infer the variance of a population based on a sample of scores, then the formula for a sample variance is used. It contains a correction factor in the denominator to make it an unbiased estimate of the population variance.

The variance measure of dispersion is particularly important because it is used in many statistical formulas. It is not, however, the best descriptive measure of variability. This is because the variance is a squared value; it is not stated in the original units of the measured variable. The standard deviation is the square root of the variance. As a descriptive measure, the standard deviation improves on the variance by converting the measure back to the original units of the measured variable.

If the scores are normally distributed, the standard deviation can be used to analyze probabilistically a data set. The 68-95-99.7 rule states that the mean plus and minus one standard deviation encompasses roughly 68% of the total number of scores in a distribution; plus and minus two standard deviations include approximately 95% of the total number of scores; and plus and minus three standard deviations comprise virtually all scores (99.7%) in a normally distributed data set.

Transforming the original scores of a distribution by the four basic arithmetic operations will have a predictable effect on the mean and variance. Adding or subtracting a constant to each score will alter the mean by that constant. There will be no effect on the variance. Multiplication or division of every score by a constant will alter the mean accordingly and will alter the variance by the constant squared or square-rooted, respectively.

Deciding the most appropriate measure of variability to use depends on various features of the distribution. Factors such as extreme scores, sample size, arbitrary end points of a distribution, common practices, and the importance of a stable estimate of the population variability will influence a researcher's decision as to which measure of variability is most desirable.

Using Microsoft® Excel and SPSS® to Find Measures of Variability

Excel

General instructions for data entry into Excel can be found in Appendix C.

Data Analysis

1) Once the data is entered, select **Data Analysis** and then **Descriptive Statistics**. Click **OK**.
2) Highlight only the scores and put those quadrant numbers into the **Input Range**.
3) Select a location for the output. Use the **Output Range** box if needed.
4) Make sure to click **Summary Statistics** before clicking **OK**.

This should generate a table with the *sample* standard deviation, *sample* variance, and range. To get the *population* standard deviation and variance:

1) Once the data is entered, select a quadrant to receive the value and click the ***fx*** key to the immediate left of the input box. Type in Population variance and select. There are several similar options. Select **VAR.P**.
2) Highlight only the scores in the population distribution, and record them in the **Input Range** box.
3) Click **OK**.
4) Repeat for population standard deviation but use **STDEV.P**.

To find the interquartile range and semi-interquartile range, we will need to do some calculations. But first, follow the steps below:

1) Click the ***fx*** key to the immediate left of the input box at the top of the spreadsheet.
2) In the **Search Box** type **quartile** and hit **Go**. It should pop up in the box below. Select it.
3) Highlight only the scores and put the quadrant numbers into the **Array** box.
4) In the **Quart** box type in "3" (for third quartile). Record that score.
5) Repeat the process, this time typing in "1" in the **Quart** box (for first quartile). Record that score.
6) The interquartile range is the third quartile minus the first.
7) The semi-interquartile range is found by dividing the interquartile range by two.

SPSS

General instructions for inputting data into SPSS can be found in Appendix C.

Data Analysis

1) Once the data has been entered, click **Analyze** on the tool bar, and select **Descriptive Statistics** and then **Frequencies**.
2) Move the column label containing the data we wish to analyze from the left box to the **Variable** box. Also, unclick the **Display Frequency Tables** box in the lower left corner.
3) Before we run it, click on the **Statistic**s box in the upper right corner. Here we will find boxes labeled in the lower left-hand corner under **Dispersion** labeled **Std. deviation**, **Variance**, and **Range**. Click on them. In addition, in the upper left-hand corner, we will find in an area marked **Percentile Values** a box labeled **Quartiles**. Click that box as well, and then click **Continue**.
4) Now we are ready to click **OK**. The first box in the resulting output will give us the *sample* standard deviation, *sample* variance, range, and some quartiles. The Interquartile range is the 75th quartile value minus the 25th quartile value. Divide the resulting value in two to get the semi-interquartile range value.

Unfortunately, SPSS does not generate population variances and population standard deviations. However, we could use SPSS to generate ΣX and ΣX^2 – this will simplify a hand calculation. For ΣX, simply click **SUM** in the **Statistics** Box. To find ΣX^2, we will need to create a new variable and simply square each score – then find the sum of that new variable.

Key Formulas

Range

$Range = X_H - X_L$ (Formula 4.1)

Interquartile range, IQR

$IQR = Q_3 - Q_1$ (Formula 4.2)

Semi-interquartile range, SIQR

$$SIQR = \frac{Q_3 - Q_1}{2} \quad \text{(Formula 4.3)}$$

Mean deviation for population, MD

$$MD = \frac{\Sigma \mid X - \mu \mid}{N} \quad \text{(Formula 4.4)}$$

Mean deviation for sample, MD

$$MD = \frac{\Sigma \mid X - M \mid}{n} \quad \text{(Formula 4.5)}$$

The definitional formulas for the variance

Population variance, σ^2

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N} \quad \text{(Formula 4.6)}$$

Sample variance

$$s^2 = \frac{\Sigma(X-M)^2}{n-1} \quad \text{(Formula 4.7)}$$

The deviation score formulas for the variance

Population variance

$$\sigma^2 = \frac{\Sigma x^2}{N} \quad \text{(Formula 4.8)}$$

Sample variance

$$s^2 = \frac{\Sigma x^2}{n-1} \quad \text{(Formula 4.9)}$$

The sum of squares formulas for the variance

Population variance

$$\sigma^2 = \frac{SS}{N} \quad \text{(Formula 4.10)}$$

Sample variance

$$s^2 = \frac{SS}{n-1} \quad \text{(Formula 4.11)}$$

The computational formulas for the variance

Population variance

$$\sigma^2 = \frac{\Sigma X^2 - \left[(\Sigma X)^2/N\right]}{N} \quad \text{(Formula 4.12)}$$

Sample variance

$$s^2 = \frac{\Sigma X^2 - \left[(\Sigma X)^2/n\right]}{n-1} \quad \text{(Formula 4.13)}$$

Any formula for the standard deviation is the square root of the variance formula.

Key Terms

Measures of variability (or dispersion)
Range
Quartile
Percentile
Interquartile range
Semi-interquartile range
Mean deviation
Variance
Definitional formulas
Sum of squares (SS)
Computational (raw score) formulas
Standard deviation
68-95-99.7 rule

Questions and Exercises

1 Given two samples, one in which $n = 36$, the other where $n = 60$, which distribution would have the larger variance? Which variance is likely to be closer in value to the population variance? Which sample is likely to have the larger range?

2 Assume two samples: $M = 78$ and $M = 155$. Which sample would have the larger variance?

3 A school psychologist wants to inform a teacher about the mean and standard deviation of the students' IQ scores. The scores are below; assume they are a sample.

IQ scores: 98, 111, 102, 100, 101, 109

a What is the mean?
b What is the standard deviation?

4 Calculate the range, variance, and standard deviation of this sample of scores.

2, 4, 7, 4, 8, 5, 1, 4, 4, 5

a What is the range?
b What is the variance?
c What is the standard deviation?

5 A researcher who uses heart rate as the dependent variable finds the 75th percentile to be a heart rate of 111 and the 25th percentile to be at 81.
a Compute the *IQR*.
b Compute the *SIQR*.

6 For the following populations of scores, find the mean deviations:
a 5, 7, 9, 9, 13, 14, 15, 16
b 23, 25, 31, 34, 36, 39, 44, 56, 63, 69
c 6, 8, 3, 9, 1, 4, 7, 4, 1, 1, 8, 2

7 Which of the following variance definitions is correct?
a The average of the deviations scored squared.
b The average of the absolute value of the deviations.
c The average of the deviation scores square rooted.
d The average of the squared deviation scores.
e The average deviation score.

8 What does a distribution with a mean of 50 and standard deviation of zero look like?

9 For each situation, specify whether we should use s or σ.
- **a** A set of coaches are interested in the variability of their basketball team's scores over the season.
- **b** A clinician is evaluating a new treatment for sexual dysfunctions.
- **c** A teacher is interested in providing feedback to students about class performance on the midterm exam.
- **d** A manufacturer takes a sample of light bulbs to estimate the variability of their life.

10 Why is the formula for a sample variance different from the formula for a population variance?

11 Calculate the variance and standard deviation for this population of scores.

22, 32, 21, 20, 19, 15, 23

12 Which distribution of sample scores has the larger variance?

Distribution *A*: 2, 4, 5, 1, 1, 2, 3, 9
Distribution *B*: 34, 39, 34, 35, 33, 32

13 A negatively skewed distribution has a mean of 500 and a standard deviation of 100. Given what we have learned in this chapter, is it possible to determine the percentage of scores that fall between 400 and 600? If so, what is it?

14 What is the main disadvantage in using the range as a measure of dispersion?

15 As a descriptive statistic, is the variance or the standard deviation a better measure of variability? Why?

16 What is the standard deviation of this population of scores?

9, 7, 10, 14, 12, 9, 16, 13, 11

17 If a normal distribution has a mean of 50 and a standard deviation of 10, what scores encapsulate the middle 68% of the distribution? The middle 95% of the distribution? The middle 99.7% of the distribution?

18 What if a normal distribution has the same mean as in question 17, but had a standard deviation of 2. What scores would encapsulate the middle 68, 95, and 99.7% of scores?

19 If a normal distribution has a mean of 80 and 68% of the scores are between 68 and 92, what is the variance of that distribution?

20 If a normal distribution has a variance of 100 and 95% of the scores are between the values of 120 and 160, what is the mean?

21 For a set of 10 000 scores that is normally distributed and has a μ of 100 and a σ of 15, about how many of the scores will be:
a Greater than 130?
b Greater than 115?
c Greater than $\pm 3\ \sigma$ away from 100?
d Greater than $\pm 2\ \sigma$ away from 100?

22 If a distribution has a $M = 4.5$ and $s^2 = 1.6$, what would be the M and s^2 if all the raw scores have 10 added to them?

23 Refer to the data found in Chapter 3, Box 3.2. Compute the standard deviations of the experimental and control groups, for each phase of the study.
a Baseline
b Post-testing

24 An experiment is conducted to evaluate the effectiveness of two different attitude change techniques. The dependent variable is attitudes toward immigrants. In the following table, higher numbers reflect more positive attitudes.

Technique *A*		Technique *B*	
Pretest	**Posttest**	**Pretest**	**Posttest**
3	7	2	4
4	4	3	2
5	6	4	5
2	5	3	3

Technique *A*
Calculate:
a Pretest M
b Pretest s^2
c Pretest s

d Posttest M
e Posttest s^2
f Posttest s

Technique *B*
Calculate:
a Pretest M
b Pretest s^2
c Pretest s
d Posttest M
e Posttest s^2
f Posttest s

25 Complete the following table. $\mu = 50$ and $\sigma = 5$. The constants specified are used to transform the scores of the distribution.

$X + 10$	$X - 10$	$X \cdot 10$	$X \div 10$
$\mu = ?$	$\mu = ?$	$\mu = ?$	$\mu = ?$
$\sigma^2 = ?$	$\sigma^2 = ?$	$\sigma^2 = ?$	$\sigma^2 = ?$

26 What if a newcomer to American football decided to record the yardage gained or lost on each play of a football game in terms of feet instead of the more typical measure of yards; what would the coach need to do with the data to compare the team's performance with previous games?

27 Three friends sampled students at their university to see how much variability there is in daily time spent on social media. One asked 25 people and got an σ of 10 minutes; another asked 50 people and got an σ of 15 minutes; and a third asked 500 people and got an σ of 25 minutes. All three attempted to randomly sample the student body. What is our best guess of the actual population standard deviation?

Computer Work

Determine the range, interquartile range, and semi-interquartile range for each of the following sample distributions.

28 Scores:

13	5	11	17	8	10	13	12	15
15	18	19	16	14	11	12	11	10

(*Continued*)

(Continued)

14	4	5	15	11	10	19	13	14
7	8	5	11	11	9	9	14	15
8	11	17	17	10	9	8	16	14
7	18	6	17	18	18	11	6	13

29 Scores:

43	45	51	27	48	27	43	22	25
45	38	19	26	24	56	42	53	47
54	48	25	39	51	30	29	33	39
27	58	35	33	21	39	35	34	35
18	19	57	51	40	29	28	46	26
37	55	26	47	35	46	53	36	23

Determine the mean, variance, standard deviation, and range for each of the following sample distributions.

30 Scores:

3	5	3	7	9	10	2	12	15
1	8	9	6	4	11	1	11	10
1	4	5	5	3	10	9	13	14

31 Scores:

102	100	99	81	75	113	100
106	114	82	79	88	111	104
100	106	85	99	82	101	100

32 Scores: (For this question only, let us diverge from our commitment to only go out two decimal places – instead, let us go out four, since the values are so small.)

0.1070	0.2190	0.1917	0.2120	0.2016	0.1432	0.1939
0.0988	0.2002	0.1859	0.0847	0.1965	0.1492	0.1861
0.0854	0.1656	0.1776	0.1517	0.1942	0.1812	0.1911

33 Scores:

1020	1000	990	810	750	1130	1000
1060	1140	820	790	880	1110	1040
1000	1060	850	990	820	1010	1000

34 Multiply each value in the data set for Work Problem #32 by three. Find the new mean, variance, standard deviation, and range.

35 Divide each value in the data set for Work Problem #32 by three. Find the new mean, variance, standard deviation, and range.

5

The Normal Curve and Transformations

Percentiles and *z* Scores

5.1 Percentile Rank

How did Andrew do on his last exam? This simple question raises the issue of how best to convey a person's level of performance. Stating that Andrew received a score of 35 may not be particularly helpful since it fails to provide context for the score. Stating that Andrew's 35 was a *B* is better because it provides a rough indication of how he did with respect to some absolute standard. Stating the mean of the distribution, and perhaps the lowest and highest scores of the distribution, would be additionally helpful in defining context; however, this information will not *specifically* locate his score in the distribution. Locating the score based on its *relation to the other scores* in the distribution would further assist in giving meaning to the score of 35.

One method used to specify the relative position of a score in a distribution involves transforming it into a percentile rank. The **percentile rank** of a particular score states the percentage of scores that fall at or below that score in the distribution. For instance, if 54% of the scores of the distribution fall at or below the score of 35, the percentile rank corresponding to the score of 35 is 54%. However, if instead we had a percent in mind and wanted to find the corresponding score, we will be asking for a *percentile*. These terms are very similar.

Perhaps keeping in mind the two different types of questions these concepts answer will help. One question starts with a score and asks for the percent at or below it (e.g. "What is the percentile rank for the score of 35?"). The other starts with a desired percent of the distribution and asks for the corresponding score (e.g. "What score is at the 75th percentile?"). Formulas related to each of these questions are below.

Computing the Percentile Rank of a Score

The formula for transforming a score to its percentile rank is very easy to use.

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

Formula for finding the Percentile Rank of *X*, *PR*

$$PR\ of\ X = \left(\frac{B + 1/2E}{N}\right) \cdot 100 \quad \text{(Formula 5.1)}$$

where

B = the number of scores *below* the given score X
E = the number of scores *exactly* the same as X (if there is only one X score, then $E = 1$)
N = the total number of scores in the distribution

■ **Question** *For the following distribution, what is the percentile rank of 16?*

12, 13, 13, 14, 16, 18, 22

Solution $N = 7$, $B = 4$, $E = 1$. Using Formula 5.1, we find that

$$PR\ of\ 16 = \left(\frac{4 + 1/2(1)}{7}\right) \cdot 100$$

$$PR\ of\ 16 = \left(\frac{4 + 0.50}{7}\right) \cdot 100$$

$$PR\ of\ 16 = (0.64)100$$

$$PR\ of\ 16 = \mathbf{64\%}$$

This means we can say that 64% of the scores of this distribution fall at or below a score of 16. The score 16 is at the 64th percentile. ■

For small samples like in the example above, the percentile rank concept becomes a bit fuzzy. On the face of it, we see that the score of 16 is the fifth number up from the bottom. Since we are looking to find the percent of scores at or below 16 and there are seven numbers, we should be diving 5 by 7 to get 71%, correct? Well, we need to remember that the score of 16, if on a continuous scale, is actually at the midpoint of the interval between 15.5 and 16.5. In a larger sample (with, say, 1000 values), we might have multiple 16's. These "16's" are assumed to be evenly distributed. That is, we would assume that about half of them, if measured more precisely, would have values between 15.5 and 16, therefore at or below 16. The other half, if measured more precisely, would be assumed to have values between 16 and 16.5, therefore above 16. This is why we use half of this frequency number when calculating percentile rank.

Finding a Score Value Given the Percentile Rank

Suppose we administer an achievement test to a group of students. Formula 5.1 can be used to transform each student's score to a percentile rank; this, in turn, allows us to determine how any one student scored on the test with respect to

the group. However, what if we wanted to work things the other way? Instead of using a score to find a percentile rank, we would use a percentile rank to find a score. Formula 5.2 is used to identify the score that is a given percentile rank.

Formula for finding X given a Percentile Rank, X_p

$$X_p = L + \left(\frac{(N)(P) - F}{f}\right) \cdot h \quad \text{(Formula 5.2)}$$

where

X_p = the score at a given percentile
N = the total number of scores in the distribution
P = the desired percentile, expressed as a proportion
L = the exact lower limit of the class interval
F = the sum of all frequencies below L
f = the number of scores in the critical interval
h = the width of the critical interval

Although Formula 5.2 requires us to determine six values to find the score at a given percentile rank, determining these values is really quite easy. The computational steps of Formula 5.2 are specified in the context of the next worked example.

■ **Question** *Two hundred twenty-four students are administered an achievement test. The grouped frequency distribution of the obtained scores is presented in the following table. What score is at the 85th percentile?*

Class Interval	Frequency	Cum f
450–499	15	224
400–449	**29**	**209**
350–399	46	180
300–349	65	134
250–299	32	69
200–249	20	37
150–199	9	17
100–149	8	8

Solution

Step 1. Determine N. The highest number in the *cum f* column is the total number of scores in the distribution. $N = 224$.

Step 2. Identify the critical interval within which lies the score at the 85th percentile. We want to find the score below which falls 85% of the total

number of scores. Eighty-five percent of 224 = (0.85)(224) = 190.40, or rounded, 190. The interval 400–449 contains the 181st through the 209th scores. The 190th score is somewhere in this interval.

Step 3. Determine *L*. The exact lower limit of the critical interval is 399.5.

Step 4. Determine *F*. *F* is the sum of the scores below the critical interval. Simply look at the cumulative frequency just below the critical interval: $F = 180$.

Step 5. Determine *f*. The number of scores in the critical interval is 29. It is assumed that the scores within the interval are evenly distributed.

Step 6. Determine *h*. The real limits of the critical interval are 399.5–449.5. The interval width, *h*, is 449.5 – 399.5 = 50.

Step 7. Determine *P*. *P* is the desired percentile rank, stated as a proportion. $P = 0.85$.

Step 8. Plug the preceding values into Formula 5.2.

$$X_p = L + \left(\frac{(N)(P) - F}{f}\right) \cdot h$$

$$X_{85} = 399.50 + \left(\frac{(224)(0.85) - 180}{29}\right) \cdot 50$$

$$X_{85} = 399.50 + \left(\frac{190.40 - 180}{29}\right) \cdot 50$$

$$X_{85} = 399.50 + [(0.36)(50)]$$

$$X_{85} = 399.50 + 18$$

$$X_{85} = 417.50 \textit{ or } \mathbf{418}$$

The score that is at the 85th percentile is 418. Alternatively, 85% of the achievement scores fall below a score of 418. ■

Notice that inner workings of Formula 5.2 bear a striking resemblance to the intuitive calculation of the median. This should come as no surprise since the median is the 50th percentile. If we wanted to find the median of the foregoing distribution, we would begin by multiplying 224 by 0.50 instead of 0.85. It is important to note that every percentile rank should be considered an approximation due to the assumption that scores are evenly distributed across an interval.

Some Characteristics of Percentile Ranks

Consider the following worked example.

■ **Question** *LaMarr and Margaret are in different statistics classes. They each scored 42 on the midterm. Who did better?*

LaMarr's Class

50	49	49	47	44	42	42	42	42	41	39	37	37	36

Margaret's Class

44	44	43	42	41	40	40	39	39	35	32	30

Solution Using Formula 5.1,

$$PR\ of\ X = \left(\frac{B + 1/2E}{N}\right) \cdot 100$$

$$LaMarr's\ PR = \left(\frac{5 + 1/2(4)}{14}\right) \cdot 100 = \mathbf{50}\%$$

$$Margaret's\ PR = \left(\frac{8 + 1/2(1)}{12}\right) \cdot 100 = \mathbf{71}\%$$

Even though both students received identical scores, Margaret's performance could be considered superior since her percentile rank was higher, that is, if we are primarily interested in performance as performance relative to others in a distribution. Of course, we are not always interested in measuring performance in this manner. For example, if the tests taken in the two courses were identical, it might be reasonable to conclude that neither student outperformed the other. ■

Suppose we had the task of admitting students into our university. It would probably be a mistake to base our decision for admittance on only students' percentile ranks taken from high school grade point averages. Surely, some high schools have different degrees of rigor associated with their classes. Some data sets, even if they are ostensibly measuring the same thing, may not be equivalent. This is one reason why undergraduate admission committees also consider students' performances on nationally standardized tests (e.g. the Scholastic Aptitude Test [SAT]). The SAT is the same measure throughout the country. By using the percentile ranks of SAT scores, a student from Redding, California, can be compared with a student from Ypsilanti, Michigan.

However, the manner in which the percentile rank locates scores has an important weakness. Percentile ranks are based solely on the rank ordering of scores. (Recall the limitations associated with ordinal scales discussed in Chapter 2.) Similar to the median, a percentile rank is determined merely by the relative position of the scores. The magnitude of the difference, if it can even be determined, is not taken into account. The next worked example highlights this difficulty.

■ **Question** *What are the percentile ranks for the score of 80 in both distributions A and B?*

Solution

Distribution *A*	Distribution *B*
82	90
81	89
81	87
81	82
80	**80**
60	79
60	79
59	79
58	77
56	76
40	76

$$\textit{Distribution } A:\ PR = \left(\frac{6 + 1/2(1)}{11}\right)\cdot 100 = \mathbf{59}\%$$

$$\textit{Distribution } B:\ PR = \left(\frac{6 + 1/2(1)}{11}\right)\cdot 100 = \mathbf{59}\%$$

Even though the percentile rank of 59% is the same, the distributions are clearly different. The score of 80 is 2 away from the top and 40 away from the bottom in Distribution *A*, whereas it is 10 away from the top and 4 away from the bottom in Distribution *B*. If the underlying scale is ordinal, then this is an unresolvable problem. The percentile rank does accurately communicate the relative position of the value 80 in both distributions, and no improved descriptor can be calculated. Percentile ranks only consider the number of scores below a given value, not the magnitude of differences between those scores. ■

One clear advantage of percentile ranks is that they are versatile; they can be used with any shaped distribution, skewed, or otherwise. However, just as an ordinal scale can communicate relative position but not the magnitude of differences between values, so also are the limitations of percentile ranks. A different system, one that measures *how much* more or less one value is from another, will need to be employed to locate more precisely a given score within a distribution of scores.

5.2 The Normal Distributions

Chapters 2–4 introduced us to the notion of a normal distribution (curve). In this section, the normal curve will be discussed in much greater detail. From this point in the text through Chapter 16, the normal curve will be heavily relied upon and extensively used, so much so that data sets will be assumed to be normally distributed unless there is specific information to the contrary.

The Importance of Normal Distributions

Normal distributions are of fundamental importance in the field of statistics for two reasons. First, it is a very common distribution shape. Measurements of many naturally occurring phenomena, including psychological concepts like intelligence, anxiety, mood, and so on, are normally distributed. Second, if we were to take a sample of scores from any shaped population, calculate M, then replace the scores and take another sample of the same size, calculate M, replace them, and so on until we had an exceedingly large number of sample means; those *means* would be normally distributed. This second observation forms an important theoretical basis for most statistical analyses used to test hypotheses. This point will be developed extensively in Chapter 7.

Characteristics of Normal Distributions

For a distribution to be called normal, it must conform to a certain mathematical model:

$$y = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X-\mu)^2/2\sigma^2}$$

where

y = the ordinate on the graph, that is, the height of the curve for a given X
X = any given score
μ = population mean
σ^2 = population variance
π = the value of pi: 3.1416 (rounded)
e = 2.718 (rounded), the base of the system of natural logarithms

Do not experience "formula shock"; we will likely never use this equation. However, this formula can be used to make some valid points about normal distributions. The formula for a normal curve is a general formula; it is not tied to a specific set of scores. All the values in the equation are fixed, except for X, μ, and

σ^2, which will vary from distribution to distribution. To draw any curve, we need to know, for each X, how far up on the graph to go to plot a point. This distance along the ordinate (y) reflects the relative frequency of scores for the given X. The distance along the ordinate is different for every combination of μ and σ^2. The mean locates the center of the distribution on the abscissa, and the variance indicates the degree of dispersion among the scores. Once μ and σ^2 are specified, as X scores inserted into the equation increasingly deviate from μ, y becomes smaller. What this means is that scores *close to* the mean occur more frequently (higher on the ordinate) and scores *far away from* the mean occur with less frequency (lower on the ordinate). It follows that there are an infinite number of normal distributions since μ and σ^2 can take any value. Hence, one refers to the *family of normal distributions*. Nonetheless, all the normal distributions in the family share five characteristics:

1) A normal distribution is unimodal, meaning it has one hump.
2) A normal distribution is symmetrical. This means that the right half of the curve is a mirror image of the left half. If we were to fold the curve at the midpoint, the two sides of the distribution would coincide.
3) A normal distribution has the same value for the mean, median, and mode. This follows from the fact that a normal distribution is unimodal and symmetric.
4) A normal distribution is **asymptotic**; the tails of the distribution never touch the abscissa. This is merely a theoretical point, but dictated by the mathematical model. This means a normal curve will always have some relative frequency associated with every value of X, even those extremely far away from the mean. However, just because theoretically any X value can be placed in the equation does not mean that any X score can be found in reality. For instance, if we are plotting the height of adult males, there will be a very small and yet nonzero frequency associated with the value of 20 ft tall, even though in reality there is absolutely no chance of ever finding a person of that height.
5) In a normal curve, approximately 68% of the scores in the distribution lie between $\mu \pm \sigma$, approximately 95% of scores in the distribution lie between $\mu \pm 2\sigma$, and approximately 99.7% of scores in the distribution lie between $\mu \pm 3\sigma$. This is the 68-95-99.7 rule mentioned in Chapter 4.

There may be other distributions having some of the characteristics of the preceding list, but only the family of normal distributions will share all five characteristics. It is easy to forget that there is a family of normal distributions because all statistics books use very similar drawings to depict a normal curve. Moreover, a normal curve is often referred to as *the* normal curve, as if there is just one. Keep in mind that any curve that possesses all of the foregoing five characteristics is a normal curve. Figure 5.1 illustrates three "members" in the family of normal distributions.

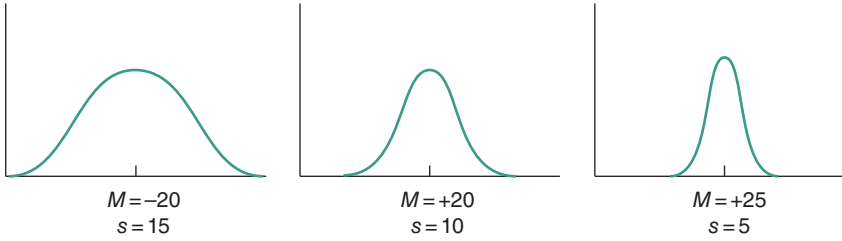


Figure 5.1 Three normal distributions that differ in their *M*'s and *s*'s.

The concept of the normal curve has broad application. Behavioral and social scientists base many of their statistical analyses on the normal curve, yet the discovery of the normal distribution emanates from a mathematician's interest in gambling. Spotlight 5.1 traces some notable historical advances in the use of the normal curve.

Spotlight 5.1 Abraham De Moivre and the History of the Normal Curve

The discovery of the normal curve is usually attributed to Abraham De Moivre (1667–1754), being traced to a publication of his from 1733 (De Moivre 1738/1959; English Translation). He was a friend of people like Edmond Halley (of Halley's Comet fame) and Sir Isaac Newton and was held in high esteem by the intellectual class of his time. Apparently, Newton occasionally replied to questions with, "Ask Mr. De Moivre, he knows all that better than I do" (Walker, 1934, p. 322). De Moivre's discovery grew out of his interest in the probability of chance occurrences, in which an event could take on one of two values. Imagine we wanted to know the probability of getting between 500 and 600 heads after tossing a coin 1000 times. Given repeated tosses, if the number of tails is plotted on the horizontal axis and the probability of obtaining any number of tails is plotted on the vertical axis, as the number of tosses increase, the graph begins to take on the shape of a normal curve. How De Moivre arrived at the formula for the normal curve is unclear, since it was the writing style of the day to publish results and conceal methods. It would be interesting to know De Moivre's thought process because he used two constants from areas not associated with statistics: π, the ratio of the circumference of a circle to its diameter, and e, a constant that is used in calculating financial growth rates, exemplified by interest-bearing accounts.

Although the context of De Moivre's discovery had to do with binomial events (i.e. two possible outcomes) and gambling procedures, the curve has extensive application to errors of measurement, a fact that has made it useful for the work of observing the movements of planets and stars. In addition, the curve approximates all sorts of raw score distributions, such as the heights of

adult males in a given homogeneous population, obtained IQ scores for a given population, some personality variables, and innumerable naturally occurring variables, including things like the weight of a given species of wild rabbits and even the widths of foreheads of crabs. It was De Moivre who worked out the area under the curve for distances up to three standard deviations. Because he worked out the formula for the normal curve and the areas under the curve, De Moivre is duly credited with the normal curve's discovery. However, other mathematicians deserve credit for popularizing it.

The first person to extend the normal curve to continuous measures was the English mathematician Thomas Simpson (1755). Suppose we are interested in determining the position of a star. Each independent observation will yield a slightly different number. The amount of variation among the numbers would be a function of the reliability of our instrument. However, which observation should we trust? It seems obvious today that the best way to handle this problem would be to minimize the error by taking the mean of all observations. However, this was not the practice prior to Simpson's recommendation. He, however, understood how the normal curve could be used to counterbalance positive and negative error out of a final judgment.

The idea of the normal distribution was extended even further when the French mathematician Pierre Laplace (1749–1827) proved the Central Limit Theorem (a concept we will explore more in Chapter 7). This theorem is the single most important theorem in statistics. It allows for the use of sampling distributions in hypothesis testing. We will learn about the central role of this theorem in later chapters. In essence, the Central Limit Theorem states that the means of many samples from a population will be normally distributed. This allows us to use the normal distribution to figure the probability of obtaining a mean by chance.

The German mathematician Carl Gauss (1777–1855) also popularized the normal distribution. In fact, the normal distribution is also called the Gaussian distribution. Gauss was one of the greatest mathematicians who ever lived and one who showed promise right from the start. For instance, at the age of 3, it is said he discovered an error in his father's calculations of employee wages! Later in Gauss' life, when working at the University of Gottingen, Napoleon's armies were advancing on the city. Laplace is reported to have contacted Napoleon, his longtime friend, asking him to spare Gottingen because "the foremost mathematician of his time lives there" (Dunnington, 1955, p. 251). Gauss' most lasting contribution was the use of statistics to relocate an asteroid. With only a few observations with which to work, Gauss predicted exactly where the asteroid Ceres would reappear. He was using the *method of least squares*, a method he invented that eventually found its way into modern statistics.

Many mathematical curves today serve as statistical models. However, De Moivre's normal curve serves as the cornerstone of descriptive and inferential statistics.

Area Under the Normal Curve

We will recall from Chapter 2 that a simple frequency distribution can be depicted as a histogram or a frequency polygon. The frequency of every value of X is represented on the graph. Accordingly, we can say that all of the scores in a distribution fall in the area under the curve. This allows us to begin to think of a normal curve using terms related to probability or likelihood. Chapter 6 will focus on probability theory, but for the time being we can understand probability to mean the likelihood that a given event (e.g. flipping a coin and getting a "head") will occur or not occur. This likelihood is quantified using the range of 0–1. A probability of 0 means the event in question cannot occur, while a probability of 1 means the event in question is most certain to occur. Probability theory is most useful when thinking about events that may or may not occur. For instance, an event that is just as likely to occur as it is not to occur (say, getting a "heads" on a coin flip) has a probability of .50. Since the area under the curve includes all of the scores of a distribution, the probability that a score from the distribution will be found under the curve is 1. Furthermore, since a normal curve is symmetrical, the probability that a score selected at random will be greater than the mean is equal to the probability that a score selected at random will be less than the mean. However, since the normal curve is mathematically so well understood, we will learn how to determine the probabilities associated with any value found under a normal curve.

Throughout this chapter and those that follow, references will be made to both the *percentage* of scores and the *probability* of a given score occurring. Please understand that these concepts are used interchangeably such that any point made about a percentage is true for probability. In this way, saying that 50% of scores fall above the mean is the same as saying the *probability* is .50 of randomly selecting a score above the mean. Now that the basic concepts of a normal curve have been addressed, we can turn our attention to a transformation method that allows us to locate precisely the position of a score within a normal distribution

5.3 Standard Scores (*z* Scores)

It is difficult to overstate the importance of standardizing scores to the practice of statistical analysis. The ability to standardize allows us to take data from any set of normally distributed scores, no matter the particular value of the mean or the variance, and think about that distribution in a similar way as any other normally distributed data set. Some may argue it even allows us to compare apples with oranges. (See Box 5.1 for further development of this concept.) The concepts covered in this section of the text are critical for full comprehension of later material.

Box 5.1 With *z* Scores We Can Compare Apples and Oranges

Is he taller than he is heavy? This question, at first glance, seems to be nonsensical, like comparing apples with oranges. The reason the question appears to be unanswerable is that height and weight are different variables and measured in different units. How can we say that 6 ft 2 in. is more or less than 145 lb? However, in the world of statistics we *can* compare the relative position of scores in different distributions by using standardized scores. The *z* score transformation will convert original scores, from different scales, to a common unit. The common unit is the *z* score, which is the number of standard deviations a raw score is from the mean of a given distribution. Now if we were told that a man's height transforms to a *z* of +1.3 and his weight to a *z* of −.42, could we answer the question, "is he taller than he is heavy?" If we first qualified our statement by saying that we are comparing two values relative to the distribution from which they came, then "yes"; we could answer affirmatively. When his height is transformed into a *z* score, the mean and standard deviation of a distribution of heights is used to make the transformation. His weight is transformed into a *z* score using the mean and standard deviation from a distribution of weights. In this way, to say that he is taller than he is heavy is to say that his transformed height value locates him *higher* on the *z* distribution of heights than his transformed weight score locates him on the *z* distribution of weights.

We can also use *z* scores to compare things like test performances on two different tests. Suppose a roommate is gloating a bit because they scored an 88 on a history exam while the other roommate only scored an 82 on their psychology exam. However, we suspect that the history exam was much easier than the psychology one. If we knew the means and standard deviations of both exams (and if we can assume both sets of tests were normally distributed), we could see which of the roommates performed better in relation to the rest of their respective classes.

This way of comparing scores from different scales of measurement is very useful in the social and behavioral sciences as well as in the field of education. We can ask, for instance, if a person is more depressed than anxious, more paranoid than manic, or better at math than at reading. Although the scales of the tests are designed to tap different traits and abilities, and each scale has its own mean and standard deviation, by standardizing the raw scores an examiner can easily make cross-scale comparisons.

A ***z* score** is a measure of how many standard deviations a raw score is from the mean of the distribution. Given a normal distribution, suppose the mean is 20 and the standard deviation is 4. A score of 24 is one standard deviation above the mean. Consequently, a score of 24 would be 1 *z* score above the mean

($z = +1$). What z score would be assigned to a score of 16? Since 16 is one standard deviation below the mean (20 – 4), a score of 16 would transform to a z score of –1.

A z score is also called a standard unit or **standard score**. When all of the raw scores from a *normal distribution* have been transformed into z scores, the resulting distribution is called the **standard normal distribution**. The standard normal distribution is a special distribution; it has a mean of 0 and a standard deviation of 1. This point is so important – it bears repeating; any normal distribution of raw scores if converted into z scores, no matter the mean or the standard deviation, will take the shape of the standard normal distribution, having a mean of 0 and a standard deviation of 1. This makes the standard normal distribution very special.

Two z score formulas are provided; one is used to transform the scores of a population, while the other is used to transform the scores of a sample.

Formulas for transforming an X Score into a z Score

Population	*Sample*
$z = \frac{X - \mu}{\sigma}$	$z = \frac{X - M}{s}$
(Formula 5.3a)	(Formula 5.3b)

where

X = the raw score to be transformed
μ = the mean of the population
σ = the standard deviation of the population
M = the mean of the sample
s = the sample standard deviation

■ **Question** *For a distribution with $\mu = 4.80$ and $\sigma = 2.14$, what is the z* score *of a raw score of 6?*

Solution

$$z = \frac{X - \mu}{\sigma}$$

$$z = \frac{6 - 4.8}{2.14} = \mathbf{0.56}$$

Given the characteristics of this distribution, a score of 6 is 0.56 standard deviations above the mean. ■

The following table lists several characteristics of z scores and the standard unit normal curve.

Important Facts About *z* Scores

1) A *z score* distribution is established by transforming every raw score into a *z* score.
2) A *z* score distribution always has a mean of 0 and a standard deviation of 1.
3) All raw scores that fall below the mean have some *z* value that is negative; all raw scores that fall above the mean have some *z* value that is positive.
4) A raw score that is one standard deviation from the mean has a *z* score of either ±1, depending on whether it is above or below the mean.
5) A raw score that is the same as the mean has a *z* value of 0.
6) The *z* score distribution will have the same shape as the raw score distribution.

Area Under the Curve and *z* Scores

All of the scores in a distribution are contained in the area under the curve. When the distribution is normal, half the scores are above the mean and half the scores are below the mean. In Chapter 4, we learned that approximately 68% of the scores in a normal distribution fall between plus and minus one standard deviation of the mean. In the standard normal distribution, approximately 68% of the *z* scores will fall between a *z* score of ±1 (see Figure 5.2). Statements of percentages can be translated to probability statements. For instance, the probability that a score selected at random will have a positive *z* score is 0.50. The probability that a randomly selected raw score will correspond to be *z* score between ±1 is approximately 0.68.

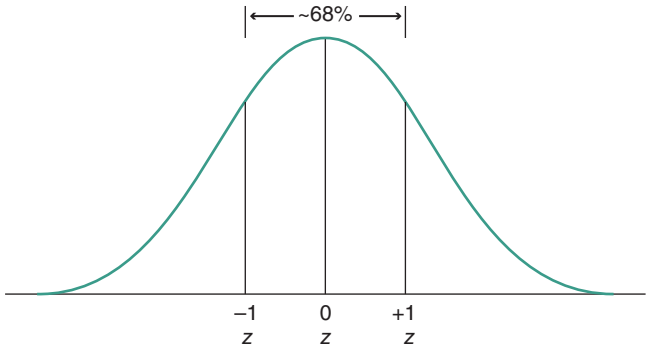


Figure 5.2 Approximately two-thirds of the scores of a normal distribution fall between *z* scores of ±1.

It is mathematically possible to specify the probability that a score will be drawn from a specified range of scores under the curve. However, we have been mercifully spared the arduous task of calculating these probabilities. We can simply make use of the *z* table found in Appendix A (Table A.1). With the aid of this table, we can answer such questions as "What percentage of scores fall below a given *X* score?" or "What is the probability that a score taken at random will fall between any two scores?" The *z* table, however, can only be used when working with data from a normal distribution.

Using the *z* Table

The following is a portion of the *z* table found in Appendix A (Table A.1).

(*A*)	(*B*)	(*C*)
z	Area Between Mean and *z*	Area Beyond *z*
⋮	⋮	⋮
1.00	.3413	.1587
1.01	.3438	.1562
1.02	.3461	.1539
1.03	.3485	.1515
1.04	.3508	.1492
⋮	⋮	⋮

Column *A* of the table lists *z* scores. Column *B* provides the probability that a single score will fall between the mean of the distribution and the *z* value in column *A*. By moving the decimal point two places to the right, the numbers in column *B* would represent the percentage of scores falling between the mean and a given *z* score. Column *C* specifies the probability that a score will fall beyond a particular *z* score, that is, between that score and the end of the distribution on whichever side the *z* score in question falls. In this way, Columns *B* and *C* serve to cut one side of the distribution into sections with the sum of the area of the two sections always equaling 0.50. This means that for any *z* value, the corresponding areas found in column *B* plus column *C* will sum to 0.50. Finally, notice that the *z* table does not depict negative *z* scores. Recall that a normal distribution is symmetrical, and, therefore, the area between the mean and a *z* score of, say, +1 is the same as the area between the mean and a *z* score of −1. Having a table depicting probabilities for only one-half of a perfectly symmetrical distribution is sufficient.

Figure 5.3 shows a more precise measurement of the percentages of scores falling between various points of a normal distribution. Several worked examples will follow in order to familiarize us with the use of the *z* table.

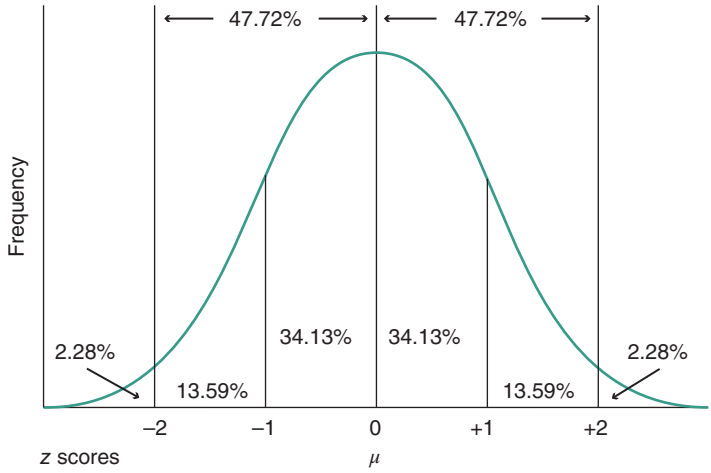


Figure 5.3 The percentage of scores that lie between various points of a normal distribution. This figure represents a more precise representation of the 68-95-99.7 rule presented in Chapter 4.

■ **Question** *What is the probability that a randomly selected score will fall between the mean and a z score of 0.39 (Figure 5.4)?*

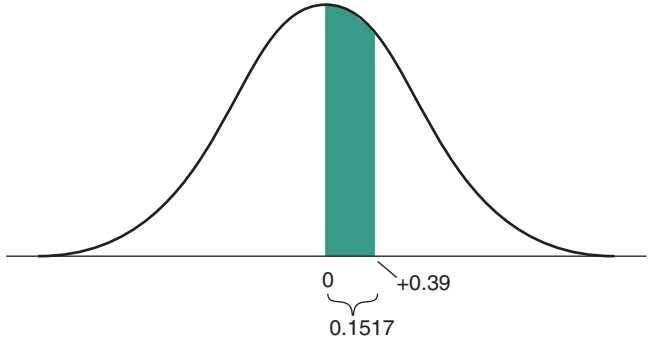


Figure 5.4 The probability that a randomly selected score will fall in the shaded area is 0.1517 or 15.17%.

Solution 0.1517 ■

■ **Question** *What percentage of scores fall between the mean and a z score of 1 (Figure 5.5)?*

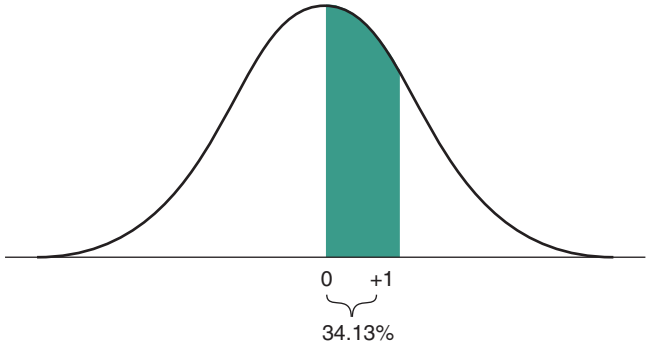


Figure 5.5 The shaded area includes 34.13% of the scores.

Solution 34.13% ■

■ **Question** *What percentage of scores fall between ±1 z score (Figure 5.6)?*

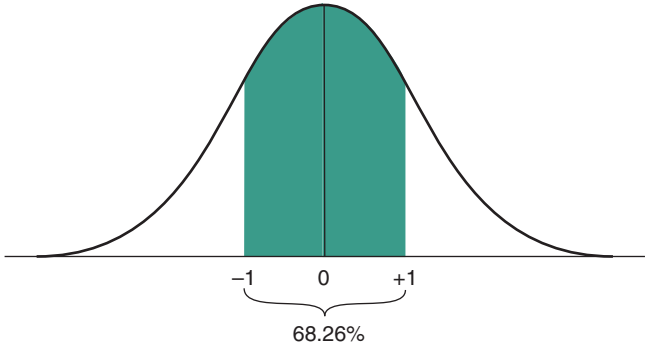


Figure 5.6 Just over 68% of scores fall within ±1 *z* score.

Solution 34.13 + 34.13% = **68.26%** ■

■ **Question** *What is the percentage of scores that fall between a z score of +0.25 and +1.20 (Figure 5.7)?*

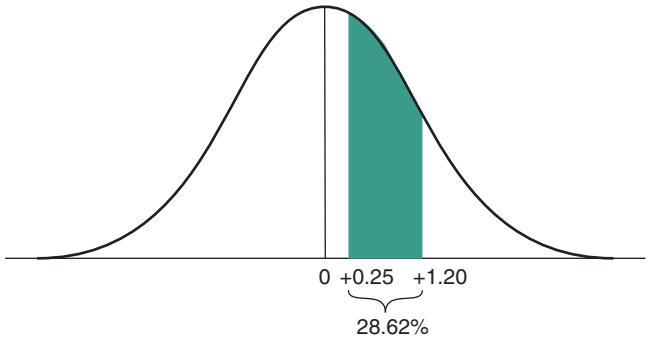


Figure 5.7 The shaded area contains 28.62% of the scores.

Solution The percentage of scores between the mean and a z of +1.20 is 38.49. This also includes the unwanted area between the mean and the z of +0.25. Subtracting the percentage of scores falling between the mean and +0.25 from the percentage of scores found between the mean and a z value of +1.20 will isolate the proper area:

$$38.49\% - 9.87\% = \mathbf{28.62}\% \blacksquare$$

■ **Question** *What is the total percentage of scores that fall above a z score of +1.96 and below a z score of –1.96 (Figure 5.8)?*

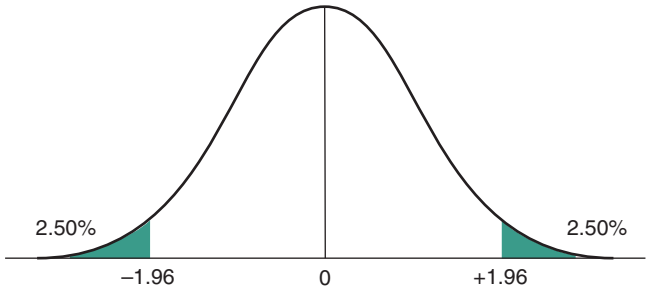


Figure 5.8 The shaded areas contain a total of 5% of the scores.

Solution Use the third column of the table when looking up 1.96. There is 2.50% of scores in each tail of the distribution beyond a z of 1.96. Therefore, the total percentage of scores is 5%. Stated differently, the probability that a score drawn at random will fall *beyond* a z score ±1.96 is .05. Moreover, 95% of all scores fall within the boundaries of ±1.96 z scores. ■

■ **Question** *Find the z score cutoffs within which fall 90% of the scores of a distribution (Figure 5.9).*

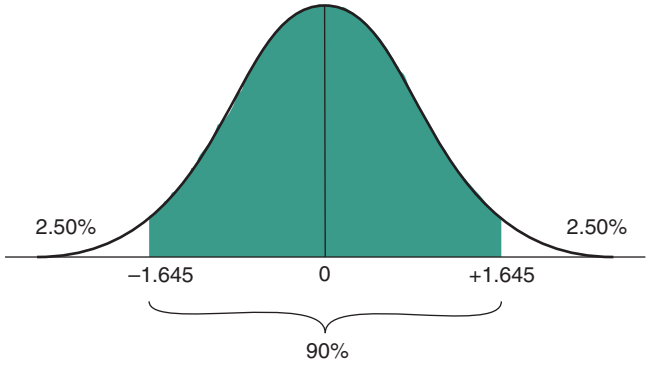


Figure 5.9 Approximately 90% of the scores fall between *z* scores of ±1.645.

Solution First, enter the second column of the table. Next, find the *z* score in column *A* that comes closest to the probability value of .4500. The *z* score we need is between 1.64 and 1.65. We could use either 1.64 or 1.65 and simply note that the area identified will be either slightly smaller or slightly larger than 90%. Alternatively, since the two values seem to be equally close to .4500 (in actuality, because we are talking about the area under a *curved* line, they are not exactly equally close), we could take the midpoint between 1.64 and 1.65 and state that approximately 90% of the scores are within ±1.645. ■

Using the *z* Score Formula

■ **Question** *Given a distribution in which M = 25 and s = 5, what percentage of scores fall between 25 and 32?*

Solution To use the *z* table, the raw scores of 25 and 32 must be transformed to *z* scores. Using Formula 5.3b (the symbols *M* and *s* should tip us off that the scores are from a sample), we find

$$z = \frac{X - M}{s} = \frac{25 - 25}{5} = \frac{0}{5} = 0.$$

$$z = \frac{X - M}{s} = \frac{32 - 25}{5} = \frac{7}{5} = 1.40.$$

Now that we have converted the raw scores into *z* scores, the question can be rephrased as "what percentage of scores fall between a *z* score of 0 (the mean) and 1.40?" The answer is 41.92%. (Next to a *z* score of 1.40, see the value in column *B*.) ■

■ **Question** *Given M = 100 and s = 25, what percentage of scores fall between 75 and 125?*

Solution Well, the standard deviation is 25 units of whatever is being measured. The score of 125 is one standard deviation above the mean, while the score of 75 is one standard deviation below the mean. By now, we should know that approximately 68% of the scores fall within plus and minus one standard deviation of the mean (68.26%, to be exact). ■

Up to now, we have been using z scores to find the percentage of scores falling within a given area of the normal curve, or the probability that a given score will fall within an area. However, sometimes questions are asked that require z scores to be converted into raw scores. Formula 5.4a and b enables us to accomplish this conversion.

Formulas for transforming z to an X Score

Population	*Sample*
$X = \mu + z\sigma$ (Formula 5.4a)	$X = M + zs$ (Formula 5.4b)

■ **Question** *A teacher administers a placement test in order to assign each student to one of three classrooms: an accelerated, a remedial, or a regular class. The regular class will have those students who obtained scores falling within the middle 60% of the distribution. All students scoring in the upper 20% of the distribution will be assigned to the accelerated class, and those receiving scores in the lower 20% of the distribution will be assigned to the remedial class. The mean of the class distribution is 75 with a standard deviation of 7. What raw score cutoffs should be used to make the assignments (Figure 5.10)?*

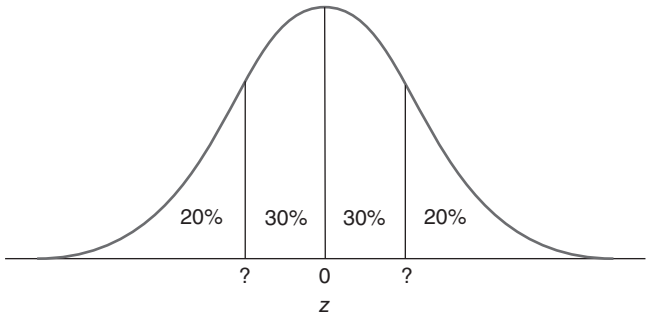


Figure 5.10 What are the cutoffs that bracket the middle 60% of scores in the distribution?

Solution This problem, at first glance, may seem overwhelming. If we represent the problem visually, we can simplify matters. Our illustration should look like Figure 5.10. Since we want the middle 60% of the distribution, we need a z score that has 30% of the scores between it and the mean. Since it is assumed that the distribution of placement test scores is normal, the same percentage of scores that fall between the relevant positive z score and the mean will fall between the *same* negative z score and the mean. Instead of using the first column of the z table, the column of z scores, use the second column to find the percentage closest to 30. The percentage 29.95 is as close as this table allows. The z score that corresponds to 29.95 is +0.84. This means that approximately 30% of the raw scores fall between the mean and a z of +0.84. Since the distribution is symmetrical, another 30% of the scores fall between the mean and a z score of −0.84. Therefore, the middle 60% of the distribution falls between z scores of ±0.84.

We are now in a position to convert the z scores to raw scores. Since we are looking for the raw score cutoffs that correspond to a positive *and* negative z score, two separate calculations are required, both using Formula 5.4 (whether we use version a or b depends on how we define the class, population or sample, but either one will yield the same result):

$$\begin{aligned} \textit{Upper Cutoff} &= 75 + (+0.84)(7) \\ &= 75 + 5.88 \\ &= \mathbf{80.88} \end{aligned}$$

$$\begin{aligned} \textit{Lower Cutoff} &= 75 + (-0.84)(7) \\ &= 75 - 5.88 \\ &= \mathbf{69.12} \end{aligned}$$

Assuming the placement test yields whole number results, we can apply the findings to the distribution of test scores in the following manner: Students scoring 81 or higher are in the top 20% and should be assigned to the accelerated section, students scoring 69 or below are in the lower 20% and should be assigned to the remedial section, and students with scores from 68 to 80 should be assigned to the regular section. ■

Using *z* Scores to Calculate Percentile Rank

Recall that the percentage of scores falling below a given score is the percentile rank of that score. We have just learned that any score can be transformed into a z score. The z table can enable us easily to calculate percentile ranks via z scores. Bear in mind, however, that percentile ranks can only be computed with z scores if the data set is normally distributed.

■ **Question** *What is the percentile rank of the score 15, when M = 18 and s = 4, assuming a normal distribution?*

Solution The *z score* of 15 is

$$z = \frac{15-18}{4} = \frac{-3}{4} = -0.75$$

Use the third column of the *z* table. This allows us to determine the percentage of scores *above* a *positive z* score or *below* a *negative z* score. Figure 5.11 depicts the shaded area that we are going to identify in the third column.

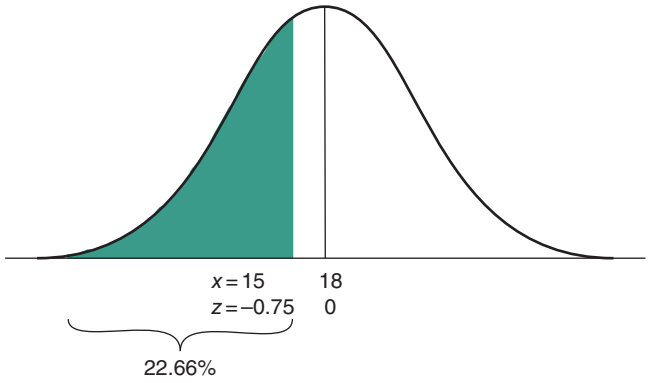


Figure 5.11 In finding the percentile rank of 15, the shaded area is the percentage of scores that fall below a raw score of 15.

The *z* score that corresponds to the raw score of 15 is –0.75. The percentage of scores that fall below a *z* of –0.75 is 22.66. Hence, the percentile rank of 15 is 22.66%. ■

■ **Question** *What is the percentile rank of 30, when M = 27 and s = 2, assuming the data set is normally distributed?*

Solution The *z* score associated with a raw score of 30 is

$$z = \frac{30-27}{2} = \frac{3}{2} = 1.5$$

The percentage of scores between the mean and a *z* of 1.50 is 43.32. However, the percentile rank includes *all* the scores below *X*, so we need to add the lower half of the distribution to 43.32. Therefore, the percentile rank of 30 is 43.32 + 50 = 93.32%. We can calculate percentile ranks using either the second or the third column of the *z* table. If we draw a picture of the normal curve, shade the appropriate area, and understand what the second and third columns of the *z* table are giving us, the proper arithmetic operations will be obvious. ■

Identifying the Interquartile Range

Recall from Chapter 4 that the interquartile range marks the middle 50% of the distribution. It is a descriptive measure of variability that is unaffected by extreme scores. The way in which we go about finding the interquartile range is similar to the way we solved the problem of assigning students to advanced, regular, and remedial classes.

■ **Question** *A distribution has $\mu = 80$ and $\sigma = 5$. What is the interquartile range?*

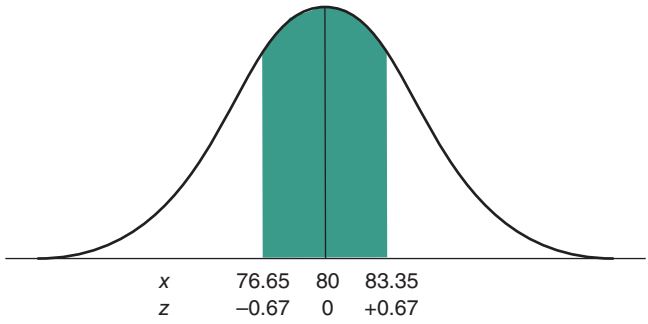


Figure 5.12 The shaded area defines the interquartile range.

Solution Figure 5.12 shows the middle 50% of the distribution. We need to identify the z scores that bracket the middle 50%. Enter the second column of the z table and find the percentage closest to 25%. A z of 0.67 is as close as our table gets us. Since the distribution is symmetrical, 50% of the scores fall between z scores of ±0.67. Now convert the z scores to raw scores using Formula 5.4a

Upper Cutoff	Lower Cutoff
$X = 80 + (0.67)(5)$	$X = 80 - (0.67)(5)$
$= 80 + 3.35$	$= 80 - 3.35$
$= 83.35$	$= 76.65$

The interquartile range is $83.35 - 76.65 =$ **6.70**. ■

One issue needs to be addressed before we continue. Up to this point, we have been claiming to identify the portion of the curve *above* or *below* given values in a data set; however it is unclear how to classify scores that correspond

perfectly with the *z* score in question. In other words, if we are asked to find the percent of scores in a data set below a *z* of –1.96, column *C* in the *z* table would direct us to answer 2.5%; however what about a raw score that corresponds perfectly with a *z* of –1.96? Is that raw score part of the lowest most 2.5% of the data set or part of the upper 97.5%? Where do we put it? This is not an easy question to answer. It depends, in part, on whether the raw scores are understood to be from a discrete or a continuous measure. The conventional way to handle this situation is to suggest that the point itself – in this case, the raw score corresponding perfectly with a *z* value of –1.96 – should be included in the area of the curve being identified. In this way, it is appropriate to say that 2.5% of the scores in the distribution have a *z* score of –1.96 *or* below (not just below –1.96). Up until now, we have only been using language of being *above* or *below* a given point, but it is also appropriate to use the phrases "at or above" and "at or below."

The *z* score system is not the only method of standardizing scores. Another standardizing method is the *T* score system. It is very similar to the *z* score system. However, it features a mean of 50, a standard deviation of 10, and does not have any negative values. We will not cover the *T* distribution since use of it is largely restricted to a handful of specific psychometric measurements (e.g. the Minnesota Multiphasic Personality Inventory [MMPI] uses *T* scores). Information about this standardizing system is likely to be presented when students are studying these particular psychometric measurement systems in other content-related classes.

Summary

Percentiles and *z* scores are statistical transformations of original scores. They provide information about where a score stands in relation to other scores in a given distribution. The percentile rank of a score is expressed as the percentage of scores in the distribution that fall below that score. The percentile rank of a score is based on the rank order of scores; it does not take the distance between scores into consideration. The *z* score transformation avoids this problem by using the variability of the distribution in the transformation formula. A *z* score is the number of standard deviations a raw score is from the mean of the distribution. All *z* scores above the mean are positive; those below the mean are negative. The *z* score distribution has a mean of 0 and a standard deviation of 1. Formulas can be used to transform raw scores into *z* scores and conversely to find the raw score values that correspond with specific *z* scores. A *z* table can be used to find probabilities associated with selecting scores at random above and below *z* scores. However, a *z* distribution will only be normal if the raw score distribution is normal.

Using Microsoft® Excel and SPSS® to Find *z* Scores

Excel

General instructions for data entry into Excel can be found in Appendix C.

Data Analysis

1) Input a data set. (For practice, we can use one of the data sets in the "Work Problems for the Computer" for this chapter.)
2) Since Excel first needs to know the mean and standard deviation, we will need to find that first. So, select **Data Analysis** and then **Descriptive Statistics**. Click **OK**.
3) Highlight all of the scores in the distribution and put those quadrant numbers into the **Input Range** box.
4) Select a location for the output. Use the **Output Range** box if needed.
5) Make sure to click **Summary Statistics** before clicking **OK**. This should generate a box of descriptive statistics, including the mean and standard deviation of the distribution (the first and fifth statistic generated, respectively).
6) Now select the location for the *z* score computation for a given raw score number we wish to transform (typically this is the cell directly adjacent to the right of the raw score), and click the ***fx*** key to the immediate left of the input box at the top of the spreadsheet.
7) Excel uses the term **standardize** for *z* scores. Search for and select this term.
8) In the **X** box, select the raw score we wish to transform.
9) In the **Mean** and **Standard-dev** box, select the appropriate values from the Summary Statistics box we just constructed and click **OK**. The *z* score should show up in the selected box.
10) To transform all of the raw scores into *z* scores, we can use the "autofill" function. However, first highlight the *z* score already transformed and place a $ in front of the number of the coordinate for both the mean and standard deviation components of the equation (since we do not want those values to change). For instance, if our mean value is found in cell E5, change that to E$5; if our standard deviation value is found in cell E9, change that to E$9. This will keep this value constant for all of the autofill calculations. Then highlight the cell containing the *z* score and move the cursor to the lower right of the cell until a "+" appears. Then drag down to a cell corresponding to the last raw score cell and release. This should produce a column of *z* scores corresponding to the adjacent raw scores.

SPSS

General instructions for inputting data into SPSS can be found in Appendix C.

Data Analysis

1) Input a data set. (For practice, we can use one of the data sets in the "Work Problems for the Computer" for this chapter.)
2) Once the data has been entered, click **Analyze** on the tool bar, select **Descriptive Statistics** and then **Descriptives**.
3) Move the column label containing the data we wish to convert into *z* scores from the left box to the **Variable** box. Also, make sure the **Save standardized values as variables** box in the lower left corner is checked. This is vitally important.
4) Click **OK** (no other work is needed).
5) The output screen will generate some descriptive statistics. However, once we go back to the data file, we will see a new variable (named identical to the selected variable with a "Z" in front of it) with each raw score transformed into its corresponding *z* score.

Key Formulas

Formula for finding the Percentile Rank of *X*, *PR*

$$PR\ of\ X = \left(\frac{B + 1/2E}{N}\right) \cdot 100 \quad \text{(Formula 5.1)}$$

Formula for finding X given a Percentile Rank, X_p

$$X_p = L + \left(\frac{(N)(P) - F}{f}\right) \cdot h \quad \text{(Formula 5.2)}$$

Formulas for transforming an *X* Score into a *z* Score

Population

$$z = \frac{X - \mu}{\sigma} \quad \text{(Formula 5.3a)}$$

Sample

$$z = \frac{X - M}{s} \quad \text{(Formula 5.3b)}$$

Formulas for transforming *z* to an *X* Score

Population

$$X = \mu + z\sigma \quad \text{(Formula 5.4a)}$$

Sample

$$X = M + zs \quad \text{(Formula 5.4b)}$$

Key Terms

Percentile Rank	**Standard Score**
***z* Score**	**Standard Normal Distribution**

Questions and Exercises

1 Using the following frequency distribution, what is the percentile rank of a score of:
a 56
b 60
c 54
d 49

X	*f*
62	3
60	4
58	7
56	12
54	10
49	7
44	6

2 What is a *z* score, conceptually?

3 Why are some *z* scores positive values, while others are negative?

4 Think of two examples of variables that are believed to be normally distributed across a population. Defend the answers.

5 Think of two examples of variables that are not believed to be normally distributed across a population. Defend the answers.

6 Transform these scores of a population distribution into *z* scores.

Raw Scores: 4, 5, 7, 9, 10, 11

7 Given $M = 14$ and $s^2 = 16$, what is the *z* score of a raw score of 11?

8 In a distribution where $M = 25$ and $s = 3$, what raw score corresponds to a *z* score of 0.36?

Assume normality for all remaining questions.

9 If a distribution has a mean of 130 and a standard deviation of 13, what is the probability of randomly selecting a score above 140?

10 When $M = 34$ and $s = 3$, what percentage of scores are lower than 27?

11 What is the total percentage of scores that lie beyond z scores of ±1.96?

12 What percentage of scores fall between the z scores ±1.28?

13 What is the z value when the probability of selecting a score at random is:

- **a** At or below $z = 0.4207$
- **b** At or below $z = 0.3821$
- **c** At or above $z = 0.3192$
- **d** At or above $z = 0.0694$
- **e** At or below $z = 0.1151$
- **f** At or above $z = 0.2946$
- **g** At or above $z = 0.4641$
- **h** At or below $z = 0.4247$
- **i** At or above $z = 0.2119$

14 What is the probability of randomly drawing a score between the z scores +0.56 and −1.2?

15 In a distribution with $M = 78$ and $s = 7$, what is the probability of selecting a score between 72 and 80?

16 In a distribution having a mean of 123 and a variance of 49, what is the total percentage of scores falling above 130 and below 116?

17 If a standardized anxiety questionnaire has a mean of 25 and a standard deviation of 5, what is the probability that an individual selected at random will score between 20 and 30?

18 A standardized test of reasoning ability has a mean of 70 and a standard deviation of 7. The principal of a school would like to identify the best and worst students, as defined by their scores on the test. The best students are those with a percentile rank of 90 and above, and the worst students are those with a percentile rank of 10 and below. What are the raw score cutoffs the principal should use to identify the two groups of students?

19 Transform the following population of raw scores into z scores.

2, 4, 5, 6, 8, 9

20 For a distribution with $M = 48$ and $s = 4$, what is the percentile rank of:
- **a** 43
- **b** 57
- **c** 48
- **d** 50
- **e** 47

21 A 100-point final exam is administered in a class where $\mu = 78$ and $\sigma = 7$. What score did these four students receive?
- **a** Laurie, with a percentile rank of 95%.
- **b** Jennifer, if she is in the 80th percentile.
- **c** Jim, who scored better than 30% of the other students.
- **d** Gus, with a percentile rank of 45%.

22 For a distribution with $M = 35$ and $s = 3$, find the percentage of scores that are:
- **a** At or above $z = +1.20$
- **b** At or above $z = -0.36$
- **c** At or below $z = -0.56$
- **d** At or below $z = -0.79$
- **e** At or below $z = -1.10$
- **f** At or below $z = +0.98$
- **g** At or above $z = +0.13$

23 Professor Seitz gives a final exam to his abnormal psychology class and finds that $\mu = 56$ and $\sigma = 5$.
- **a** If the passing score is 38, what percentage of students will fail?
- **b** If Professor Seitz wants the "C" category to span the middle 30% of the distribution, what would be the cutoffs?
- **c** What score would serve as the cutoff for an "A" if only the top 10% of the class is to receive an "A?"

24 A student receives a score that corresponds to a percentile rank of 80%.
- **a** What z score corresponds to this rank?
- **b** Given the information available here, can we determine the raw score?

25 A score from a population that is 10 points below the mean corresponds to a z score of −2.50. What is the population standard deviation?

26 A sample score that is 5 points above the mean corresponds to a z score of 2.00. What is the sample standard deviation?

27 For a population with a standard deviation of 15, a raw score of 51 corresponds to a z of −1.00. What is the population mean?

28 For a sample with a standard deviation of 5, a raw score of 31 corresponds to a z of 2.00. What is the sample mean?

29 For a population with a mean of 60, a raw score of 61 corresponds to a z of 0.20. What is the population standard deviation?

30 For a sample with a mean of 75, a raw score of 60 corresponds to a z of −2.00. What is the sample standard deviation?

31 For a given sample distribution, a raw score of 35 corresponds to a z of −1.00 and a raw score of 40 corresponds to a z of −0.50. Find the mean and standard deviation for this sample.

32 For a given population, a raw score of 72 corresponds to a z of 0.20 and a raw score of 84 corresponds to a z of 0.80. Find the mean and standard deviation of the population.

33 For a given sample distribution, a raw score of 16 corresponds to a z of −2.00 and a raw score of 23.5 corresponds to a z of 3.00. Find the mean and standard deviation of the sample.

34 For a given population, a raw score of 77 corresponds to a z of 2.50 and a raw score of 41 corresponds to a z score of −5.00. Find the mean and standard deviation of the population.

35 Suppose miles traveled per year by American drivers is normally distributed with a mean of 25 000 miles and a standard deviation of 6 000 miles. If we wanted to find the miles traveled that will cut the distribution into five equally populated segments, what are the miles traveled that define the bottom 20%, the next 20%, and so on up to the top 20%?

36 Suppose the average American household generates 45 lb of garbage per week with a standard deviation of 11 lb. Suppose the local government wants to levy a tax on the worst offenders (top 15%) and offer a tax rebate as incentive for its most conscientious citizens (bottom 28%). What weekly garbage amounts correspond to these cutoffs?

37 For a sample distribution with a mean of 25 and a standard deviation of 4, what raw score corresponds to a z score of −1.75, and what percent of sample scores will be greater than that raw score?

38 For a population of scores with a mean of 99 and a standard deviation of 9, what raw score corresponds to a z score of 1.33, and what percent of scores will be greater than that raw score?

39 For a sample distribution with a mean of 150 and standard deviation of 15, what percent of scores will fall between the values of 170 and 175?

40 For a population of scores with a mean of 1 and a standard deviation of 0.15, what percent of scores will fall between the values of 0.6 and 0.7?

41 Suppose Andrew and Lisa wanted to compare how well they each performed in their respective soccer games. They are, however, on different teams. Since both are defenders and neither score very much, they decided to compare the number of completed passes. Andrew completed 54 passes; his team average was 44 completed passes per player with a standard deviation of 6. Lisa completed 48 passes; her team average was 38 passes with a standard deviation of 7. Which one performed better relative to their teammates?

42 Suppose both Sarah and Justine think the other wastes too much time. Sarah feels Justine spends too much time on social media compared with others, while Justine feels Sarah spends too much time figuring out what to wear compared with others. Suppose further we know that people average 65 minutes per day on social media (σ = 20 minutes) and 15 minutes per day deciding what to wear (σ = 4 minutes). Justine spends 90 minutes a day on social media and Sarah spends 20 minutes deciding what to wear. Who wastes more time compared with the rest of the population?

43 Using the following grouped frequency distribution, what is the percentile rank of a score of 172?

Class interval	Frequency
180–184	7
175–179	11
170–174	16
165–169	15
160–164	11
155–159	9
150–154	7

44 For the following grouped frequency distribution, find the percentile rank for an X of 40, 50, 65, and 90.

Class interval	Frequency
26–29	17
22–25	13
18–21	23
14–17	27
10–13	25
6–9	12
2–5	9

Computer Work

45 For the following population of scores:

12	15	34	23	32	12	22	21	19	25	14	11	12
11	10	14	15	13	12	16	18	21	29	32	31	30
24	30	29	28	26	21	19	17	16	15	11	10	17
32	30	29	29	28	27	21	14	21	18	16	16	11
20	23	14	15	17	11	21	32	20	20	25	15	17
14	15	23	26	30	24	19	23	22	21	24	17	15

Find μ, σ^2, σ and convert all raw scores into z scores.

46 For the following sample of scores:

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

Find M, s^2, s and convert all raw scores into z scores.

47 For the following sample of scores:

112	175	344	123	327	412	122	217	419	125	147	411	112
112	108	145	125	183	152	126	188	251	229	382	351	320
243	309	629	283	269	216	193	197	166	153	119	106	173
324	130	279	429	128	277	421	114	217	184	116	167	411
520	223	148	155	127	181	251	322	280	250	225	185	157
164	153	239	266	303	294	196	233	229	621	243	197	156

Find M, s^2, s and convert all raw scores into z scores.

48 For the following population of scores:

4.2	8.5	3.4	2.3	3.2	2.2	2.2	2.8	6.9	2.5	1.4	9.1	2.2
5.6	9.1	2.4	4.5	6.3	1.2	3.6	9.8	2.5	2.9	3.2	3.8	3.1
2.4	3.0	2.9	2.8	2.6	2.9	4.9	1.7	4.6	7.5	1.2	7.0	3.7
3.2	3.1	2.9	2.9	2.8	2.7	2.5	2.4	2.3	8.8	3.6	6.6	4.5
2.0	2.3	3.4	5.5	7.7	8.1	2.6	3.2	2.0	2.1	2.5	5.5	6.7
7.4	1.5	2.3	2.6	3.1	2.4	7.9	2.3	2.2	2.9	2.4	4.7	7.5

Find μ, σ^2, σ and convert all raw scores into z scores.

Part 3

Inferential Statistics

Theoretical Basis

6

Basic Concepts of Probability

6.1 Theoretical Support for Inferential Statistics

The concepts presented in this chapter and Chapter 7 build on what we have already learned regarding descriptive statistics and will, in conjunction, present the theoretical support for performing inferential statistics – the subject matter of the remainder of the text. **Inferential statistics** are a collection of mathematical techniques that use probability theory and hypothesis testing logic to draw inferences about the characteristics of a population of scores from the characteristics of a sample of scores. The primary types of inference to be made concern whether or not a sample seems to come from a known population *or* whether or not two or more samples seem to come from the same population. Operationally, inferential statistics are used to analyze the merits of different hypotheses that have been presented and for which quantitative information has been carefully and systematically gathered.

This chapter will introduce us to the basics of probability theory – its terminology, formulaic principles, and conceptual limitations. Chapter 7 will introduce us to concepts and terminology related to hypothesis testing as well as a description of the sampling distribution concept and an explanation of its theoretical importance. Together, this chapter and Chapter 7 help us understand the relationship between the features of randomly drawn samples and the features of populations from which they are drawn.

Probability theory can help us speculate about the nature of samples that are drawn from well-defined populations. For instance, imagine we are blindly selecting 10 jelly beans from a jar that contains an equal number of red and black jelly beans. We can imagine this jar to be a population of jelly beans and the selection of 10 to be a sample. Although we will not be able to predict the exact makeup of our sample, we might suggest that it should be close to about 5 of each color. Although the exact composition of our sample cannot be perfectly predicted, we can gather some reasonable expectations based on our knowledge of the population from which it is drawn. Now, what if there

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

was another jar of jelly beans from which we blindly selected 10 jelly beans, but this one contained 90% red and only 10% black jelly beans. Just as before, we would not be able to predict how many red and black jelly beans we would have in our sample, but because the population is well defined, we could use our knowledge of the population to make some predictions. One, for instance, might be that we will almost certainly have more red than black jelly beans in our sample. By knowing the composition of the population, we can gain a sense of how likely it will be to get different kinds of samples. Probability theory helps us to better understand what to expect when selecting samples from well-defined populations.

Inferential statistics reverses this logical flow of inference. Instead of inferring the features of a sample that is drawn from a well-defined population, inferential statistics starts with access to a well-defined sample, and by using probability theory as well as various hypothesis testing concepts, we are enabled to draw inferences about the features of the population from which the sample came. For instance, if we looked at our sample of 10 jelly beans and saw that they were all red, this information should be helpful to us as we think about the features of the jar of jelly beans from which this sample came. Furthermore, if we knew that there were two jars of jelly beans as described earlier, we might feel pretty strongly about from which jar our sample came. If, however, our sample contained seven red and three black jelly beans, we may not feel very certain at all. Additional ideas and terminology concerning how to properly test hypotheses will need to be introduced to help us set up ways of making decisions.

This chapter and Chapter 7 contain numerous concepts that are theoretically important. Although full comprehension of these theoretical concepts is not absolutely necessary to be able to "run" inferential statistical tests, a clear understanding of these concepts is necessary for a deep appreciation of how the inferential tests themselves work, what the recommended conclusions of the test mean (and do not mean), and why. For these reasons, we need to seriously grapple with the concepts and arguments presented in this chapter and Chapter 7 and return to them as needed as we make our way through the rest of the material in the textbook.

The chapters following 6 and 7 can be roughly clustered into four different cohesive groups of analysis: (i) Part 4 – z tests and t tests, (ii) Part 5 – analysis of variance tests, (iii) Part 6 – correlation and regression analyses, and (iv) Part 7 – nonparametric tests. In terms of theoretical rationale, all of them are anchored in the twin pillars of probability theory and hypothesis testing. In this sense, the order of presentation for these groups of tests does not matter. In fact, many textbooks have a different organization scheme. The specific order of presentation chosen for this text is based in pedagogical reasons. In this sense, once we feel comfortable with the material contained in this chapter and

Chapter 7, we should feel some degree of comfort moving directly to any subsequent chapter matching our interest, if so desired. Furthermore, if we find ourselves struggling to understand either the theoretical rationale underlying an inferential procedure or the conclusions we can or cannot draw from the results of an inferential test, review of the material in this chapter and Chapter 7 is recommended.

6.2 The Taming of Chance

Historically the process of decision making in situations of uncertainty was often dealt with by techniques designed to let the fates or the gods control the outcomes; for instance, procedures like the casting or drawing of lots is often referenced in ancient writings. With the rise of Christianity in the West, this practice was largely replaced through the direct petitioning to the Christian God for guidance in the making of weighty decisions. However, starting in the seventeenth century, several natural philosophers and mathematicians started to speculate about how to deal with more mundane uncertainties associated with daily life (e.g. Reeves, 2015). One such activity that spurred a good deal of interest concerned gambling. Many historians of science point to an exchange of letters between Blaise Pascal and Pierre De Fermat regarding how to settle out the stakes involved in an interrupted gambling venture that could not be finished as a critical event in the development of probability theory (e.g. Katz, 1993; Weisberg, 2014). We can easily imagine a gambling situation where one contestant would be ahead and would be more likely to win if the contest was to continue, but that this outcome would not be certain. Splitting the winnings in half would seem unfair to the person who is ahead. However, giving this person the full amount would seem unfair to the other contestant who, although sitting in a disadvantaged position, had not yet lost and had some hope of recovering to win. The discussion started by Pascal and Fermat generated a lot of further discussion over the next 300 years regarding how to quantify the uncertainty in situations such as these. (See also Box 2.2 presented earlier in this text.)

A major breakthrough came when it was realized that "games of chance" were probably not that at all but that the outcomes of rolls of the dice and the location of a card in a deck when shuffled were regulated and law-like, if only detailed information about preconditions and processes were available. However, given that such precise information could never be fully known, the key was to become *willfully ignorant* of the myriad of small effects that influence outcomes and choose to focus on the more general truths. (See Box 6.1 for further information about the use of willful ignorance in probabilistic thinking.) If we should choose to become ignorant of any particular roll of the dice and instead focus on what tends to happen across several events (e.g. several rolls

Box 6.1 Is the Scientific Method Broken? Uncertainty, Likelihood, and Clarity

An aspect of probabilistic thinking that seems to have been lost in most modern discussions of probability, and that may be partly responsible for the reproducibility problem in the social and medical sciences, is the realization that uncertainty is not merely the quantification of likelihood, but is also influenced by a clear understanding of the situation; let us use the term "clarity" for a lack of a better one. Now "likelihood" (or "risk" as it is sometimes called) is usually understood as something that can be quantified numerically, like the odds of selecting a spade from a deck of cards. This predisposes that the conditions are well understood as well as the relative frequencies of favorable and unfavorable options to be known. Every judgment call dealing only with likelihood has some known degree of risk to be wrong. Modern probability theory, since the 1930s, has almost exclusively focused on only this aspect of uncertainty. However, uncertainty, which is what probability theory was designed to address, is not restricted to merely this more objective quantification. Clarity is also a necessary component. By clarity, it is meant the degree to which the situation one is speculating about is being properly conceptualized.

For instance, when we say that the likelihood of getting snake eyes when rolling two dice is 1/36, we are assuming that we have full clarity regarding the situation (e.g. what is known as "snake eyes," the features of the two dice, and so on). We realize that gambling on this outcome has a certain amount of risk associated with it (actually a lot of risk associated with it!), but this risk is at least quantifiable. However, if we think about it for a few minutes, it should be easy to see that we may be overlooking several aspects of the probabilistic situation out of convenience. For instance, what if each side of each die is not equally likely? What if there are actually 10 sides to each die? What if the numbers on the die do not start with 1? Now, this is a simple example where there is most likely a lot of initial clarity regarding the situation; or if not, issues revolving around the clarity of the situation could be easily resolved by simply looking at the dice to be rolled. However, in many research situations, the degree of clarity that a researcher has regarding the situation is not as easy as inspecting the dice. Further compounding the situation, a researcher's lack of clarity may be best understood as a qualitative variable as opposed to a quantitative one, and as a result, not something that could even be factored into modern probability theory. Acknowledging this limitation, however, might leave the researcher without a clear path forward in their efforts to investigate. So, oftentimes researchers, perhaps unknowingly, engage in what some theorists call "willful ignorance" regarding issues of clarity so as to focus only on the issue of likelihood when testing their hypotheses. By essentially ignoring issues of clarity and focusing exclusively on likelihood, modern probability theory has created a false sense of confidence in the stated outcomes of some statistical analyses. Some argue that our failure to address this fundamental issue has led to the challenging state in which many areas of scientific investigation currently find themselves (e.g. Byers, 2011; Weisberg, 2014).

To help illustrate this further, imagine a medical situation where a patient has arrived with a set of symptoms and an unknown diagnosis. There are two judgments that need to be made: First, what is the nature of the illness, and second, what form of treatment would be best? It might be helpful to consider

the first of these two judgements as being representative of the issues of clarity. The doctor, relying on clinical expertise, past experiences, logic, and judgment, moves to increase clarity by rendering a diagnosis. Yes, some probabilistic information may have been used in this analysis, but primarily this was not an exercise in probability. Now, the second question of treatment may be much more representative of the issues related to likelihood and probability. Based on the outcomes of previous studies employing different treatments, the doctor may choose to prescribe one form of treatment over another due to the likelihood of a favorable outcome derived from a quantification of relevant data.

Research situations oftentimes involve both forms of uncertainty: likelihood and clarity. Moreover, when research replication efforts take place, issues of clarity have oftentimes not properly been noted or documented by the original researchers and are not properly considered by the replicating researcher. This disconnect can lead to contradictory findings when replication takes place and thus great confusion as to what is actually going on. In the interest of accuracy and the public confidence in science, it would be wise for researchers to begin to consider seriously the issue of clarity when setting up their investigations, writing up their procedures, and drawing conclusions from their data.

of the die), we can construct a fairly accurate representation of the set of outcomes and their likelihood. For example, Figure 6.1 shows a representation of the set of outcomes associated with the rolling of two standard dice. Historians suggest this gained insight to be critical in the development of probability theory.

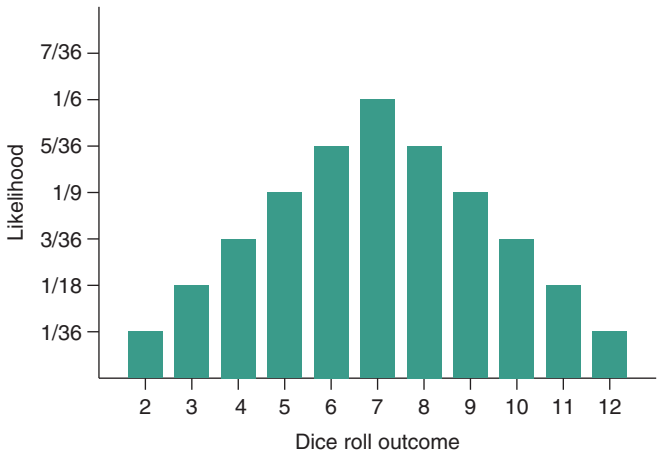


Figure 6.1 A probability distribution representing a single roll of a pair of six-sided dice.

6.3 What Is Probability?

Probability, as mathematicians use the term, can be defined as the likelihood of an event occurring. This likelihood can be represented by a number between 0 and 1. An event is anything that can either occur or not occur. For example, when we flip a coin, it either comes up heads or it does not come up heads (i.e. it comes up tails). So, in a coin flip, heads is an event. Of course, tails is also an event. When we roll a die, it either comes up 4 or it does not come up 4. Therefore, 4 is an event. Events that are single occurrences like a "4" on a die roll or a "head" on a coin flip are often called elementary events. An exhaustive set of elemental events is referred to as a set. Now back to events; an event can also be understood as a collection of elemental outcomes; we call these complex events. For example, in rolling the die, getting an even number is an event because we will either get an even number or not. Getting a number less than 4 is also an event. Drawing the King of Spades from a deck of cards would be an elementary event, but determining the probability of drawing a spade (of any value) would be a complex event.

When we assign a probability (i.e. a number between 0 and 1) to an event, we are stating how likely that event is to occur. If the probability is .5, it means that the event is as likely to occur as it is to not occur.[1] A probability greater than .5 means that it is more likely to occur than not occur, and a probability of less than .5 means that the event is more likely to not occur than to occur. Events with a probability of 0 mean they cannot occur,[2] while events with a probability of 1 mean that they most certainly will occur.

A good place to start to understand how probabilities are determined is to look at situations where each elemental event is equally likely – for example, a roll of a die, a drawing of a card, or a flip of a coin. In each of these scenarios, it makes sense to think that each elementary event is equally likely to occur. In situations like this, relative frequency is used to determine likelihood. By relative frequency we mean the comparison of the number of favorable events with the total number of possible events. Perhaps we can see that the likelihood of any elemental event is equal to the reciprocal of the total number of events in the set.

---

1 In this chapter and throughout the rest of the book, when we are dealing with probabilities, quantities that cannot exceed 1, we will typically not place a "0" in front of the decimal.

2 Theoretically, this description is a too simplistic. Events with the probability of virtually 0 happen all the time. For instance, imagine a six-figure number (e.g. 587 202); now imagine a two-digit number with two decimal places (e.g. 33.91). Now divide the six-figure number by the double-digit number with two decimal places. There are literally millions and millions of different answers that can be achieved as a result of this exercise. The answer found for the two numbers chosen as examples is 17 316.484 812 739 6. The chance of any person going through this simple exercise and getting that exact number is so close to zero as to be virtually zero – and yet it occurred. It may help to distinguish between events with probability equaling zero on the one hand and logically impossible events, like encountering a square circle, on the other. These are different concepts.

For example, the likelihood of rolling a die and getting a 4 is 1/6 or .17 (actually it is .166 repeating; but we will keep our answers to two decimals). There are six different events and one of them is a 4. This is also the frequency of getting a 3, by the way. In fact, each of the six elemental events has the same likelihood of occurring. This is not always the case, of course. Sometimes each elemental event is not equally likely. For instance, when we examine what will happen when a baseball player takes their turn batting, there are several different elemental events that could take place; for simplicity purposes we can think of the set of events as being (i) an out, (ii) a walk, (iii) a single, (iv) a double, (v) a triple, or (vi) a home run. Each of these six events, however, is not equally likely. It would be inappropriate for us to conclude that the likelihood of the batter getting a home run is one chance out of six. Likewise, it would be inappropriate for us to decide that the chance of rain today is .5 because after all there are only two choices – it will either rain or not rain. In these latter cases, while we may have correctly identified the set of elemental events, we have failed to realize that each event is not equally likely.

When we examine finding the probability of a complex event if all elemental events are equally likely, we encounter our first significant probability formula. First, however, we need to introduce some notation information as well as a new term. In probability formulas the uppercase letter P is used to represent "probability" and can be read as "The probability of…" Events are placed within parentheses. In this way, one can express the probability of flipping a coin and getting heads as $P(heads)$. In our coin flipping example the $P(heads) = .5$. Often mathematicians use an abstract symbol like a capital letter to represent an event. For the probability of event A, we could write $P(A)$. Finally, the word "favorable" is used to identify the event or events that are being considered. So, assuming all elemental events are equally likely, the formula for determining the likelihood of an elemental event occurring that is part of a favorable set of events is the ratio of favorable events compared to the set of all possible events.

Probability of favorable event

$$P = \frac{(number\ of\ favorable\ events)}{(total\ number\ of\ events)} \quad \text{(Formula 6.1)}$$

when all elemental events are equally likely.

If, for instance, we wanted to know the likelihood of getting a spade with the draw of a single card from a deck of 52, since all elemental events are equally likely, the probability would be 13/52 (or .25 or 25%; all of these ways of representing this relationship are acceptable). If the two Joker cards were also included in the deck, the ratio would be 13 favorable events over a total of 54 possible events (13/54, or .24).

Using the notion of relative frequency, we can say that a probability of .7 means that the event in question has a 7 out of 10 chance of occurring. Likewise, a probability of .25 would be a 1 out of 4 chance of occurrence. We should be aware that the analogy breaks down if we try to apply this way of literal thinking to all things. For example, we can say that there is a .7 probability of measurable rain tomorrow. We should note, however, that no 10 elementary events of which 7 are favorable exist. We cannot think of this situation in relative frequency terms. Yes, it can be argued that "something like" on 7 out of 10 days with meteorological features like tomorrow measurable rain will be recorded. But what does that really mean? Not much! Do we really have access to 10 days like tomorrow? No, we do not. However, we can still think of the .7 probability as a 7 out of 10 probability for the purposes of understanding what the number means. In order to make use of probabilistic information, it is not required for the user to actually identify the elemental events as real entities.

6.4 Sampling with and Without Replacement

Sampling with replacement is a method of sampling wherein a member of a population is randomly selected and then returned to the population before the next member is selected. This is not a difficult concept. Suppose we want to know the probability of selecting a red card from a deck of playing cards. Since half of the cards are red and half of the cards are black, the probability of selecting a red card at random is .50. If we return the selected card to the deck, the probability of selecting another red card is still .50. However, suppose we do not return the card to the deck and ask, "Now what is the probability of selecting a red card at random?" Since there are now more black cards than red cards in the deck, the probability of selecting a red card is *not* .50 (it is a bit less). In the latter example, we have sampled without replacement. Sampling without replacement is a method of sampling in which a member of a population is not returned to the population before selecting another member of the population.

Sampling with or without replacement has obvious implications for the probability of occurrence of subsequent events. The distinction between sampling with replacement and sampling without replacement is also important to mathematical statisticians interested in hypothesis testing. For example, the sampling distribution concept, a centerpiece of Chapter 7, is derived from a sampling *with* replacement procedure. As we move forward in the chapter, we will be limiting ourselves to only sampling with replacement – the procedure that is most commensurate with an introduction to probability concepts.

6.5 A Priori and A Posteriori Approaches to Probability

Up to this point we have approached probability from an **a priori or classical approach**. This approach is based on a logical analysis of the probabilistic situation and the relative frequencies that are predicted. It is not based on the accumulation of data. Using logic alone we can claim that we should get 100 "4's" if we rolled a single die 600 times.

The **a posteriori approach**, on the other hand, is an empirical approach to probability. It requires the collection of data. Using the preceding example, how would we determine the probability of getting a 4? We would need to collect some data. We could, for instance, roll a die 600 times and record how many times we got a 4. Suppose we did just that and got a "4" 90 times. Note that this a posteriori method of determining the likelihood of rolling a "4" is .15, close, but not identical, to the .17 probability determined by the a priori method. Theoretically, it is believed that if we had rolled the die an infinite number of occurrences, this a posteriori approach would have yielded a probability of .17 (or, stated more precisely, a ratio of 1/6).

For the previous problem of determining the getting a "4" on a roll of a die, nothing is gained by using the a posteriori approach because reason alone (the a priori method) could solve the problem. However, in actual research, the a priori approach is often inappropriate because, without collecting data, we do not know the number of favorable events in a population (recall the concept of "clarity" presented in Box 6.1). For example, suppose we would like to know the probability that those who are hospitalized have a diagnosis of depression. There is no way to use logic alone to determine the answer. We would have to take a random sample (preferably a large one) of all hospitalized patients and observe the proportion of patients in our sample who received a diagnosis of depression. We could then make a statement about the probability of those that are hospitalized having a diagnosis of depression.

Since the purpose of this chapter is to provide a brief introduction to some of the basic concepts of probability, considering only the a priori approach will suffice for this chapter. The a posteriori approach will be further explored, implicitly, when hypothesis testing is introduced in Chapter 7.

6.6 The Addition Rule

The **addition rule** is used to determine the probability of occurrence of one of many possible events. It is typically applied when the question has the word "or" in it. For example, "What is the probability of rolling a die and obtaining a 4 *or* a 6?; What is the probability of drawing a club *or* a heart from a deck of cards?" We can represent this concept by using the term $P(A \text{ or } B)$. [Note that in mathematics, this is usually written as $P(A \cup B)$, but we will just use the word "or" rather

than introduce a new symbol.] Determining the proper formula to use when answering an "or" question depends upon another concept – mutual exclusivity. **Mutually exclusive events** occur when one event precludes the occurrence of another event. (The term "disjoint" may also be used to represent this concept.) For example, when we roll a die once, it is impossible to obtain a 4 and a 6; one precludes the other. When we select a card from the deck, it cannot have more than one suit (we can either get a spade, heart, club, or diamond; the suits cannot co-occur within the same elemental event). If the two events are mutually exclusive, the formula for the addition rule follows.

Addition rule formula for two mutually exclusive events

$P(A \text{ or } B) = P(A) + P(B)$ (Formula 6.2)

Formula 6.2 is read as "The probability of either event A or event B occurring equals the probability of event A occurring *plus* the probability of event B occurring."

■ **Question** *Assume that we roll one die. What is the probability of coming up with a 2 or a 5?*

Solution

Step 1. First, determine $P(A)$. Let us call rolling a 2 event A. Is each event equally likely? Assuming the die is a fair die, we can assume this. So, we will use Formula 6.1 to answer this question.

$$P(A) = \frac{(number\ of\ favorable\ events)}{(total\ number\ of\ events)} = \frac{1}{6} = .17$$

The probability of rolling a 2 is .17. (This value is more accurately stated as the fraction 1/6.)

Step 2. Determine the probability of event B occurring. Let us call rolling a 5 event B. Since each event is equally likely, let us use Formula 6.1 once again.

$$P(B) = \frac{(number\ of\ favorable\ events)}{(total\ number\ of\ events)} = \frac{1}{6} = .17$$

The probability of rolling a 5 is .17. (This value is more accurately stated as the fraction 1/6.)

Step 3. Since events A and B are mutually exclusive, we can use Formula 6.2 to determine the probability of rolling a 2 or a 5.

$$P(A \text{ or } B) = P(A) + P(B)$$

$$P(A \text{ or } B) = 1/6 + 1/6 = \mathbf{2/6} \text{ or } \mathbf{.33}$$

The probability of rolling a 2 or a 5 on a single toss of the die is .33. ■

The addition rule for mutually exclusive events is generalizable to situations where we want to determine the probability of occurrence of one of several events. The general equation for the addition rule with more than two mutually exclusive events follows.

Addition rule formula for more than two mutually exclusive events

$$P(A \text{ or } B \text{ or } C \text{ or } \ldots Z) = P(A) + P(B) + P(C) + \cdots + P(Z) \quad \text{(Formula 6.3)}$$

where

$P(Z)$ = the probability of occurrence of the last event

Formula 6.3 is merely an extension of the addition rule formula for two mutually exclusive events. As a consequence, the computational steps follow the same format as outlined in the preceding worked example. Below are a couple worked examples.

■ **Question** *What is the probability of randomly selecting a 3, 7, or 9 from a deck of cards?*

Solution

Step 1. First determine the $P(A)$. Let us call drawing a 3 event A. Since each elemental event is equally likely, we can use Formula 6.1.

$$P(A) = \frac{(number\ of\ favorable\ events)}{(total\ number\ of\ events)} = \frac{4}{52} = .0769$$

The probability of randomly selecting a 3 from the deck is .0769. (We should feel free to use a couple more decimal places when dealing with probabilities, since the values only range from 0 to 1.)

Step 2. Determine the probability of event B occurring. Let us call drawing a 7 event B. Since each elemental event is equally likely, once again we can use Formula 6.1.

$$P(B) = \frac{(number\ of\ favorable\ events)}{(total\ number\ of\ events)} = \frac{4}{52} = .0769$$

The probability of randomly selecting a 7 from the deck is also .0769.

Step 3. Determine the probability of event C occurring. Let us call drawing a 9 event C. Since each elemental event is equally likely, once again we can use Formula 6.1.

$$P(C) = \frac{(number\ of\ favorable\ events)}{(total\ number\ of\ events)} = \frac{4}{52} = .0769$$

The probability of randomly selecting a 9 from the deck is also .0769.

Step 4. Use the addition rule to sum the separate probabilities.

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$$

$$P(A \text{ or } B \text{ or } C) = .769 + .769 + .769 = .23$$

This means the probability of selecting a 3, 7, or 9, on a single draw, is .23. ■

Let us try one more example. This time the problem will involve 4 elemental events, and we will take some liberties with the process.

■ **Question** *Assume that we roll one die. What is the probability of coming up with a number less than 5?*

Solution

Step 1. Since all elemental events can be considered equally likely, and since each elemental event is mutually exclusive, we can use Formula 6.3.

$$P(1 \text{ or } 2 \text{ or } 3 \text{ or } 4) = P(1) + P(2) + P(3) + P(4) = (1/6) + (1/6) + (1/6) + (1/6) = 4/6 \text{ or } .67$$

The probability of rolling a number less than 5 on a single toss of the die is .67. ■

We must watch out since not all cases of $P(A \text{ or } B)$ can be worked that easily; sometimes events A and B can co-occur. For example, what is the probability of drawing a heart or a queen? Well, the probability of drawing a heart is 13/52 (or 1/4), and the probability of drawing a queen is 4/52 (or 1/13). But the probability of drawing a heart or a queen is 16/52; this is not the sum of 13/52 and 4/52 (which is 17/52). What is the difference? In the first example, the two events (jack and queen) cannot both occur at the same time. However, in the second example, the two events (heart and queen) can co-occur; one can draw both a queen and a heart at the same time. In card language, that card is the queen of hearts; and in mathematical language, that card represents a co-occurrence. An event being a "heart" is not mutually exclusive from an event being a "queen." If we leave the formula as is, the queen of hearts card will be counted twice, once when we tally up all of the queens and again when we tally up all of the hearts. We need to change our formula to keep from double-counting events that qualify as being both an event A and event B. In so doing we will need to introduce a new concept, the probability of both events A *and* B co-occurring. (This idea will be more fully explained in the following sections of the chapter.) Here is the formula for finding the likelihood of events A or B occurring if they are not mutually exclusive.

Addition rule formula for two events

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \quad \text{(Formula 6.4)}$$

$P(A \text{ and } B)$ can be read as "The probability of both event A and event B co-occurring." In the above example of selecting a queen or a heart, the probability of both a queen and a heart co-occurring is 1/52. This helps us understand why the correct answer is not 17/52, but rather 16/52. Perhaps it is helpful to point out that Formula 6.2 is just a special case of the more general Formula 6.4. Formula 6.4

always works; but Formula 6.2 works when $P(A \text{ and } B) = 0$. In these situations the last expression in Formula 6.4 simply falls out of the formula, leaving us with Formula 6.2.

■ **Question** *Suppose we are about to win a game if we roll either an even number or a number greater than 4. What are our chances of winning?*

Solution

Step 1. First determine the $P(A)$. Let us call rolling an even number event A. Let us further assume that we know that the chance of rolling an even number is .5.

Step 2. Second determine the $P(B)$. Let us call rolling a number greater than 4 event B. Let us further assume that we know that the chance of rolling a number greater than 4 is .3333.

Step 3. Finally determine $P(A \textit{ and } B)$. This, when translated into our problem, means the chance of rolling a number that is both an even number and a number greater than 4. Let us assume we know that the chance of this occurring is .1667. (The number 6 is the only value on a standard die that is both an even number *and* a number that is greater than 4. Since all elemental events are equally likely, the chance of rolling a 6 is 1/6 or .1667.)

$$P(A \textit{ or } B) = P(A) + P(B) - P(A \textit{ and } B)$$

$$P(A \textit{ or } B) = .5 + .3333 - .1667 = \mathbf{.6667} \textit{ or } \mathbf{2/3}$$

The probability of rolling a die and getting either an even number or a number greater than 4 is .6667. We have a 2/3rd chance of winning on the next roll. ■

The addition rule for more than two events when the events are not mutually exclusive is much more complicated. Since this chapter is merely an introduction to probability, this topic will not be covered.

6.7 The Multiplication Rule

The addition rule is used when we want to determine the probability of one of two or more events occurring. The **multiplication rule** is used when a problem is framed as the probability of event A *and* event B occurring or $P(A \text{ and } B)$. [In mathematics, the $P(A \textit{ and } B)$ is usually written $P(A \cap B)$, but for the same reason as above, we will use the word "and."] Of course, the multiplication rule can be extended to problems that address more than two events co-occurring, just as the addition rule can be used when there are more than two events being considered. However, since this chapter is merely an introduction to probability, we will only consider the co-occurrence of two events. We will, however, consider the multiplication rule under two different conditions: when events are independent of each other and when they are dependent.

Probabilistic independence between events is found when knowledge of the occurrence of one event has no effect on determining the probability of occurrence of a second event. For instance, if we wanted to determine the likelihood of selecting a card from a deck of cards that is both a spade and a face card, the occurrence of a spade does not change the likelihood that the card will be a face card. If we did not know the card selected is a spade, the probability of a face card (we will include the ace for the sake of argument) is 16/52 or 4/13. If, however, we know the card selected is a spade, the probability that the card is also a face card has not changed – it is still 4/13. In this case, we can say that event A (spade) and event B (face card) are independent of each other; the occurrence of one event did not change the likelihood of the other event occurring. (Even though we presented this relationship from the perspective of selecting a face card, the same relationship can be shown from the perspective of selecting a spade. Namely, there is a 1 in 4 chance of selecting a spade knowing nothing else. If we learn that we have a face card, the ratio stays the same – we still have a 1 in 4 chance of selecting a spade. If event A is independent of event B, then event B is independent of event A – this is a shared property.) The formula for the multiplication rule for independent events follows.

Multiplication rule for two independent events

$P(A \text{ and } B) = P(A)P(B)$ (Formula 6.5)

Formula 6.5 is read as "The probability of events A and B *co*-occurring is equal to the probability of event A occurring multiplied by the probability of event B occurring." Be sure to note that this formula is true only if the two events are independent. We will develop a more general rule for multiplication in the next section.

Let us take a look at an example. If we roll a die and flip a coin, certainly the outcome of each is independent of the outcome of the other. So, what is the probability of getting both a heads on the coin flip *and* a 5 on the roll of the die? Well, the $P(heads) = 1/2$ and $P(5) = 1/6$. So, using our rule, $P(heads \text{ and } 5) = P(heads)P(5) = (1/2)(1/6) = 1/12$. If we think about this, we can see that it is correct. The elementary events in this example are the combined events of the two actions – one from the coin flip and one from the roll of the die. There are 12 of them (e.g. *head and 1*, *head and 2*…, *tail and 1*, *tail and 2*…). Furthermore, all of the elementary events are equally likely, so the probability of any one of them (e.g. *head and 5*) is 1/12, just as we found using the rule.

Let us look at another example. Suppose we know that the probability of Lisa earning an "*A*" in a collegiate course is .9, and we know that the probability of Jason earning an "*A*" in the course is .4. Suppose we also know that these students have no interaction at all; in other words, their grades should be independent. So, the probability that both will get an A is $(.9)(.4) = .36$. As we would expect, it must be less than the probability of either one of them getting an "*A*."

■ **Question** *What is the probability of randomly selecting a 4 and an 8 on two successive draws from a deck of cards? Since sampling with replacement is used, one card is randomly drawn from the deck and then put back into the deck, and then a second card is randomly selected.*

Solution

Step 1. First determine the $P(A)$. Let us call drawing a 4 event A. Since all cards are equally likely to be selected, we can use Formula 6.1.

$$P(A) = \frac{(number\ of\ favorable\ events)}{(total\ number\ of\ events)} = \frac{4}{52} = .0769$$

The probability of randomly selecting a 4 is .0769 or about 7.7%.

Step 2. Determine $P(B)$. Let us call drawing an 8 event B. Note that $P(B)$ is unaffected by step 1. This is because the card drawn for step 1 has been replaced (the "sampling with replacement" procedure); therefore, the second draw is from a complete deck of cards.

$$P(B) = \frac{(number\ of\ favorable\ events)}{(total\ number\ of\ events)} = \frac{4}{52} = .0769$$

The probability of randomly selecting a 8 is .0769 or about 7.7%.

Step 3. The multiplication rule is now applied.

$$P(A\ and\ B) = P(A)P(B)$$

$$P(A\ and\ B) = (.0769)(.0769) = \mathbf{.0059}$$

The probability of drawing a 4 replacing the card and then drawing an 8 is .0059. Stated differently, only 59 times out of 10 000 would these two events be expected to occur in this sequence. ■

Formula 6.5 can be extended to any number of independent events. For example, if we wanted to know the probability of drawing an ace, a spade, and the 6 of clubs on three successive draws, we would merely multiply the three probabilities of occurrence.

Probabilistic dependence between events is found when knowledge of the occurrence of one event changes the determination of the probability of occurrence for the second event. For instance, if we wanted to determine the likelihood of selecting a person from a population that is both a biological female and under 6 ft tall, the occurrence of a biological female would change the likelihood that the person will also be under 6 ft tall. If we stipulate that about 8% of people are 6 ft tall or taller, we would have about a 92% chance of randomly selecting a person who was less than 6 ft tall. But if we stipulate that the person we have selected is a biological female, the likelihood of having a person who is under 6 ft tall has just jumped to about 99%. In this case, the occurrence of one event (female) changed the likelihood of the other event occurring. Yes, this is a

bit confusing because both events are technically occurring at the same time (at the point of selecting the individual), but the way to calculate the likelihood of this co-occurrence necessarily involves exploring the relationship between the two events. To determine the probability of two dependent events, we need to use the more general multiplication rule. (Why we use the more general multiplication rule for dependent events will be clarified at the end of the next section.)

Multiplication rule for two events

$P(A \text{ and } B) = P(A|B)P(B)$ (Formula 6.6)

The symbol $P(A|B)$ is read as "The probability of event A occurring given the occurrence of event B". [It does not mean that $P(A)$ is to be divided by $P(B)$.] The symbol $P(A|B)$ is also referred to as a conditional probability, a concept that will be further explored in the next section of this chapter. A worked problem clarifies the use of Formula 6.6.

■ **Question** *What is the probability of randomly selecting a person from the campus student population that is both a psychology major and a biological female?*

Given *Suppose it is known that 10% of the student population are psychology majors, 80% of the psychology majors are biological females, and 60% of the entire student population are biological females.*

Solution

Determine $P(A)$, $P(B)$, and $P(A|B)$. Let us call being a biological female $P(A)$. Let us call being a psychology major $P(B)$. This would mean that $P(A|B)$ would mean the probability of being a biological female given that one is a psychology major. (This is the only way to assign the events for this problem – we do not have enough data to find $P(A|B)$ if events A and B are switched.) Since we are given all of the needed values, there is no need for preliminary steps.

If being a psychology major and being a biological female were independent of each other (where the occurrence of one of these events did not change the rate of occurrence for the other), then we could use Formula 6.5 and determine that

$$P(A \text{ and } B) = P(A)P(B)$$
$$P(A \text{ and } B) = (.6)(.1) = .06$$

However, these events are not independent and so we must use Formula 6.6. In this case

$$P(A \text{ and } B) = P(A|B)P(B)$$
$$P(A \text{ and } B) = (.8)(.1) = .08$$

The probability of randomly selecting a student on campus who is both a psychology major and a biological female is 8%. Since there is a dependency between events A and B, we should *not* use Formula 6.5, and it would *not* be true that the probability of randomly selecting a student on campus who is both a psychology major and a biological female is 6%. ■

Formula 6.6 allows us to incorporate into the calculations the realization that most psychology majors are biological females. And so, of the population of students, the likelihood of event A and event B co-occurring in the same person is increased a bit by the fact that events A and B already seem to co-occur a bit more than one might expect if the two events were randomly represented in the population (or said in other words, if the two events were independent of each other). Of course in other situations, events A and B may be dependent, but in such a way that the occurrence of A *decreases* the likelihood of event B occurring. The key term regarding whether events are independent of each other or not is whether the likelihood of one event *changes* (increases *or* decreases) if the other event is known to have occurred.

6.8 Conditional Probabilities

As was stated earlier in the chapter, a **conditional probability** is an expression of likelihood given that another particular event has occurred. For instance, we can easily see that the probability of randomly selecting a person from a population who is pregnant will change if we know ahead of time if the person selected is a biological male or female. The $P(A|B)$ if event A equals "being pregnant" and B equals "being a biological female" is undoubtedly a low number, but $P(A|B)$ would be zero if event B was changed to "being a biological male." In a sense, we have been doing conditional probabilities all along. Oftentimes, the conditions surrounding a probability are simply assumed. For instance, what is the probability of rolling a die and having it land on a 4? The commonsense response is that the probability would be 1/6. But what if we were later told that the die we were rolling was 10-sided? Well, we might be a bit upset to have not been told that up front, but we would instinctively know that we need to change our answer. It helps to think of conditional probabilities as specifications of the situation. Sometimes the situation is so well known as to be assumed, but other times we need to be clear about the nature of the situation. By the way, oftentimes probability theorists will refer to gaining the specifics of the probabilistic situation as understanding the **sample space**. (For example, learning how many sides are on the die that is being rolled, learning how many times it will be rolled,

and so on.) Determining the probability of rolling a die and getting a 4 changes depending on the number of sides a die has because the sample space in which a 4 can be found changes.

The formula for determining conditional probabilities is simply a reworking of the multiplication rule for two dependent events. The formula is as follows.

Conditional probability formula

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \quad \text{(Formula 6.7)}$$

By simply dividing both sides of the formula for the multiplication rule for two dependent events by $P(B)$, we can transform Formula 6.6 into Formula 6.7.

Let us try our hand at determining conditionals by looking at a form of a problem that is often used when teaching probability – the ball-in-urn scenario. Imagine an urn (it is a bit of a mystery why probability people like to call baskets "urns," but far be it for us to change the status quo!) filled with balls. These balls have different colors and also have either an "*X*" or a "*Y*" on them. Let us further stipulate that each ball is equally likely to be drawn. Now suppose the urn we are working with has the following contents:

20	Red – *X*
20	Red – *Y*
10	Green – *X*
50	Green – *Y*

Conveniently, the balls add up to 100. On the basis of the given information, we should be able to see the following:

P(*Red*)	=.4 (Formula 6.1)
P(*Green*)	=.6 (Formula 6.1)
P(*X*)	=.3 (Formula 6.1)
P(*Y*)	=.7 (Formula 6.1)
P(*Red or Green*)	=.6 + .4 = 1 (Formula 6.2)
P(*X or Y*)	=.3 + .7 = 1 (Formula 6.2)
P(*Red or X*)	=.4 + .3 – .2 = .5 (Formula 6.4)
P(*Green or X*)	=.6 + .3 – .1 = .8 (Formula 6.4)
P(*Red or Y*)	=.4 + .7 – .2 = .9 (Formula 6.4)
P(*Green or Y*)	=.6 + .7 – .5 = .8 (Formula 6.4)

Furthermore, we should be able to see that the following are true:

P(*Red and X*)	=.2 (given)
P(*Red and Y*)	=.2 (given)
P(*Green and X*)	=.1 (given)
P(*Green and Y*)	=.5 (given)

Let us try to find some conditionals.

■ **Question** *Given the probabilistic situation above, what are the following conditional probabilities? P(Red|X), P(Red|Y), P(Green|X), and P(Green|Y)?*

Solution

For *P*(*Red*|*X*), first assign "Red" and "X" to events *A* and *B*. Let us define "Red" as event *A* and "*X*" as event *B*. Use Formula 6.7 to determine the correct value.

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(Red|X) = \frac{P(Red \text{ and } X)}{P(X)} = \frac{.2}{.3} = \mathbf{.67}$$

The probability of selecting a red ball given an "*X*" ball has been drawn is .67 or 67%.

For *P*(*Red*|*Y*), first assign "Red" and "*Y*" to events *A* and *B*. Let us define "Red" as event *A* and "*Y*" as event *B*. Use Formula 6.7 to determine the correct value.

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(Red|Y) = \frac{P(Red \text{ and } Y)}{P(Y)} = \frac{.2}{.7} = \mathbf{.29}$$

The probability of selecting a red ball given a "*Y*" ball has been drawn is .29 or 29%.

For *P*(*Green*|*X*), first assign "Green" and "*X*" to events *A* and *B*. Let us define "Green" as event *A* and "*X*" as event *B*. Use Formula 6.7 to determine the correct value.

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(Green|X) = \frac{P(Green \text{ and } X)}{P(X)} = \frac{.1}{.3} = \mathbf{.33}$$

The probability of selecting a green ball given an "*X*" ball has been drawn is .33 or 33%.

For *P*(*Green*|*Y*), first assign "Green" and "*Y*" to events *A* and *B*. Let us define "Green" as event *A* and "*Y*" as event *B*. Use Formula 6.7 to determine the correct value.

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(Green|Y) = \frac{P(Green \text{ and } Y)}{P(Y)} = \frac{.1}{.7} = .\mathbf{14}$$

The probability of selecting a green ball given a "*Y*" ball has been drawn is .14 or 14%. ■

Before we move on to the final section of this chapter, we need to go back and address the topic of independence again, but now from a more informed vantage point. When we introduced the concept of independence, we stated that two events are independent if the occurrence of one event does not change the likelihood of occurrence for the other event. Determining if two events are independent is often times not a straightforward decision. With a little reflection we may be able to see that selecting a spade and selecting a face card are independent events, but it may be much more difficult to determine merely from what is given in the problem above if a red ball is independent of a ball with an "*X*" on it. To help matters, it is important to realize that Formula 6.6 is the more generally true formula for determining multiplication problems in probability (the "and" problems). Formula 6.5 is merely a special case version of Formula 6.6. That special case is for situations where events *A* and *B* are independent. Look at the two formulas. Here we see that the only difference is that Formula 6.5 refers to *P*(*A*), while Formula 6.6 refers to *P*(*A*|*B*). Here we also see that if two events are independent, then the *P*(*A*) should equal the *P*(*A*|*B*). That is, if two events are independent, the likelihood that event *A* will occur is the same likelihood if we previously know that event *B* has or has not occurred. Putting this into the card drawing question, the probability of selecting a spade is 1 out of 4 (25%). This likelihood does not change if we are told ahead of time that the card is a face card; the probability is still 1 out of 4 (or 25%). Therefore, this means we can use the two formulas in combination if we need to determine if two events are independent of each other. We can simply run both formulas and see if they yield the same result. If the two formulas produce the same answer, then we will know that *P*(*A*) = *P*(*A*|*B*), and so events *A* and *B* must be independent of each other. If, on the other hand, the formulas yield different answers, then we know that

$P(A) \neq P(A|B)$, and therefore events A and B are dependent upon one another (and so the answer to Formula 6.6 is the correct one). In this way we can use these formulas in combination to help us figure out if two events are independent.

■ **Question** *Given the following probabilistic situation involving balls in an urn, are events Green and X independent? Are events Yellow and Y independent?*

15	Green – X
15	Green – Y
10	Red – X
20	Red – Y
25	Yellow – X
15	Yellow – Y

Solution

Step 1. To determine if events "Green" and "X" are independent, we need to first determine the values corresponding to Formulas 6.5 and 6.6. Let us assign "Green" to be event A and "X" to be event B. These are $P(Green\ and\ X)$, $P(Green)$, $P(X)$, $P(Green|X)$.

$P(Green\ and\ X)$	=.15 (given)
$P(Green)$	=.3 (Formula 6.1)
$P(X)$	=.5 (Formula 6.1)
$P(Green|X)$	=.15/.5 = .3 (Formula 6.7)

Step 2. Then we need to "run" both formulas. (Recall that we assigned "Green" to be event A and "X" to be event B.)

$$P(A\ and\ B) = P(A)P(B)$$

$$P(Green\ and\ X) = P(Green)P(X) = .3(.5) = .15 \qquad \text{Formula 6.5}$$

$$P(A\ and\ B) = P(A|B)P(B)$$

$$P(Green\ and\ X) = P(Green|X)P(X) = .3(.5) = .15 \qquad \text{Formula 6.6}$$

Both formulas yield the same answer (.15). This means "Green" and "X" are independent. The likelihood of a "Green" ball occurring is not influenced by the occurrence of a "X" ball (and vice versa).

Now let us apply this same procedure to the second question: Are "Yellow" and "Y" independent?

Step 1. To determine if events "Yellow" and "*Y*" are independent, we need to first determine the values corresponding to Formulas 6.5 and 6.6. Let us assign "Yellow" to be event *A* and "*Y*" to be event *B*. These are *P*(*Yellow and Y*), *P*(*Yellow*), *P*(*Y*), *P*(*Yellow*|*Y*).

P(*Yellow and Y*)	=.15 (given)
P(*Yellow*)	=.4 (Formula 6.1)
P(*Y*)	=.5 (Formula 6.1)
P(*Yellow*|*Y*)	=.15/.4 = .375 (Formula 6.7)

Step 2. Then we need to "run" both formulas. (Recall that we assigned "Yellow" to be event *A* and "*Y*" to be event *B*.)

$$P(A \text{ and } B) = P(A)P(B)$$

$$P(\text{Yellow and } Y) = P(\text{Yellow})P(Y) = .4(.5) = .2 \qquad \text{Formula 6.5}$$

$$P(A \text{ and } B) = P(A|B)P(B)$$

$$P(\text{Yellow and } Y) = P(\text{Yellow}|Y)P(X) = .375(.5) = .19 \qquad \text{Formula 6.6}$$

The formulas do not yield the same answer (.2, .19). The values are close, but technically, "Yellow" and "*Y*" are not independent. The likelihood of a yellow ball occurring is influenced by the occurrence of a "*Y*" ball (and vice versa). ■

6.9 Bayes' Theorem

Up to this point most of the formulas in this chapter are what mathematicians call commutative. That is, we can "commute" or "move around" values and still get the same answer. For instance, *P*(*A or B*) = *P*(*B or A*). It does not matter which of the two events in question we call *A* and which we call *B* – we will get the same answer. This is also true of the multiplication rule, whether or not events *A* and *B* are independent. The *P*(*A and B*) = *P*(*B and A*). This is not true, however, for conditionals. *P*(*A*|*B*) is almost certainly not equal to *P*(*B*|*A*); only by pure coincidence might these be the same value. A little reflection can show us this rather quickly. For instance, the probability of being a biological female if one is a psychology major (perhaps around .7 at many universities) is not the same thing as the probability of being a psychology major if one is a biological female (this must be under .1 at most universities). Alternatively, take this example, the probability that a baseball player chews gum [*P*(*gum chewer*|*baseball player*)] is almost certainly not the same value as the probability of those who chew gum playing baseball [*P*(*baseball player*|*gum chewer*)]. These are very different questions.

Unfortunately, it is a common human foible to think that when we know one, we also know the other. This tendency is referred to in the academic literature as the "confusion of the inverse" or "conditional probability fallacy" (e.g. Plous, 1993; Villejoubert & Mandel, 2002). Oftentimes people will be given one bit of information, for example, the degree to which political liberals vote democratic [or $P(vote\ democratic|liberal)$], and unknowingly transpose the relationship [$P(liberal|vote\ democratic)$]. In the first case the answer might be around .9 (of those who consider themselves politically liberal, about 90% vote democratic); but the other conditional might be very different (of those who vote democratic, only about 60% might consider themselves politically liberal). Of course some conditionals may be more implicitly understood as only working one way (e.g. if we suppose that 80% of pickup truck owners are males, this does not mean that 80% of males own pickup trucks), but a lack of familiarity with the variables can often lead to inadvertently flipping the relationship between the conditional and the event in question.

However, if we know just a few more bits of information, we can figure out $P(B|A)$ if we know the $P(A|B)$ thanks to the work of an eighteenth-century English cleric named Thomas Bayes (learn more about Bayes in Spotlight 6.1). But before we look at the classic version of **Bayes' theorem**, we need to make a couple more observations. Both observations concern the "not" concept in probability. Up to this point we have been asking questions about how likely something is to occur, but of course we could frame the probability question from the perspective of how likely something is *not* to happen. For instance, we could ask, "what is the probability that "*notA*" is going to happen?" To help clarify our thinking here, we need to recognize that the probability of event *A* and event *notA* combined is always 1. Together they make up what probability theorists refer to as an "exhaustive set." We can explore this a bit by asking ourselves questions like "what is the probability that we are wearing blue socks or not wearing blue socks?" Well, that probability is 1, correct? What is the probability that it is going to rain tomorrow or not rain tomorrow? Again, the probability is 1. If the probability that a randomly selected person is of retirement age is .2, then the probability that a randomly selected person is not of retirement age must be .8. Therefore, if we know the probability of event *A*, we can always deduce the probability of event *notA*.

The last needed observation merely extends this line of thinking into the world of conditionals. The "not" concept can also be used as a descriptor for the sample space, the condition for the event. We can ask ourselves what is the likelihood that we will enjoy our meal if we go to Restaurant *A* [$P(enjoy\ meal|Restaurant\ A)$], but we can also ask ourselves about the probability that we will enjoy our meal if we do not go to Restaurant *A* [$P(enjoy\ meal|not\ Restaurant\ A)$]. When it comes to these probabilities, the pair of occurrences based on the conditional and the not conditional rarely add up to 1. For instance, in this case it may be fairly likely that we will enjoy our meal whether

Spotlight 6.1 Thomas Bayes and Bayesianism

Thomas Bayes (1701–1761) was a nonconformist (a term used for those who had problems with the Church of England), English cleric, statistician, and philosopher (Bellhouse, 2001). Although his interests were broad and his writings ranged from theology to a defense of Newton's ideas regarding calculus, he is most well known for a posthumously published paper by a friend in which he formulated a specific case of the theorem that now bears his name (Bayes' theorem; see Section 6.9). His theorem solved the problem of inverse probability (also known as the "confusion of the inverse" or "conditional probability fallacy"). As a result of his broad contributions to mathematics, Bayes was elected as a Fellow of the Royal Society sometime in the mid-1700s and prior to his death in 1761. This was the most prestigious British association for individuals who had been deemed to have made substantial contributions to the improvement of what was called "natural knowledge."

The currently used term Bayesianism comes not only from Bayes' own writings but also from the work of a French scholar named Pierre-Simon Laplace (1749–1827) who used Bayes' ideas to develop a way to think probabilistically about events that may not be part of a known "reference class" (e.g. Stigler, 1986) or what we have previously referred to in our text as the "sample space." This way of thinking allowed probability theorists to reason about the accuracy of various speculative hypotheses by first assigning prior probabilities, which were to be later updated to posterior probabilities in the light of new and relevant data in a recursive system of thinking. What we now call Bayesianism (or Bayesian probability) is the standard set of procedures and formulae utilized for this sequence of calculations.

we go to Restaurant A or somewhere else (at least we can hope). There is no need for these two events to sum to one; they are not complementary. However, given the proper information, we can determine $P(A|notB)$. For instance, what is the probability of getting the queen of spades given that we did *not* get a red card or $P(queen\ of\ spades|not\ red\ card)$? Because we understand the sample space well enough, we can figure this out (1 favorable card out of 26 total cards = 1/26).

Now we are ready to learn Bayes' famous theorem. If we know $P(A|B)$ and want to determine the $P(B|A)$, we will additionally need the $P(B)$, $P(notB)$, and $P(A|notB)$. (Actually, all we need is one or the other of the first two. For example, if we know $P(B)$, we can deduce the $P(notB)$; and vice versa.) Following is Bayes' theorem.

Bayes theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|notB)P(notB)} \text{ (Formula 6.8)}$$

Let us try our hand at a typical Bayes' theorem problem.

■ **Question** *Imagine it is true that 1% of 40-year-old women who participate in a routine screening have breast cancer. Further imagine that 80% of women with breast cancer will receive a positive reading from the mammogram screen procedure. However, 9.6% of women without breast cancer will also receive a positive reading from the mammogram screening procedure (this is sometimes referred to as a "false positive" result). Now suppose a 40-year-old woman is told that her mammogram screening is positive for breast cancer. What is the likelihood that she actually has breast cancer?*

Solution

Step 1. Let us first transpose the variables in our example into the terms used by Bayes' theorem, namely, $P(B|A)$, $P(A|B)$, $P(B)$, $P(notB)$, and $P(A|notB)$. Recall that we are trying to determine the probability that a person with a positive mammogram reading does indeed have breast cancer. This can be stated in probability language as $P(breast\ cancer|positive\ reading)$. Since the formula is set up to find $P(B|A)$, this means that event A corresponds to the event *positive reading* and event B corresponds to the event *breast cancer*.

Step 2. This means the following assignments should be true:

$P(A\|B)$	$=P(positive\ reading\|breast\ cancer)$
$P(B)$	$=P(breast\ cancer)$
$P(A\|notB)$	$=P(positive\ reading\|not\ breast\ cancer)$

It follows then that

$P(A\|B)$	$=.8$
$P(B)$	$=.01$
$P(A\|notB)$	$=.96$

And we can deduce that

$P(notB)$	$=.99$

Step 3. Use Bayes' theorem to solve the equation

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|notB)P(notB)}$$

$$P(B|A) = \frac{.8(.01)}{.8(.01) + .096(.99)} = \mathbf{.078}\ \textit{or about}\ \mathbf{7.8}\%$$

This may seem surprising to us. We probably thought the chance of this lady actually having breast cancer was much higher. However, as we stop to think about it, we may realize that about 10% of the 40-year-old ladies getting screened who do not have breast cancer (which is about 99% of them) are going to get a positive mammogram. That's a large number of false positives. This is intentional, as the inconvenience and sense of alarm that may result from receiving a false positive pale in comparison to the need to avoid false negatives. In reality, positive screens for breast cancer result in secondary screening procedures that are more sensitive and designed to distinguish between these initial false positives and true positives. ■

Summary

This chapter and Chapter 7 serve as the theoretical "bridges" that connect descriptive statistics to inferential statistics, the remaining material in the textbook. Inferential statistics, the ability to draw inferences about populations based on known properties of samples drawn from those populations, is dependent upon several concepts related to probability theory and hypothesis testing.

Probability theory started in the seventeenth century by several key thinkers who decided it was best to approach situations with a sense of willful ignorance regarding specific outcomes and the many idiosyncratic issues associated with them and to rather focus on determining likelihood over multiple trials. Out of this thinking emerged modern probability theory.

Probability can be understood mathematically as a proportion that ranges from 0 to 1. A probability of 0 means that an event is certain to not occur; a probability of 1 means that an event is certain to occur. A distinction is made between sampling with and without replacement. Sampling with replacement is a method of sampling whereby a member of a population is randomly selected and then returned to the population before the next member is selected. Sampling without replacement is a method of sampling in which a member of a population is not returned to the population before selecting another member of the population. Since hypothesis testing concepts are based on determining likelihood when in situations with replacement, this chapter will restrict itself to these situations.

There are various formulas that can be used to determine specific probabilities: the basic probability formula, the "or" formulas, the "and" formulas, the conditional probability formula, and Bayes' theorem. To distinguish between the "or" formulas, the concept of "mutual exclusivity" is needed. To distinguish between the "and" formulas, the concept of "independent" is needed. Bayes' theorem allows us to avoid the problem of inverse probability (also called the "confusion of the inverse" or "conditional probability fallacy").

Key Formulas

Probability of favorable event

$$P = \frac{(number\ of\ \ favorable\ events)}{(total\ \ number\ of\ \ events)} \quad \text{(Formula 6.1)}$$

Addition rule formula for two mutually exclusive events

$$P(A\ \ or\ \ B) = P(A) + P(B) \quad \text{(Formula 6.2)}$$

Addition rule formula for more than two mutually exclusive events

$$P(A\ \ or\ \ B\ \ or\ \ C\ \ or \ldots Z) = P(A) + P(B) + P(C) + \cdots + P(Z) \quad \text{(Formula 6.3)}$$

Addition rule formula for two events

$$P(A\ \ or\ \ B) = P(A) + P(B) - P(A\ \ and\ \ B) \quad \text{(Formula 6.4)}$$

Multiplication rule formula for two independent events

$$P(A\ \ and\ \ B) = P(A)P(B) \quad \text{(Formula 6.5)}$$

Multiplication rule formula for two events

$$P(A\ \ and\ \ B) = P(A|\,B)P(B) \quad \text{(Formula 6.6)}$$

Conditional probability formula

$$P(A|B) = \frac{P(A\ and\ B)}{P(B)} \quad \text{(Formula 6.7)}$$

Bayes' theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|notB)P(notB)} \quad \text{(Formula 6.8)}$$

Key Terms

Inferential statistics
Probability
A priori (classical) approach
A posteriori approach
Addition rule
Mutually exclusive events
Multiplication rule
Probabilistic independence
Probabilistic dependence
Conditional probability
Sample space
Bayes' theorem

Questions and Exercises

1 Together, this chapter and Chapter 7 allow the researcher to (please select the best answer):
- **a** Understand how to run inferential statistics
- **b** Determine the nature of samples from populations and the nature of populations from samples

c Think philosophically about numbers
d Understand the relationship between numbers and people

2 Assuming all events are equally likely, please determine the following probabilities:
a Selecting a spade from a deck of 52 cards
b Rolling a 6 on a 10-sided die
c Randomly selecting a pawn from a set of chess pieces
d Selecting North Carolina from a lottery draw involving all 50 states
e Missing a pop quiz in a Psychology 100 class the one day class it is skipped

3 Which of the following events are mutually exclusive?
a Being a resident of Country A; being a resident of Country B
b Making one toss of a die and obtaining either an even number or a 2
c Drawing a 5 and a 2 on a single draw from a deck of cards
d A person has black hair and blue eyes
e Obtaining a 1 and a 6 when rolling two dice at once
f Being a dog; being a cat
g Being pregnant; being not pregnant
h Being a Yankees fan; being a Red Sox fan
i Wearing an official Star Trek shirt; getting a date

4 Suppose there is a bin with 40 red marbles and 60 white marbles. In each of the following problems, selections are made blindly, and the marbles are randomly distributed throughout the bin.
a What is the probability of picking a red marble?
b What is the probability of drawing a white marble?
c What is the probability of drawing either a red or a white marble?
d What is the probability of picking two red marbles in a row, with the first marble replaced?
e What is the probability of drawing two white marbles in a row, with the first marble replaced?

5 For each of the following situations, specify whether the events are independent or dependent.
a The weekly state lottery is conducted by drawing six numbers from a bin that has 54 balls, each with a number. Is the successive selection of balls during the drawing an instance of independent or dependent events?
b The six balls selected one week and the six balls selected the next week.
c Drawing two 5's from a deck of cards without replacing the first card.
d Drawing a 3 from a deck of cards, replacing it, and then drawing a 6.

6 Suppose we have an urn with the following set of balls:

10	Red – X
20	Red – Y
10	Green – X
30	Green – Y
15	Yellow – X
15	Yellow – Y

Find:

a $P(Red)$
b $P(Green)$
c $P(not\ Red)$
d $P(Red\ or\ Yellow)$
e $P(Red\ or\ X)$
f $P(Red\ and\ X)$
g $P(Red|X)$
h $P(X|Red)$
i Are events Red and X independent?
j Why or why not?

7 Suppose we have an urn with the following set of balls:

20	Green – X
15	Green – Y
10	Red – X
15	Red – Y
30	Yellow – X
10	Yellow – Y

Find:

a $P(Green)$
b $P(Y)$
c $P(Red\ or\ Yellow)$
d $P(Red\ or\ Yellow\ or\ Green)$
e $P(Red\ or\ X)$
f $P(not\ Green)$
g $P(X\ or\ Y)$
h $P(Yellow\ and\ X)$
i $P(Yellow\ and\ Red)$
j $P(Red|Y)$

k $P(Y|Red)$
l Are events X and *Green* independent? Why or why not? If not, what probabilities might we change to make them independent?
m Are events *Yellow* and Y independent? Why or why not? If not, what probabilities might we change to make them independent?

8 Suppose that we have an urn with red and green balls with an X or Y on them (i.e. four different kinds of balls, namely, Red – X, Red – Y, Green – X, Green – Y). Suppose that we know the following:

$P(Red) = .2$
$P(Red\ and\ X) = .1$
$P(X) = .3$

Find:
a $P(Green)$
b $P(Red\ or\ X)$
c $P(Red|X)$
d $P(X|Red)$
e Are events Red and X independent?
f Why or why not? If they are not independent, change one of the probabilities given to make them independent.

9 Suppose we have an urn with the following set of balls:

15	Green – X
15	Green – Y
10	Red – X
20	Red – Y
25	Yellow – X
15	Yellow – Y

Find the following probabilities:
a $P(Green)$
b $P(Y)$
c $P(Red\ or\ Yellow)$
d $P(Red\ or\ X)$
e $P(not\ Green)$
f $P(X\ or\ Y)$
g $P(Yellow\ and\ X)$
h $P(Yellow\ and\ Red)$
i $P(Yellow|Y)$

j $P(Y|Yellow)$

k Are events X and *Green* independent? Why or why not? If not, what probabilities might we change to make them independent?

l Are events *Yellow* and Y independent? Why or why not? If not, what probabilities might we change to make them independent?

10 Suppose the probability of being farsighted is .1. Suppose also that the probability of a farsighted person being dyslexic is .05 and the probability of a person who is not farsighted being dyslexic is .025 (1/2 as likely). What is the probability that a person with dyslexia is farsighted?

11 In baseball, suppose we are told that the probability of scoring a run on a double is .54. That is, given a play has generated a double, 54% of the time at least one run will score. However, we want to know how often when a run scores, it was generated by a double. This is not the same question. Do we see the difference? We are told that the probability of runs scoring on plays that are not doubles is .11 and the probability of hitting a double is 18%.

12 Suppose we are interested in finding out more about who buys our company's product – Nutrinut Peanut Butter. We know that about 44% of peanut butter buyers who come from families with 4 members or more choose Nutrinut over the other brands (that is, given a peanut butter buying 4-person family or more, 44% of them choose Nutrinut), but we want to know what percentage of Nutrinut buyers are from 4-person families or more. (Do we see how this is quite a different question?) We know that those who choose Nutrinut who are not in 4-person families or more is 36%, and we know that the percentage of peanut butter buyers who are in 4-person families or more is 57%. What percentage of Nutrinut buyers are in 4-person families or more?

Hint: $P(\text{Nutrinut}|\text{PB buyers from 4pf}) = .44$

13 One cab company in our city is named "Blue Cab Co." And they have had some complaints about the driving behavior of their employees. But we know that all cab companies have some drivers who are a bit reckless. We know that the probability of getting a reckless driver if we are in a Blue Cab car is .25, but what we want to know is if we have a reckless driver, how likely is it a Blue Cab that we are in? Do we see the difference? We know the probability of getting a reckless driver if we are not in a Blue Cab car is .15, and we know the probability of getting a Blue Cab car is .4. So, what is the probability of being in a Blue Cab car if we have a reckless driver?

14 In the dice game called craps, we win on the first roll if we get a 7 or an 11. (Two dice are rolled, and the numbers are added.) What is the probability we will win on the first roll? Hint: First find the probability of a 7 and then of an 11. To do each of these, find what combinations give us each number and what the probability of that combination is. Then find the probability of a 7 or an 11.

7

Hypothesis Testing and Sampling Distributions

7.1 Inferential Statistics

As mentioned in the previous chapter, inferential statistics originally developed, in part, from thinking about how to settle fairly interrupted gambling endeavors. The modern era of inferential statistics began in the late nineteenth century. One driving force was agricultural companies and breweries interested in assessing the influence of various treatments on crop yields. Another driving force was the desire by a growing number of social scientists to measure various economic, sociological, and psychological phenomena like employment rates, population growth, mental capabilities, and various developmental markers in children. As a result of these research interests, most of the basic statistical tools to be introduced starting in Chapter 8 and running through the remainder of the text were developed within the span of just a few decades, from the 1880s to the 1920s.

As mentioned in Chapter 6, inferential statistics, based on probability theory and the logic of hypothesis testing, is used to make inferences about the characteristics of a population from the characteristics of a random sample drawn from the population. For example, what if we wanted to know the level of reading skills among the high school students of a large city? We could proceed to test every student in the city (a costly and time-consuming effort), or we could test a random sample of all the students and use their scores to infer the reading skills of the entire student population. Given real-world limitations, ever-present budgetary constraints and the desire to obtain answers to questions as efficiently as possible, the use of inferential statistics has become a necessity for today's behavioral and social science researchers.

In inferential statistics, *the* key phrase is "random sample." A **random sample**, as noted in Chapter 1, is a sample of scores taken from a population in such a way that each score in the population has an equal chance of being included in

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

the sample. Random sampling maximizes the likelihood that the sample is *representative* of the population.

Inferential statistics requires random sampling, but this is not always easy to achieve. Consider a typical psychology experiment. The sample of participants is university students enrolled in an introductory-level psychology course who have volunteered for the experiment to fulfill a course requirement. What is the population to which we can generalize from this sample? Strictly speaking, the population is students who attend that particular type of university and who opt to take an introductory-level psychology course. This rather narrow population may or may not line up well with the population of "all university students" or even less so the population of "people in general." Because of the large number of psychology studies conducted with university students, we have to wonder to what extent the findings of many psychological studies apply only to university undergraduates. (This criticism has often been leveled at the field of psychology and is one that psychologists in recent years are working hard to correct.) The problem here is not one of mathematics, but rather one of logic. One cannot, based on the findings from a study, make statements about a group of people who are different (in important ways) from the participants in the study. In inferential statistics, the researcher is bound by the degree to which samples are representative of the populations wished studied. Given the uncertain nature of generalizing research findings to others *not* represented in the study sample, researchers are required to limit their conclusions to only populations that are well represented in the study sample.

Types of Inferential Estimation

Recall that a parameter is a numerical characteristic of a *population* (e.g. mean, standard deviation, variance, etc.), whereas a statistic is a numerical characteristic of a *sample*. Therefore, parameter estimation uses data from a sample to infer the value of a population parameter. There are two kinds of estimation: point estimation and interval estimation. Suppose we take a random sample and compute the mean. If someone were to ask us to estimate the mean of the population, we could use our sample mean to make a **point estimation.** Any sample statistic (e.g. mean, median, variance, standard deviation) can be used to make a point estimation of a population parameter. The other kind of estimation is called an **interval estimation** (or **confidence interval**). In this procedure, two values are stated within which it is believed the actual population value falls. With interval estimations, a formula is used to determine both the values creating the interval and the degree of confidence that should be given to the claim that the population value falls within this interval.

7.2 Hypothesis Testing

A **research** (or *scientific*) **hypothesis** is a formal statement or expectation about the outcome of a study. They are usually stated in terms of independent and dependent variables, and the relationship between them, or in terms of two variables and the degree of association between them. These statements are oftentimes derived from relevant preexisting theories and/or based on relevant previous research findings. For example, we might read something like the following, "Given that time pressures diminish the tendency to help (e.g. Darley & Batson, 1973; Moore & Tenney, 2012), participants in the low stress situation are predicted to be more helpful than participants in the high stress situation." The accumulation of knowledge in the behavioral and social sciences relies heavily on the process of formulating theories, stating hypotheses, gathering data to test hypotheses, revising theories, making new hypotheses, and conducting more research. Although **hypothesis testing** is conducted in many different contexts, all instances of hypothesis testing share common characteristics: the use of probability theory and inferential statistical concepts to extrapolate from sample data to relevant populations as well as efforts to quantify the possibility of making decision errors associated with these probability-based inferences.

The Use of Sample Data to Make Inferences About Populations

If the investigator of a study is *only* interested in drawing conclusions about the participants in the study, then hypothesis testing is irrelevant. Confining one's interest to the participants in the study defines those participants as the population; thus, no inference is required. One goal of the researcher, however, is to acquire general knowledge about our world. The behavior of our research participants is interesting only insofar as it allows us to make statements about the behavior of people who are *not* in our study – a larger population. Representative samples of participants are used to generalize study results to this larger group.

In order to use samples to infer the characteristics of populations, mathematical tools to accomplish the task are needed. Because, in a sense, the researcher is forced to work at the level of samples, anything concluded about populations always involves a degree of uncertainty. Inferences about population parameters are, therefore, necessarily probabilistic in nature. The statistical methods of hypothesis testing allow us to use sample data to make probabilistic statements about the credibility of research hypotheses, hypotheses that are always stated in population terms.

Decision Errors: An Unpleasant Fact of Life in Hypothesis Testing

Whenever a conclusion is drawn about a population based on sample data, there is a chance the conclusion will be wrong. For example, suppose we took two samples from a population and gave one sample technique *A* to manage their anxiety regarding an upcoming timed test and the other was given technique *B*. Further, suppose that those given technique *A* were subsequently shown to be less anxious and therefore more accomplished on the task than those given technique *B*. From this we might conclude that technique *A* is more effective than technique *B* for the entire population from which we drew our samples. Although this is probably an accurate conclusion to draw, there is always a chance we may end up being wrong. Despite our best efforts, the samples may not be representing the population well, and our conclusion about which technique is best for the population is dubious. For this reason, there are concepts and principles developed later in the text (first introduced in Chapter 8) to help the researcher think about, reduce, and quantify the level of uncertainty associated with any conclusions that are made. (One of these principles is replication. See Box 7.1 presented later in this chapter for a more developed argument in favor of replication.)

Research Hypotheses Versus Statistical Hypotheses

As previously stated, a *research hypothesis* is a statement based on relevant previous findings and/or a theory regarding the expected outcome of a study. It is the thesis that prompts the study; or in other words, it is the study's reason for being. A **statistical hypothesis**, however, serves as the vehicle for evaluating a research hypothesis. This is a numerical statement regarding the potential outcome features of a study. Some statistical hypotheses rely on previously known population parameters. These are sometimes referred to as single-sample research designs. For example, suppose an existing validated stress questionnaire has a known mean of 50. (This is important; the population mean is known ahead of time.) In other words, it is known that the average person will receive a score of 50. A community psychologist is interested in the emotional effects of natural disasters. Soon after an earthquake, 100 randomly sampled people are asked to complete the stress questionnaire. The research hypothesis may be stated as, "Natural disasters create stress reactions among the victims." (Note: research hypotheses are almost always based on differences; that is, something is supposed to change.) The statistical hypothesis can be stated as, "Does the mean stress score of the sample suggest the population mean from which this sample came is different from 50?" The job of the researcher, then, is to use gathered sample data to decide if there is enough evidence to conclude that the population mean is probably not 50.

Other statistical hypotheses are based on similarity or differences between two groups. These are sometimes referred to as two-sample research designs.

An example from medicine might be, "Heart patients who receive a beta-blocker will experience fewer cardiac arrhythmias than patients who receive a placebo." The corresponding statistical hypothesis might read, "Is there a difference in the mean number of arrhythmias between the 'beta-blocker' population and the 'placebo' population?" Notice that in these situations there are no known means; rather the question at issue has to do with whether there is a difference between the two population means ("beta-blocker" and "placebo"). The job of the researcher, then, is to use gathered sample data from both populations to decide if there is enough evidence to conclude that the population means are probably not the same.

A majority of current experimental research, however, involves more than two samples. For example, an experimental psychologist might formulate the research hypothesis: "Does the magnitude of reinforcement influence the number of trials it takes to learn a task?" The psychologist might then examine the influence of five different incentive conditions on learning. The statistical hypothesis might be, "Is there a difference among the populations in the mean number of trials required to learn the task?"

In the behavioral and social sciences, hypothesis testing does not always take place within an experimental context. The correlational approach is an alternative research method that differs from experimentation in that it does not attempt to exert an influence on a measured response. Because variables are not controlled, correlational research cannot identify causal relations among variables. (Correlational designs were initially presented in Chapter 1.) Instead, this approach attempts to find variables that relate to one another. An example of a correlational research hypothesis is, "First-time parents who have newborns who cry a lot are more dissatisfied with their marriages than first-time parents who have quiet babies." The investigator would search for an association between the amount of time an infant cries and the parents' reports of marital dissatisfaction. No attempt is made to manipulate some variables and hold other variables constant. (For instance, it would be unethical to randomly assign noisy and quiet babies to parents and observe the hypothesized deterioration of the marital unit.)

Although the correlational approach does not manipulate variables, it does *not* alter the fundamental characteristics of hypothesis testing. The job of the researcher is to use gathered sample data to decide if there is enough evidence to conclude that there is a relationship between the variables at the population level.

More on Statistical Hypotheses: The Null and Alternative Hypotheses

As previously noted, statistical hypotheses are numerical statements regarding the potential outcomes of an experiment. When conducting a study, statistical hypotheses always come in pairs, a null hypothesis, denoted H_0, and an alternative hypothesis, denoted H_1. In the context of an experiment, the **null**

hypothesis states that there is *no* effect of the independent variable on the dependent variable. In a correlational context, the null hypothesis states that there is no relationship between two variables. It may help to realize that the word "null" in hypothesis testing means, quite literally, *nothing* – as in, no differences between groups or no relationships between variables, as the case may be. The **alternative hypothesis** states that there *is* something going on. In the context of an experiment, it means that there is effect of the independent variable on the dependent variable. In a correlational context, it states that there is a relationship between two variables. The alternative hypothesis is the logical alternative to the null hypothesis.

Box 7.1 Is the Scientific Method Broken? The Value of Replication

This is another box in the series looking at the reproducibility crisis in the social, behavioral, and medical sciences. When researchers conclude that the null cannot be rejected (also known as "failing to reject the null hypothesis"), the study's findings are deemed "nonsignificant." This term is a way of expressing the idea that any differences between the sample means of the various conditions in a study are not substantial enough to warrant rejecting the null hypothesis of no difference. (The degree of differences needed to be found between sample means before the null hypothesis can be rejected is a topic that will be carefully explored in future chapters.) Unfortunately, most journals in the social sciences are not interested in publishing research that has not found evidence of differences between conditions, so-called nonsignificant findings. Each journal wants to include only articles that seem original and important and are likely to be read and referenced by others.

This policy, however, can create a problem. Imagine several researchers working independently of each other, each of them looking at a similar research question. Furthermore, imagine their research hypothesis is, in the end, not a very good one. That is, perhaps the null hypothesis is actually true; the independent variable has no effect on the dependent variable. However, given the sampling error that naturally occurs within a sampling distribution, suppose one of the researchers gets an extreme sample mean that prompts them to reject the null. The researcher is in error; their sample mean was very unusual. However, they do not realize this and believe they are correct in rejecting the null hypothesis and claiming to have found an effect. If all of the researchers who were looking at the same topic published their findings, readers might become suspicious of the one study showing an effect amid the many others that do not. However, readers will not be exposed to these other (nonsignificant) findings. These studies will not be published by the journals. The only study that will be published is the one that found evidence to reject the null. Once we realize this, it is easy to see how the current publication practices in the social, behavioral, and medical sciences create the possibility that the findings of a number of published studies may not be reliable.

A potential remedy for this is the replication of published studies. Unfortunately, replication is not particularly valued in the profession and so is rarely

performed. For example, according to Makel, Plucker, and Hegarty (2012), only about 1.6% of all published studies in the top 100 psychology journals from 1900 to 2012 were replication attempts. Simply stated, replications are rarely published. However, in recent years a growing number of researchers are serving the scientific community by carefully and painstakingly replicating published research findings. A good example of this growing trend is the Center for Open Science (https://cos.io/). Hopefully, the value of replication efforts will continue to rise in the social, behavioral, and medical sciences in the wake of this reproducibility crisis.

The null and alternative hypotheses are statistical hypotheses and are therefore numerical expressions. For example, if an educational enrichment program is expected to increase IQ, and we know that the average IQ in the population is 100, the null and alternative hypotheses are stated as

$$H_0: \mu = 100$$
$$H_1: \mu \neq 100$$

The null hypothesis can be determined as untenable whether the program increases *or* decreases IQ. To reject the null hypothesis in favor of the alternative hypothesis, we will need to find statistical evidence that the null is likely to be false.

Taken together, the null and alternative hypotheses are *mutually exclusive* (i.e. they cannot both be correct) and *collectively exhaustive* (i.e. one of them has to be correct). Consider the statistical hypothesis that a population mean is 50. Well, the population mean either equals 50 or it does not, symbolically either $\mu = 50$ or $\mu \neq 50$. The hypotheses are mutually exclusive in that only one of them *can* be true. They are collectively exhaustive because one of these hypotheses *has* to be true.

Notice also that statistical hypotheses are always made in reference to population parameters, not sample statistics. We will gain sample means or sample correlations from our data, but our inferential decision will always be in reference to a population parameter (usually a mean, a mean difference, or a correlation).

Truth in Hypothesis Testing

Once the data are analyzed in hypothesis testing, the researcher has to make a decision about whether or not to reject the null hypothesis. The two options available are to *reject* the null hypothesis (in favor of the alternative) or to *not reject* the null hypothesis – in other words, to *fail to reject the null hypothesis*. Notice that there is no option to accept the null hypothesis. This is because it is not possible, on the strength of sample data alone, to demonstrate that the null is true. A poorly designed study with insensitive measurement instruments may fail to detect an experimental effect or a relationship between variables.

The use of a small number of participants may make it difficult to infer accurately a true population mean or detect whether two variables are related. Even if the study is expertly designed, with a large number of participants, and the samples yield identical means, this is still an insufficient reason to conclude that the null is true. Therefore, *failure to reject the null hypothesis does not necessarily mean that the null hypothesis is true*. It could be, of course, but it does not have to be true. This is one of the necessary limitations of using samples to make inferences about populations. For this reason, when we fail to reject a null hypothesis, the conclusion should be that we simply do not know if the null is true or not; it is much more like a scientific "shrug-of-the-shoulders" than a pronouncement that the null is correct. Unfortunately, this is a common error in statistical interpretation.

On the other hand, if the statistical analysis suggests the null hypothesis should be rejected, this only means that the alternative hypothesis is *most likely* true. For example, two samples that seem to be different from each other and that lead us to reject the null hypothesis may, in reality, be drawn from the same population, thus making the null hypothesis correct.

Of course, in actuality either the null hypothesis is true or it is not. However, we are not able to make this determination with certainty from merely the features of the accessible sample or samples. The methods of hypothesis testing are probabilistic, and probability, by definition, means that there is some level of uncertainty. This may seem unsettling, but it is simply unavoidable. Even though decision errors can occur in this process, we can at least begin to quantify and understand the likelihood of making these errors. This topic will be more carefully addressed in the next chapter, once we start using inferential statistics to make judgments regarding the null hypothesis.

However, before leaving this topic, it is important to stress the posture difference between research hypotheses and statistical hypotheses. Research hypotheses are usually not stated numerically, are derived from theory and/or previous findings, and, most importantly, are almost always suggestive of a difference, an effect, or a relationship. In short, research hypotheses usually predict that something interesting is happening. Otherwise, why are we interested in performing the research? Compare this with statistical hypotheses. Here, the default position is that nothing is going on and evidence suggesting otherwise will have to be found before the null will be rejected. The statistical hypothesis initial posture is skeptical (as if to say, "show me!"). While our heads and hearts may be predicting and hoping to find some effect or relationship, the logic and mathematics of hypothesis testing start with the presumption that nothing is going on.

One final point to be made concerns the relationship between rejecting the null hypothesis and supporting the research hypothesis. Because the alternative hypothesis covers both sides of the null hypothesis, rejecting the null may or may not support the research hypothesis. It is possible to reject the null hypothesis and yet still *not* find supporting evidence for the research hypothesis.

For example, suppose the research hypothesis is that the experimental group will outperform the placebo group. Suppose further that statistical evidence is found that the placebo group actually outperformed the experimental group. Even though the null hypothesis that they are equal will be rejected, this decision does not support the research hypothesis; the findings were opposite of the predicted direction. It is important to realize that rejecting the null hypothesis does not always mean the researcher's inclinations were correct.

7.3 Sampling Distributions

Statistical hypothesis testing is the topic of interest for the rest of this textbook. The conceptual foundation of statistical hypothesis testing and the application of formulas to research data will be emphasized. These two topics are intertwined. Without understanding the conceptual basis of statistical hypothesis testing, there is no way to understand why a formula looks as it does and why the formula should be used in a given situation.

The most important concept in inferential statistics is the sampling distribution. As we read the remainder of this chapter, be forewarned that the connection between sampling distributions and hypothesis testing may not be immediately obvious. Subsequent chapters will deepen our understanding. For now, understand that if statisticians had not worked through the characteristics of sampling distributions, statistical hypothesis testing would be impossible.

Population and Sample Distributions

We know the difference between a population and a sample, the latter being a portion of the former. Every population of scores can be depicted as a frequency distribution that reflects the frequency of occurrence of every score in the distribution. Frequency distributions are also called probability distributions because the probability of selecting a given score at random depends on the frequency with which that score occurs in the population. In addition, if we know the mean and standard deviation of a normal distribution, we can make probability statements about the likelihood of selecting a score from a specified area of the distribution (remember all those z score problems?).

As has been stated already, samples have value only insomuch as they represent the population from which they come – the population we wish to study. Most importantly, when we make inferences from a sample to a population, we do not make inferences about a specific *score*. Instead, we make inferences about a population parameter from a sample statistic. For example, we may infer the mean or the variance of a population based on the mean or variance of our

sample. It is meaningless to take a single score from a sample and try to estimate a single score of a population.

In the discussion of *z* scores in Chapter 5, questions were asked like, "What is the probability of selecting a *score* of less than 20, given that the mean of the population is 26 and the standard deviation is 4?" We were able to answer such questions because we used a standardized normal distribution, the *z* score distribution. The *z* score distribution is a distribution of transformed raw *scores*. However, in hypothesis testing, we are interested in population parameters – means (usually), never individual scores. If we want to make a probability statement about a randomly selected sample *mean* falling within a specified area under the normal curve, we need to create a *normal distribution of means*, rather than of raw scores. This topic is addressed in the following section.

A Sampling Distribution of Means

Generally stated, a **sampling distribution** is a theoretical frequency distribution of a statistic (usually a mean) based on a very large number of repeated samples of some specified size, *n*. With respect to means, a sampling distribution shows the relative frequency of all possible values of sample means, given the selected sample size. A researcher never actually goes through the process of constructing a sampling distribution. As was noted above, it is a theoretical mathematical concept; however, it is critically important to the hypothesis testing process. The concept of creating a sampling distribution is easily grasped by walking through the steps that, *theoretically*, would be taken to construct it.

Step 1. Choose a population of scores. It might be a population of IQ scores, heights, weights, or scores from a population of participants who have taken some personality inventory. It can be quite literally any population of scores that are measured on an interval or ratio scale.

Step 2. Decide on a sample size, *n*. The sample size can range from 2 to infinity. (Since means are being used, a sample size of 1 has questionable meaning.)

Step 3. Take a random sample of size *n*, the sample size decided upon in step 2.

Step 4. Compute the sample mean and replace the scores back into the population. This is called *random sampling with replacement* (as referenced in Section 6.4).

Step 5. Repeat steps 3 and 4 an almost infinite number of times. *Each repeated sample must be the same size as was selected in step 2.* How do we know when we are finished? At some point, it will be impossible for us to select randomly a sample that we have not already selected. In other words, all possible samples of size *n* will have been selected; there is no combination of participants of size *n* left that we have not already drawn.

Step 6. Finally, plot the relative frequency distribution of the means. This distribution of means is the sampling distribution of the population of scores *for that chosen sample size*.

In specifying the steps of constructing a sampling distribution, the statistic of interest for us was the mean. The steps are the same no matter what statistic we select. If we want a sampling distribution of variances, use the same procedure, except compute and plot variances. We could even take two samples, calculate the means, take the difference between the two means, and establish a sampling distribution of mean differences (techniques like this will be featured in Chapters 9 and 10). There are, in fact, many occasions in hypothesis testing when the sampling distribution to be used will not be a distribution of means. Here, however, as the concept is introduced, the discussion of sampling distributions will be confined to means.

Characteristics of Sampling Distributions

The Central Limit Theorem

This section begins with a theorem, the most important theorem in the entire field of inferential statistics, formulated in 1810 by Pierre Laplace (1749–1827). It is called the **central limit theorem.**

There are several claims made by the central limit theorem that are important for understanding sampling distributions. First, if the population is normally distributed, the sampling distribution of means will be normally distributed. Second, even if the scores in the population are *not* normally distributed, assuming n is sufficiently large, the sampling distribution of means will be normally distributed. In other words, unless the raw population of scores is wildly nonnormal *and* the selected sample size is rather small, the resulting sampling distribution will be normal.

The Mean of the Sampling Distribution, μ_M

Third, the central limit theorem states that the mean of a sampling distribution is the mean of all the sample means in the distribution and is symbolized as μ_M (the population mean of all of the sample means). Further, *the mean of the sampling distribution has exactly the same value as the mean of the population of raw scores*. Therefore, $\mu = \mu_M$.

The Standard Deviation of the Sampling Distribution, σ_M

How would we compute the standard deviation of a sampling distribution? In Chapter 4 we learned that the formula for computing the standard deviation of a population of *scores* is

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}.$$

The formula for the standard deviation of a sampling distribution, σ_M, has the same form as the standard deviation of a population. This formula is a *definitional* formula since it defines the standard deviation of the sampling distribution. Although Formula 7.1 is not actually used to compute the standard deviation of a sampling distribution, it is presented here to underscore the point that the standard deviation of a sampling distribution has the same mathematical form as the standard deviation of a population.

Definitional formula for the standard deviation of a sampling distribution

$$\sigma_M = \sqrt{\frac{\sum (M - \mu_M)^2}{N_M}} \quad \text{(Formula 7.1)}$$

Since the "scores" of a sampling distribution are means, M replaces X. In addition, since the mean of a sampling distribution is μ_M, μ_M replaces μ. The denominator of Formula 7.1 is N_M; this is the number of sample means that comprise the sampling distribution. We would never use Formula 7.1 to calculate the standard deviation of a sampling distribution because N_M can approach infinity for some populations and in most cases is simply indeterminable.

The Relationship Between σ and σ_M

Whereas the mean of the sampling distribution is identical to the mean of the population, $\mu = \mu_M$, the standard deviation of the sampling distribution is not equal to the standard deviation of the population. The two measures are related, however. Formula 7.2 shows this relationship. The standard deviation of the sampling distribution of means is called the **standard error of the mean** or simply the *standard error*.

Standard error of the mean

$$\sigma_M = \frac{\sigma}{\sqrt{n}} \quad \text{(Formula 7.2)}$$

Note the denominator of the standard error. The n refers to the sample size selected when taking repeated samples during the theoretical construction of the sampling distribution. Remember, when constructing a sampling distribution, every sample needs to be the same size. Therefore, the relationship between σ and σ_M depends on the size of the samples being drawn. As a result, the variability of the sampling distribution is determined by the variability of the population distribution *and* the size of the samples used.

There is some potential for confusion when discussing the means that comprise the sampling distribution. We are used to thinking of a distribution of *scores* that has one mean. A sampling distribution is a distribution of means. The mean of the sampling distribution is the mean of all the sample means, μ_M. For example, if we move one standard error away from the mean of the

sampling distribution, we will land on a sample mean, not *the* mean of the distribution, but *a* mean that could be found when sampling from the population (using the selected sample size).

How and Why *n* Affects the Standard Error of the Mean

Suppose two different sampling distributions from the same population are established. In the first sampling distribution, the sample size is 4. In the second sampling distribution, the sample size is 25. The standard deviation of the population is 32. Using Formula 7.2,

$$\sigma_M = \frac{32}{\sqrt{4}} = 16 \quad \textit{and} \quad \sigma_M = \frac{32}{\sqrt{25}} = 6.40$$

The standard error of the mean becomes smaller as the sample size becomes larger. This is not the case with the mean of the sampling distribution; $\mu = \mu_M$ no matter what the size of n. Step 2 in constructing a sampling distribution required us to specify the size of the samples. By using the preceding equation, we can see that n influences the variability of the sampling distribution; specifically, as n increases, the standard error decreases. However, we may be wondering, "What are the features of the *real* sampling distribution?" or "What size of n corresponds to the best sampling distribution?" First, there is not just one sampling distribution for a single population. There are as many sampling distributions for a population as there are sample sizes. Statisticians describe this situation by using the term *family of sampling distributions*. Whenever we refer to a sampling distribution, we are actually referring to a sampling distribution of a particular size n. Keeping in mind that sampling distributions are theoretical, one sampling distribution is just as "real" as the next. Which one is the best? Well, there is no best one. In general, the smaller the standard error of a sampling distribution, the better for rejecting null hypotheses, but there are other factors to consider. Also, recall that as n increases, the resulting sampling distribution approaches normality *even if* the population of raw scores is not normally distributed. This means that as n increases the sampling distribution becomes more and more like a normal distribution. How large does the n have to be before one can be sure the sampling distribution is normal? Well, the standard answer is that an n of 30 will generate a normal sampling distribution. However, it depends on the shape of the population of raw scores. If it is normal, then even small n's will yield normal sampling distributions. If it is not normal, then the more non-normal it is, the higher the n will need to be to produce a sampling distribution that is normal; perhaps an n of more than 30 will be needed. One more point needs to be made that will be further developed later in the text; as the sample size increases, the sampling distribution will not only approach normality, it will approach *the standard normal distribution*; the same distribution described in detail in the z table (Table A.1).

Given two sampling distributions, why does the one with a smaller n have a larger standard error? As an example, let us contrast a sampling distribution in which $n = 2$ with one in which $n = 20$. Each one of the means of the sampling distributions is comprised of scores. When $n = 2$, there are two scores that are averaged to arrive at the mean for that sample. When $n = 20$, there are 20 scores averaged to arrive at the mean for that sample. When thinking about unusually high or low sample means, which is more *unlikely*, selecting two extreme scores to get an extreme sample mean or selecting 20 extreme scores to get an extreme sample mean? Although we do not usually get two extreme scores in a row, it is much less likely that we would get 20 extreme scores within the same sample. In other words, with a small sample, there is little opportunity for other scores to compensate for selected extreme scores; in large samples, it is more likely that other scores will be included in the sample to compensate for selected extreme scores. Therefore, a sampling distribution based on $n = 2$ will contain many more extreme means than a sampling distribution in which $n = 20$. (Try this thought experiment – imagine having a sample size that is almost the size of the population itself, say, $N - 1$. If we were to generate a sampling distribution by sampling the entire population minus one, replacing the scores, then sampling them again with all but one, and do this repeatedly, would not the resulting sampling distribution have an extremely small standard error? Every sample mean would be virtually the same score.) Having only a few extreme scores produces a population or sample distribution with a relatively small standard deviation; having only a few extreme means produces a sampling distribution with a relatively small standard error of the mean. Figure 7.1 shows how the shape of a sampling distribution is affected by changes in the size of the samples.

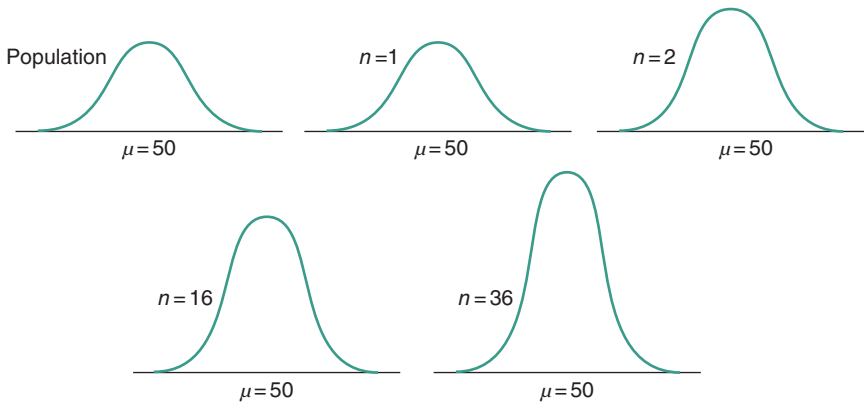


Figure 7.1 Several sampling distributions constructed with different sample sizes. As the sample size increases, the standard error, σ_M, reflected in the width of the distribution, correspondingly decreases.

The following list summarizes the main points of sampling distributions of means:

1) A sampling distribution of means is a theoretical distribution derived by computing the means of an almost infinite number of samples of size n.
2) If the scores of a population are normally distributed, irrespective of sample size, the sampling distribution of means will be normally distributed.
3) If the population distribution is not normally distributed, the sampling distribution approximates a standard normal curve as the sample size increases. The more the population distribution deviates from normality, the larger the sample size must be to establish a normally distributed sampling distribution.
4) As n increases, the shape of the resulting distribution approaches the standard normal distribution, as displayed in the z table (Table A.1).
5) The mean of the sampling distribution equals the mean of the population of raw scores, $\mu = \mu_M$.
6) The standard deviation of the sampling distribution is called the standard error of the mean. The relationship between the standard error of the mean and the standard deviation of the population is $\sigma_M = \sigma/\sqrt{n}$.

Box 7.2 will show us how we can use software programs found online to create and adjust sampling distributions.

Box 7.2 Playing with the Numbers: Create Our Own Sampling Distribution

Programs found on the Internet allow us actually to see how changing the sample size, mean, and the standard deviation of the population of raw scores change the resulting sampling distribution. Some of the ones recently found online include the StatKey Sampling Distribution for a Proportion program (www.lock5stat.com/StatKey/sampling_1_cat/sampling_1_cat.html), the Rice Virtual Lab in Statistics (onlinestatbook.com/stat_sim/sampling_dist/), and the Rossman/Chance Applet Collection (www.rossmanchance.com/applets/OneSample.html). There are others. A program that is quite flexible, however, is one created by Dr. Patrick Wessa (www.wessa.net/rwasp_samplingdistributionmean.wasp). In this program, we can input the number of replications we want to make (theoretically, sampling distributions involve the number of replications that equals the total number of samples of size n that are possible, often in the billions or more, so choose a number in the hundreds at least), the sample size (n), the mean of the population of raw scores, and the standard deviation of the population of raw scores. (It is recommended to leave the width and height of the chart as the given default values.) After imputing some values, click the "compute" button, wait for the program to run, and then scroll down to

the second chart. Here we will see a graph reflecting the various sample means given our selected sample size, mean, and standard deviation.

By changing the values of the imputed numbers, we should be able to see the general principles of the sampling distribution concept as well as the claims of the central limit theorem on display. For instance, if we increase the sample size, we should see the standard deviation of the sampling distribution (or, the standard error) tighten (and vice versa if we decrease the sample size). (If it does not look like it has tightened, check the values on the *X*-axis to see if they were adjusted to better present the data.) If we change the value of the mean, the entire distribution should shift to be centered on the new mean. If we increase the standard deviation, the resulting sampling distribution should widen (and vice versa if we decrease the standard deviation). By exploring this program (or the others previously mentioned), we should be able to see that $\mu_M = \mu$, $\sigma_M = \sigma/\sqrt{n}$, and all sampling distributions approximate a normal distribution.

7.4 Estimating the Features of Sampling Distributions

Only in rare situations will a researcher know the value of population parameters. However, population parameters are needed to determine μ_M and σ_M. Thankfully, in most cases, population parameters can be estimated from sample statistics. If μ is known, then μ_M can be computed with certainty ($\mu = \mu_M$). If μ is not known, then a sample mean (M) can be used as an unbiased estimate of μ. ("Unbiased" means that the sample mean is just as likely to be larger than the population mean, as it is to be smaller.) Of course, if σ is known, then σ_M can also be computed with certainty. However, when σ is unknown, σ must be estimated by using s (a sample standard deviation), an unbiased estimate of σ. Therefore, s_M becomes an estimate of σ_M. Formula 7.3 is the standard error of the mean estimated from a single sample.

Estimated standard error of the mean

$$s_M = \frac{s}{\sqrt{n}} \quad \text{(Formula 7.3)}$$

where

s_M = the estimated standard error of the mean
s = the sample standard error
n = the sample size

Formula 7.3 allows us to estimate the amount of variability within a sampling distribution. As an estimate of σ_M, s_M becomes more reliable as the sample size increases.

Table 7.1 clarifies the various symbols and estimates that are characteristics of sampling distributions.

Table 7.1 Important symbols of sampling distributions and population estimates.

1) M is an estimate of μ
2) s is an estimate of σ
3) μ_M is the same value as μ; μ_M is *not* an estimate of μ
4) σ_M is the standard deviation of a sampling distribution, called the standard error of the mean; it is not an estimate of anything
5) s_M is an estimate of σ_M

Sampling Distributions and Sampling Error

The notion of sampling error is fundamental in statistical inference. When discussing the mean in Chapter 3, an error was defined as the distance a score is from the mean of the distribution, $X - M$. The term error may be confusing because it usually implies a mistake; the fact that a raw score can be different from the mean should not be understood as a mistake. However, the term error has real meaning in relationship to a sampling distribution. Within the sampling distribution of means, the only value that is the same as the mean of the population is μ_M. Only those sample means that are equal to μ_M are perfect estimates of μ. Any sample mean that is different from the mean of the sampling distribution is in error in the sense that it is an inaccurate estimate of μ. A **sampling error** is the difference between a population parameter and the estimate of that parameter provided by a statistic. Thus, when estimating μ, a sampling error is $M - \mu$.

When working with a population distribution of raw scores, σ is the overall measure of error that is typically used. With a sampling distribution, the overall measure of error is σ_M. This makes perfect sense. Figure 7.2 shows two sampling distributions that have the same mean. Figure 7.2a has a smaller standard error, and as a result, the distribution is narrower than the sampling

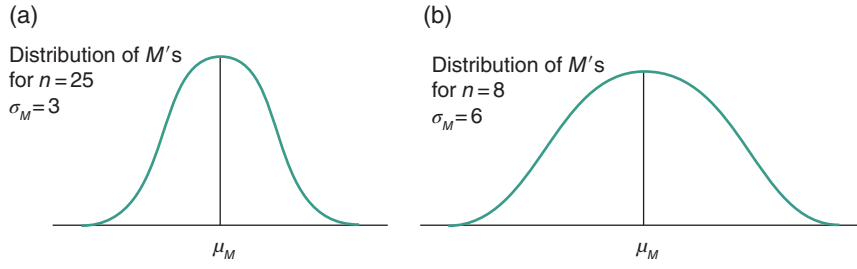


Figure 7.2 The sampling distribution in (a) has a smaller standard error than the sampling distribution in (b). The amount of sampling error is greater as σ_M increases.

distribution in Figure 7.2b, where there are more sample means at a greater distance from μ_M. Suppose we select a mean at random from each distribution in Figure 7.2. With an error defined as $M - \mu$, from which mean would we expect the larger error? Since the distribution in Figure 7.2b has more means farther away from μ_M, we are more likely to select a sample mean from that distribution that is far from the actual mean of the distribution. In other words, when comparing sampling distributions that have different standard errors, a randomly selected mean from a broader distribution will likely be a poorer estimate of μ than a randomly selected mean from a more narrow distribution.

At this point, there is no way for us to appreciate fully the importance of sampling distributions. However, it is impossible to understand the statistical foundation of hypothesis testing if we do not understand the characteristics and logic of sampling distributions. In large measure, the difference between being a statistical "number cruncher" and truly *understanding* inferential statistics depends on our ability to grasp the sampling distribution concept. The better we understand this chapter, the easier time we will have mastering much of the material presented in subsequent chapters. If there is confusion, please reread the necessary sections of the chapter to improve conceptual clarity.

Summary

The field of inferential statistics, based on probability theory and logic, is used to make inferences about the characteristics of a population from the characteristics of a random sample drawn from the population. A random sample is a sample of scores taken from a population in such a way that each score of the population has an equal chance of being included in the sample. Researchers use random sampling to obtain samples that are representative of populations. Random sampling forms the basis on which one can generalize from samples to populations.

The field of inferential statistics includes estimation and hypothesis testing. Parameter estimation uses data from a sample to infer the value of a population parameter. There are two kinds of estimation: point estimation and interval estimation or confidence intervals. Point estimation entails estimating a parameter as a single value. Interval estimation establishes a range of values within which the parameter is expected to lie. The second and most common type of inferential procedure is hypothesis testing.

Several characteristics of hypothesis testing were discussed:

1) A scientific research hypothesis is a formal statement or expectation about the outcome of a study. A research hypothesis precedes the collection of data and usually predicts a difference, effect, or relationship.

2) A statistical hypothesis is a numerical statement regarding the outcome of a study. Statistical hypotheses come in logically related pairs, the null and the alternative. They are mutually exclusive and collectively exhaustive.
3) Hypothesis testing uses samples to make inferences about populations. Statistical methods are used to make these inferences.
4) The null hypothesis states there *is no* effect of the independent variable on the dependent variable or, in correlational designs, *no* relationship between two variables. The alternative hypothesis states that there *is* an experimental effect, difference, or relationship, as the case may be.
5) Hypothesis testing results in a decision to either reject the null hypothesis or fail to reject the null hypothesis. Since any decision is probabilistic, there is always a risk of committing a decision error.

Single-sample research uses one sample to test a hypothesis about the mean of a population. Two-sample research uses two samples to test a hypothesis about the difference between two population means. When two samples are used, the investigator has an opportunity to establish the comparative effectiveness of two treatments. Many research designs are more complex, involving more than two samples. The correlational approach is an alternative research method that differs from experimentation in that it does not attempt to exert an influence on a measured response. Because variables are not controlled, the correlational approach cannot identify causal relations among variables. The correlational approach to hypothesis testing examines whether or not two variables are related.

A sampling distribution is a distribution of some sample statistic, usually the mean. It is the conceptual and mathematical cornerstone of hypothesis testing. Six facts about the sampling distribution of means were presented:

1) A sampling distribution of means is a theoretical distribution derived by computing the means of an almost infinite number of samples of size n.
2) If the scores of a population are normally distributed, irrespective of sample size, the sampling distribution of means will be normally distributed.
3) If the population distribution is not normally distributed, the sampling distribution approximates a standard normal curve as the sample size increases. The more the population distribution deviates from normality, the larger the sample size must be to establish a normally distributed sampling distribution.
4) As n increases, the shape of the resulting distribution approaches the standard normal distribution, as displayed in the z table (Table A.1)
5) The mean of the sampling distribution equals the mean of the population of raw scores, $\mu = \mu_M$.
6) The standard deviation of the sampling distribution is called the standard error of the mean. The relationship between the standard error of the mean and the standard deviation of the population is $\sigma_M = \sigma/\sqrt{n}$.

Key Formulas

Definitional formula for the standard deviation of a sampling distribution

$$\sigma_M = \sqrt{\frac{\sum (M - \mu_M)^2}{N_M}} \quad \text{(Formula 7.1)}$$

Standard error of the mean

$$\sigma_M = \frac{\sigma}{\sqrt{n}} \quad \text{(Formula 7.2)}$$

Estimated standard error of the mean

$$s_M = \frac{s}{\sqrt{n}} \quad \text{(Formula 7.3)}$$

Key Terms

Random sample
Point estimation
Interval estimation
Research hypothesis
Hypothesis testing
Statistical hypothesis
Null hypothesis
Alternative hypothesis
Sampling distribution
Central limit theorem
Standard error of the mean
Sampling error

Questions and Exercises

Most of these questions are conceptual, requiring no computations. Try to answer the conceptual questions without referring to the text.

1 How would we go about constructing a theoretical sampling distribution of means?

2 How does the population standard deviation influence the variability of the sampling distribution? What happens if it increases? What happens if it decreases?

3 What is the difference between point and interval estimation?

4 How does hypothesis testing differ from estimation?

5 How is the variability of a sampling distribution affected by the sample size?

6 What is meant by single-sample versus two-sample research designs? How is correlational research different from these designs?

7 What is the difference between a research and statistical hypothesis?

8 Define null and alternative hypotheses.

9 Give examples of null and alternative hypotheses for single-sample and two-sample research projects.

10 A normally distributed population of scores has $\mu = 100$ and $\sigma = 10$. A sampling distribution is established with $n = 9$. Describe the sampling distribution in terms of μ_M, its standard error, and shape.

11 Suppose a research hypothesis predicts that Drug *A* will generate higher performance numbers than Drug *B*.
- **a** State the null and alternative statistical hypotheses
- **b** What finding would result in a failure to reject the null?
- **c** How would a finding to fail to reject the null be interpreted in terms of the research hypothesis?
- **d** What finding would support the research hypothesis?
- **e** What finding would allow the researcher to reject the null but would not allow the researcher to support the research hypothesis?

12 Imagine it is known that American teenagers spend, on average, three hours a day on social media. Further, imagine that a community wanted to change this and took intentional steps to create numerous activities for the local teenager population that did not involve social media. Further, suppose that a researcher wanted to test the effectiveness of this community's programs by sampling the local teenager population and asking them about their social media involvement.
- **a** What would the null and alternative hypotheses look like?
- **b** What would be a finding that would result in failing to reject the null?
- **c** What would be a finding that would support the objectives of the community organizers?
- **d** What would be a finding that would reject the null but not support the objectives of the community organizers?

13 What is the relationship between a sample mean value and the standard error?

14 Suppose we know an anxiety measure has a mean of 50 with a standard deviation of 10; $\mu = 50$; $\sigma = 10$. What is the standard error if we create a sampling distribution with $n = 20$?

15 Suppose we know the average university student sleeps 6.5 hours a night during the school week with a standard deviation of 0.5 hours. What is the standard error if we create a sampling distribution with $n = 100$?

16 Suppose we know the average four-person family household generates 20 lb of garbage a week with a standard deviation of 4 lb. Suppose further that we wish to generate a sampling distribution based on a sample of 50 households. What is the mean of the population of means generated?

17 Suppose we sample the local squirrel population by trapping and releasing. Suppose further that we find our sample generates a mean weight for the squirrels of 17 ounces with a standard deviation of 2 ounces.
- **a** What is our best estimate of the population weight and standard deviation?
- **b** What is our best estimate of the standard error for a sampling distribution if we gathered 20 squirrels?

18 Suppose a local promoter, wanting to create a unique selling feature for their community, decides to try to create larger squirrels by creating and spreading genetically modified nuts throughout the community that have been supplemented with a growth hormone. Using data from Problem 17:
- **a** What would be the research hypothesis?
- **b** What would be the null hypothesis?
- **c** What would be the alternative hypothesis?
- **d** What statistical finding would support the research hypothesis?
- **e** What statistical finding would run counter to the research hypothesis?
- **f** What would be a finding that would result in failing to reject the null?

Computer Work

19 Consider the following data set as a population of scores. Compute μ and σ. Take a random sample of 5, 10, 15, and 20 scores. Compute the sample means and the standard errors using σ. Note how the sample size influences the variability of the sampling distribution that would be derived from this population. The overall measure of sampling error is the standard error of the mean. As the sample size becomes larger, sampling error becomes smaller.

A population of scores

22	**11**	**7**	**9**	**9**	**8**	**7**	**23**	**45**	**9**
23	21	8	8	5	16	9	22	17	6
12	29	6	5	9	23	7	33	24	5
15	14	9	7	3	17	8	19	15	8
11	10	8	6	1	11	9	25	35	9
10	18	6	5	4	13	8	21	20	9
14	17	5	6	2	34	2	35	35	5
35	36	8	1	3	37	1	32	33	4
33	29	7	7	6	28	9	27	26	9
27	28	5	5	4	27	8	26	25	7

Part 4

Inferential Statistics

z Test, *t* Tests, and Power Analysis

8

Testing a Single Mean: The Single-Sample *z* and *t* Tests

8.1 The Research Context

This chapter addresses the statistical analyses used in single-sample research projects. Recall that single-sample studies use one sample of participants to make an inference about whether the mean of the population is some specified value. A typical statistical hypothesis might read, "Does the obtained sample come from a population with a mean of [insert specific mean value] or from a population with a different mean?" The single-sample methodology for testing a hypothesis about a specified population mean has its place in the social and behavioral sciences. However, it is one of the least used research methods. It is certainly less commonly used than the comparison of two or more conditions or the type of study that attempts to discover the correlation between two or more variables.

Nonetheless, the relative simplicity of the method makes it ideal for the purposes of learning how to run inferential tests. Once we gain familiarity with these concepts in the context of a simple research design, we will be able to more easily understand the statistical inner workings of hypothesis testing in more complex designs – designs that are, admittedly, more commonly used in the social and behavioral sciences.

► **Example 8.1** Assume that the mean weight of newborn babies is 7 lb, with a standard deviation of 1 lb. A study is conducted with the following research question: "Do newborn babies of mothers who drink alcohol during pregnancy weigh less than the average baby?" To answer this question, the investigator would take a single random sample of newborns from mothers who consumed alcohol during pregnancy, compute the mean, and, using the methods discussed in this chapter, come to a decision about whether this sample of newborns came from a population with a mean of 7 lb. ◄

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

► **Example 8.2** Suppose a university administrator wants to know if the students coming into the institution have strong mathematical abilities. Assume the national mean of the quantitative section of the Scholastic Aptitude Test (SAT) Math is 500, with a standard deviation of 100. The administrator takes a random sample of SAT Math scores from the incoming freshman class, computes the mean, and decides whether the population of incoming students has a mean SAT Math score that is different than 500.◄

► **Example 8.3** The chairperson of a psychology graduate program is told that the average time it takes a graduate student to earn the PhD is 6.8 years, with a standard deviation of 1.2 years. The professor wonders how the students in this program compare. A random sample of 32 recent graduates is examined and found to have a mean of 5.2 years. Can the chairperson conclude that students in this program average less than 6.8 years?◄

The statistical test that is used to decide if a sample mean does or does not come from a specified population, *when the standard deviation of the population is known,* is called a single-sample ***z* test**. Since only one sample of scores is taken, the *z* test applied in this situation is a single-sample or one-sample statistical test. Furthermore, since the standard deviation of the population is known, *we can determine the actual standard error*; we do not need to estimate it using a sample standard deviation. If we did not know the population standard deviation, this would still be a single-sample research project, but we would need to estimate it and run the single-sample ***t* test** instead. This chapter discusses both tests, but we will look at the *z* test first and then *t* tests. We will discover that the arithmetic computations for each test are rather simple. We will also discuss the role of sampling distributions, the reasons the formulas for the *z* test and *t* test are expressed as they are, and the implications of shifting from a *z* test to a *t* test.

8.2 Using the Sampling Distribution of Means for the Single-Sample *z* Test

This section begins where Chapter 7 ends: the importance of sampling distributions in hypothesis testing. To recall, the following list highlights the facts about the sampling distribution of means:

1) A sampling distribution of means is a theoretical distribution derived by computing the means of an almost infinite number of samples of size *n*.
2) If the scores of a population are normally distributed, irrespective of sample size, the sampling distribution of means will be normally distributed.

3) If the population distribution is not normally distributed, the sampling distribution approximates a standard normal curve as the sample size increases. The more the population distribution deviates from normality, the larger the sample size must be to establish a normally distributed sampling distribution.
4) As n increases, the shape of the resulting distribution approaches the standard normal distribution, as displayed in the z table (Table A.1).
5) The mean of the sampling distribution equals the mean of the population of raw scores, $\mu = \mu_M$.
6) The standard deviation of the sampling distribution is called the standard error of the mean. The relationship between the standard error of the mean and the standard deviation of the population is $\sigma_M = \sigma/\sqrt{n}$.

z Scores and the Sampling Distribution of Means

The z score formula for transforming a raw score to a z score was given in Chapter 5 as

$$z = \frac{X - \mu}{\sigma}$$

A z score specifies how many standard deviations the transformed raw score is from the mean of the distribution. If every score of a population that is *normally distributed* is transformed into a z score, the result is the standard normal curve; this curve has a mean of 0 and a standard deviation of 1 (it is described by Table A.1 in the Appendix). The standard normal curve was used in Chapter 5 to solve all of the z score problems. For example, it is possible to use this table to make statements about the probability of selecting a score at random from some area under the normal curve. In every one of the z score problems we worked in Chapter 5, the focus was on *scores*. For example, a typical question was, "Given $\mu = 50$, and $\sigma = 5$, what is the probability of drawing a *score* at random above 60 or below 40?" The strategy we used to answer this type of question involved transforming the numbers 60 and 40 into z scores and then using the z table to identify the area of the curve above the z score for 60 and below the z score for 40.

In inferential statistics, a sample statistic (e.g. the mean) is used to infer a population parameter. It is meaningless to take a *score from a sample* and attempt to *infer a score from the population*. When testing a hypothesis about the value of a population mean, a sample mean is used to decide whether the population has a stated value. The decision is based on the probability of finding a sample mean of a certain value, *given a hypothesized value of the population mean*. Obviously, the z score formula cannot be used to arrive at a probability statement regarding the likelihood of drawing a *sample mean* of a certain value; *scores are* not *means*.

However, with a slight adjustment in the formula, and by employing the sampling distribution concept presented in Chapter 7, we can make probability statements about selecting a sample mean from an area under the curve of a sampling distribution.

The *z* Statistic

The z score formula transforms raw scores within a population into z scores. The formula for the z statistic transforms the means within a sampling distribution into z scores. Formula 8.1 is used for this transformation.

z Statistic

$$z_{obt} = \frac{M - \mu}{\sigma_M} \quad \text{(Formula 8.1)}$$

where

M = the mean of the sample
μ = *the* hypothesized mean
σ_M = the standard error of the mean

The z_{obt} symbol is used to indicate that the z value is obtained from a mean. The z statistic has the same basic form as the z score formula. Table 8.1 contrasts each value of the two formulas. Technically, the population mean of the z score formula is replaced by the mean of the sampling distribution. Since the mean of the sampling distribution, μ_M, is the same as the population mean from which the sampling distribution is established, the formula for the z statistic simply uses the symbol μ in place of μ_M. The sample mean M replaces the X score. In the denominator, instead of using the standard deviation of the population of raw scores, the z statistic uses the standard deviation of the sampling distribution, also called the standard error, σ_M.

If all the raw *scores* of a *normally distributed* population are transformed into z scores, the standard normal distribution is generated, and the z table

Table 8.1 A comparison of the z score formula used to transform raw scores with the z statistic used to transform a sampling distribution of means.

***z* Score formula**	***z* Statistic formula**
$z = \frac{X - \mu}{\sigma}$	$z_{obt} = \frac{M - \mu}{\sigma_M}$
X = a single score	M = a single mean
μ = the population mean	μ = the hypothesized population mean
σ = population standard deviation	σ_M = standard error of the sampling distribution

(Table A.1) can be used to identify areas under the curve and to make probability statements. When a sampling distribution of means is *normally distributed,* transforming all the means to *z* values (using Formula 8.1) produces the same standard normal distribution. The *z* table can now be used to make probability statements about *means*.

The Logic of the *z* Test

To walk through the logic of the *z* text, we can use a hypothetical study. Suppose an educational enrichment program is developed to teach math skills. The program entails the use of an interactive web-based instruction to teach basic algebra skills. A *random sample* of 36 students is selected, given the interactive web-based instruction, and tested at the end of the school year. The test has been extensively validated on a large population and is known to have a mean of 100 and a standard deviation of 10. (That is, suppose we know both μ and σ.) A statistical test, based on a single sample, will be conducted to test the hypothesis that the population mean is 100. Therefore,

$$H_0: \mu = 100$$
$$H_1: \mu \neq 100$$

We are *not* testing the hypothesis that the mean of the national exam is 100. We have already been told that. The population we are intending to make an inference about is a hypothetical population, the population of all students that *theoretically could* receive the training. When the training of the sample is completed, we must assume that this sample of participants is representative of the population of students who *could be* exposed to this training. The population of interest is *hypothetical*; it is *imaginary*. It is important for us to gain comfort with the idea of getting real samples but understanding that they may represent imaginary populations. In this situation, our sample is a subset of all these hypothetical students.

The null and alternative hypotheses in the example are statements that derive from the question, "Is the mean of the population from which the sample is taken 100?" In essence, we are asking if the treatment has any effect on math skills. If the treatment has an effect, we would expect to obtain a sample mean that is different from the hypothesized population mean of 100.

Sampling Error and Hypothesis Testing

Suppose the null hypothesis is true; that is, suppose the math enrichment program has no effect, and the population mean of students who could experience the math enrichment program is still 100. When we gather the data from our sample and compute the mean, would we expect that sample mean to be *exactly* 100?

Due to sampling error, even if the null hypothesis were true, it is extremely unlikely that the sample mean would be exactly 100. Recall that the sampling error of a mean is the distance the mean of a sample is from the mean of the population. Imagine taking repeated samples from a population, calculating the mean, and replacing them (in other words, creating a sampling distribution of means). Each mean is based on scores dispersed throughout the population. For the sample mean in question, we might randomly draw, for example, several extreme scores from the right tail of the population. The resulting sample mean, therefore, will be larger than the mean of the population. Even though the null hypothesis is true and there is no treatment effect, the random selection of scores is subject to sampling error.

Returning to the educational enrichment program, we have to decide if there is a treatment effect; that is, we have to make a decision about whether to reject the null hypothesis that $\mu = 100$. Suppose the sample mean drawn is 100.5. Are we willing to conclude that $\mu \neq 100$? Probably not. Suppose the sample mean drawn is 102. What about a sample mean of 105, 115, or 130? As we answer these questions, we are operating with some intuitive belief about the likelihood that sampling error can explain the distance between the value of the obtained sample mean and the hypothesized population mean. We might assume that a sample mean of 102 could easily occur when the population mean is actually 100, but we may also conclude that a sample mean of 130 would not be likely to occur if the population mean were 100. Thankfully, the statistical procedures involved in hypothesis testing greatly reduce this guesswork. When we conduct a z test, the sample mean is transformed into a z statistic, and this allows us to use the z table to judge the probability of obtaining that sample mean when the null hypothesis is true. If the z statistic is far away from 0, it suggests it is *very unlikely* that sampling error alone would have yielded that value. We would then conclude that the population mean from which our sample is drawn is not what is specified in the null hypothesis, and the null hypothesis would be rejected. Please reread the last three sentences; *they are the most important in the entire chapter*.

Sampling Distributions When the Null Hypothesis Is True or False

Figure 8.1 depicts two sampling distributions of a mathematics enrichment program. The sampling distribution on the left shows the mean of the population as 100. The overlapping sampling distribution on the right is from a population with a mean of 104. As we continue reading, keep the following in mind. The goal of hypothesis testing is *not* to infer the value of the mean for the hypothesized population. If we want to make an inference of that sort, use a point or interval estimation. The goal of hypothesis testing is to decide whether to reject the null hypothesis. In Figure 8.1, the sampling distribution on the left would be true if the null hypothesis is true ($\mu = 100$). In other words, if the

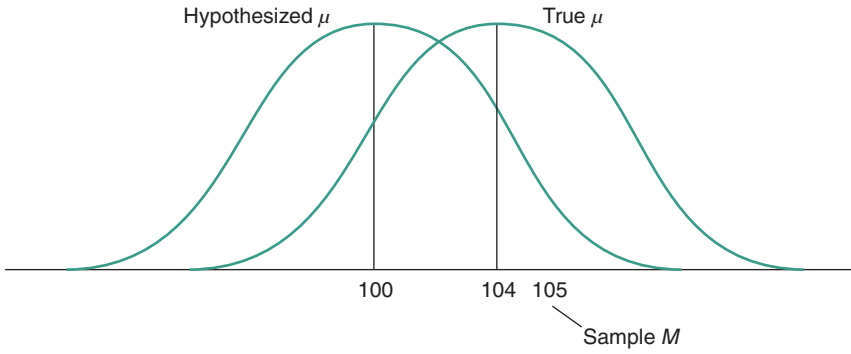


Figure 8.1 The sampling distribution on the left is from a population in which $\mu = 100$. The sampling distribution on the right is from a population in which $\mu = 104$; this represents the *true* state of affairs. A sample mean of 105 falls in the tail of the left distribution, but closer to the middle of the distribution on the right.

mathematics enrichment program has no effect, we would expect the mean of the sampling distribution to be 100. If the null hypothesis is false, then the mean of the sampling distribution is some number other than 100. In this figure the true mean of the population of those in the mathematics enrichment program is presented as 104; thus, the mean of the sampling distribution must be l04.

Now suppose we take a sample, compute the mean, and find it to be 105. If the true state of affairs is represented by the sampling distribution on the left, $\mu = 100$, then we have unintentionally and randomly oversampled from the right tail of the distribution. Since there are few means in the tail of the distribution, it is a statistically rare event to obtain a mean of 105 from a sampling distribution with a mean of 100. However, if the true state of affairs is represented by the sampling distribution on the right, $\mu = 104$, then the sample mean of 105 is not at all unusual. A sampling distribution with a mean of 104 will have many means right around the value of 104; obtaining one close to 104 is *not* a statistically rare event.

It is important to keep in mind when hypothesis testing that we do not know the mean of the population, and therefore we do not know the mean of the sampling distribution. Figure 8.1 gives us a behind-the-scenes glimpse, so to speak – a glimpse that we never have when performing actual research. With the aid of only our sample mean, we need to determine the probability of drawing that sample mean, by chance, from a distribution based on a null hypothesis that is true. If the probability is low (determining what is meant by "low" will soon be discussed), then the null hypothesis should be rejected. If we hypothesize $\mu = 100$, obtaining a sample mean of 105 *may* lead us to conclude that μ is probably not 100. As mentioned earlier, thankfully we have tools that can remove much of the guesswork. Specifically, the z test and z table are used to arrive at a decision about whether or not to reject the null hypothesis.

Using the *z* Test in Deciding to Reject the Null Hypothesis

Consider again the task of evaluating the effectiveness of the educational enrichment program. When testing the null hypothesis, we proceed *as if the* null hypothesis were true. (This is such an important point; the mathematics of hypothesis testing are set up *as if* the null is true. It is the starting position.) Since students are tested with an examination previously known (i.e. outside of this enrichment program) to have a population mean of 100 and a standard deviation of 10, the null and alternative hypotheses are stated as

$$H_0: \mu = 100$$

$$H_1: \mu \neq 100$$

A sample of 36 students is tested after completion of the program and the sample mean is found to be 105. After the null and alternative hypotheses have been stated, the next step is to transform the sample mean into a *z* statistic, using Formula 8.1:

$$z_{obt} = \frac{M - \mu}{\sigma_M}$$

$$z_{obt} = \frac{105 - 100}{10/\sqrt{36}}$$

$$= \frac{5}{1.67}$$

$$z_{obt} = \mathbf{2.99}$$

How would we interpret a z_{obt} of 2.99? If we assume that the null hypothesis is true, $\mu = 100$, then our sample mean of 105 is 2.99 standard error units above the mean of the sampling distribution. (Recall that standard error units are analogous to standard deviation units; thus, about 68% of all *z* scores will fall within ±1 standard error of the mean, about 95% within ±2 standard errors from the mean, and about 99.7 within ±3 standard errors from the mean. See the 68-95-99.7 rule presented in Chapter 4.)

Figure 8.2 shows where a sample mean of 105 lies in a sampling distribution having a mean of 100 and a standard error of 1.67. Note the *z* values that correspond to the means in the sampling distribution. Since Formula 8.1 transforms the sampling distribution to a standard normal curve with $\mu = 0$ and $\sigma = 1$, the *z* value at the mean of the sampling distribution is 0. Since the standard error of the sampling distribution is 1.67, the means of 101.67 (100 + 1.67) and 98.33 (100 – 1.67) have corresponding *z* values that are +1 and –1, respectively. The sample mean of 105 is in the right tail. If the mean of the sampling distribution is 100, what percentage of means is found at or above a *z* of 2.99? Refer to the third column in the *z* table (Table A.1). Only 0.14% of the means of the sampling distribution are found at or above a mean of 105 (a *z* of 2.99).

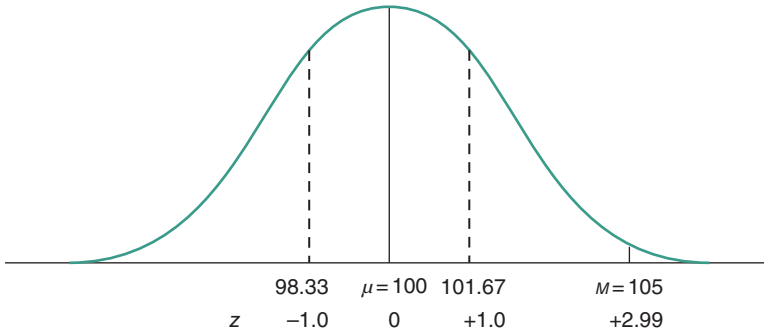


Figure 8.2 In this sampling distribution $\mu = 100$ and $\sigma_M = 1.67$. A sample mean of 105 is 2.99 standard error units above the mean.

Now there is a decision to make. We can assume that the mean of the sampling distribution (the mean of the population) is 100 and that we have experienced the highly unlikely event of oversampling from the extreme right tail of the distribution. On the other hand, we can reject the assumption that the null hypothesis is true; that is, reject the claim that our sample mean came from a population with a mean of 100. In this situation, most would agree that the null hypothesis is probably false and it should be rejected.

If it were left up to each individual researcher to decide what is and is not a rare statistical occurrence, hypothesis testing would lead to endless confusions. For this reason, in most research situations social and behavioral scientists follow a long-standing convention of using a probability value of .05 as the criterion for rejecting null hypotheses.

The Criterion for Statistical Significance: Acceptance and Rejection Regions

Over the years, statisticians have come to define the criterion for rejecting the null to be .05 (for a historical overview see Cowles and Davis, 1982). As applied to the single-sample *z* test, this means that if the probability of randomly drawing a given sample mean is less than .05, when the population mean is hypothesized to be a specified value, then we should reject the hypothesis that claims the population mean is that specified value. In the *z* distribution, what are the cutoffs beyond which lie the outermost 5% of the distribution? Stated differently, what are the *z* values that mark both the lowest 2.5% of the distribution *and* the highest 2.5% of the distribution? (Combined, these areas make up the outermost 5%.) Assuming the data is normally distributed, the *z* values of ±1.96 mark these extremes. Therefore, the probability of selecting a mean at random that will correspond to a *z* value that is equal to or falls outside of the absolute value of 1.96 is .05.

This means if the z_{obt} shows the sample mean to be more than 1.96 standard error units away from the *assumed* population mean in either direction, then we should reject the null hypothesis.

Alpha Levels and Rejection Regions

When we use the cutoffs of ±1.96, we are testing the null hypothesis at the 5% level of significance. The probability value of .05 is called the **alpha level,** symbolized as α. In hypothesis testing, the investigator chooses the alpha level. Other conventional alpha levels are .10 and .01. *Before* the data are analyzed, the researcher specifies the alpha level at which the null hypothesis is to be tested. In essence, when setting the alpha level, the researcher is defining the criterion for what will be considered a statistically rare event. An alpha value of .10 would be considered a permissive or liberal criterion, while an alpha value of .01 would be considered a conservative criterion.

As soon as an alpha level is selected, *rejection* and *fail to* reject regions are automatically determined. Figure 8.3 shows the rejection and fail to reject regions for alpha levels of .10, .05, and .01. The rejection region is always in the tails of the sampling distribution, that is, the farthest away from the mean. The absolute values of the z scores that define the rejection regions are called the **critical values.** Since 10% of the distribution falls beyond a z of ±1.65, the null hypothesis is rejected at the .10 level of significance if the absolute value of z_{obt} is equal to or falls outside of the critical values of ±1.65. If the alpha level is set at .05, the null hypothesis is rejected if the absolute value of z_{obt} is equal to or falls outside of ±1.96. Moreover, since only 1% of the sampling distribution corresponds to a z of ±2.58, the null hypothesis is rejected at an alpha of .01 if the absolute value of z_{obt} is equal to or falls outside of the critical values of ±2.58.

If alpha is set at .05 and the null hypothesis is rejected, then we would state $p < .05$, which means that the probability of obtaining a mean at or beyond the critical values *if the null hypothesis is true* is less than .05. The critical values of ±1.65, ±1.96, and ±2.58 can be verified by using the z table (Table A.1). Use the third column of the z table, and look up the alpha value divided by 2 (half in each tail of the distribution). Note that the rejection region defined by the critical values is in both tails of the distribution. The fact that the rejection region is in both tails reflects the investigator's willingness to reject the null hypothesis if the sample mean is either considerably above or considerably below the hypothesized population mean. This is known as a *nondirectional* hypothesis test or a *two-tailed* hypothesis test. In the educational enrichment program example, the z_{obt} was 2.99. Even if we set alpha at .01 prior to collecting the data, we would still be rejecting the null hypothesis because 2.99 is larger than 2.58 ($p < .01$). Now that the null hypothesis has been rejected, what does this mean about the educational enrichment program?

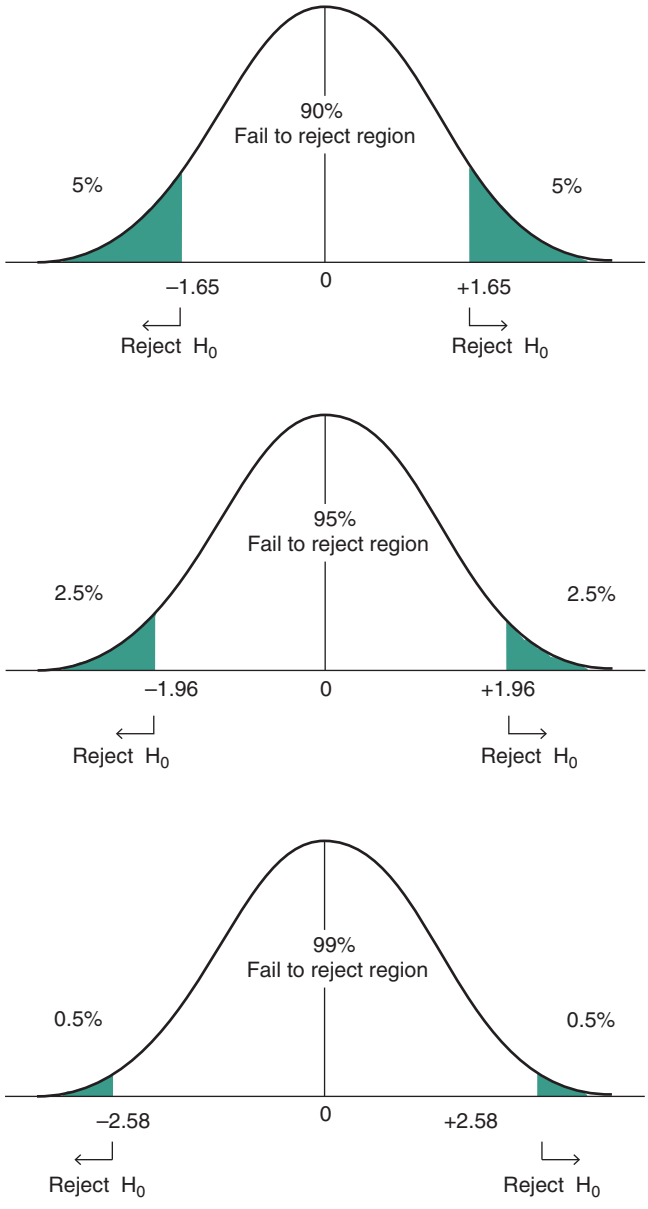


Figure 8.3 The acceptance and rejection regions for alpha levels of .10, .05, and .01.

The Implications of Rejecting the Null Hypothesis

What would we conclude if the null hypothesis were rejected? In an experiment, testing the statistical hypothesis is the process used to discover if there is a treatment effect. When the null hypothesis is rejected, it is concluded that the difference between the sample mean and the hypothesized population mean is probably not due to sampling error (chance). What caused the difference? The difference between the hypothesized population mean and the sample mean is presumably due to the experimental variable. For our example, it means the sample mean of students from the educational enrichment program was determined most likely to come from a population that had a mean that differed from 100. That is to say, given the outcome of the z test, it is reasonable to suggest that the sample of math scores came from a population with a mean that is something other (in fact, higher) than 100. Generally speaking, in the context of experimentation, any time the null hypothesis is rejected, a researcher would like to conclude that the rejection is due to the influence of the independent variable. However, we cannot automatically assume that the independent variable is responsible for the observed difference between means. In addition to sampling error, it is also possible that the experimental effect was due to a variable not controlled by the experimenter. Recall from Chapter 1 that an experimental confound presents an interpretive dilemma for the investigator since alternative explanations can be offered to explain the finding. After rejecting a null hypothesis, interpretation involves a careful analysis of the quality of the research design; a poorly designed study cannot be saved by even the most sophisticated statistical techniques.

As stated in the previous chapter, failing to reject the null hypothesis is not to accept that the null hypothesis is probably correct. Even if the sample mean perfectly equaled the hypothesized population mean, we cannot conclude that the null is correct. For one reason, the null may be false, but only by a little. For instance, suppose the enrichment program works, but it only produces a slight benefit to the students; perhaps it results in a population mean of 101. In a situation like this, getting a sample mean of 100 might be quite common, and yet clearly it does not mean that the null hypothesis is true (we just stated that the null is false and the sample is coming from a population centered on 101). Failing to reject the null simply means that the null *may* be correct; statistical evidence suggesting a rejection of the null hypothesis was not found. Think of it as a scientific version of a shrug-of-the-shoulder. Perhaps the strongest type of claim to be made when a null hypothesis is not rejected is something like the following; if the null is false, it does not look to be radically false. However, even a cautious statement like this is dependent upon other factors, like the sample size being used. When a null hypothesis is not rejected, it is best merely to state that statistical evidence to reject the null was not found and to leave it at that.

8.3 Type I and Type II Errors

The null hypothesis is rejected when it is the most reasonable conclusion given the relationship between the observed sample mean and the null hypothesized population mean. However, because inferential statistics are based on probabilities, there is always the chance of making a decision error. There are two kinds of decision errors that can be made when testing null hypotheses, rejecting the null when it should not be rejected (a false positive) and not rejecting the null when it should be rejected (a false negative).

Imagine being a member of a jury that must decide the guilt or innocence of a defendant. Independent of what is decided, the defendant is either truly guilty or truly innocent. In this situation, there are four possible outcomes. First, if the defendant is truly guilty and we decide they are guilty, we are correct. Second, if the defendant is truly innocent and we decide they are innocent, we are correct again. Third, if the defendant is truly guilty and we vote innocent, we have committed an error. Finally, if the person is truly innocent and we say guilty, another type of error is committed. Although the legal criterion for deciding guilt is "guilty beyond a reasonable doubt," the criterion we use to define "reasonable doubt" may change depending on the consequences of each kind of error. For example, if we know that the defendant will receive the death penalty if found guilty, we will want to be *very* sure of our decision when we vote guilty. Under this circumstance, we may require more evidence of the defendant's guilt before we are willing to decide guilty.

As our cutoff for defining reasonable doubt becomes more stringent, it reflects the fact that we are protecting ourselves from making a certain kind of error: the error of calling an innocent person guilty. However, by making the criterion more stringent, we increase the likelihood of the opposite error: calling a guilty person innocent. There is a trade-off between the two types of errors. Two points are important to remember. First, whether we are deciding on the guilt or innocence of a defendant, or the status of the null hypothesis, there are two types of errors we can make. These situations are analogous. Second, the criterion we use to make a decision is influenced by the relative consequences we decide are associated with committing each type of error.

In hypothesis testing, one error is called a **Type I error**. This error is committed if a true null hypothesis is rejected. The other type of error is called a **Type II error**. This error is committed when a false null hypothesis is not rejected. Table 8.2 specifies the four possible outcomes when deciding to reject or fail to reject the null hypothesis under the conditions that the null hypothesis is actually true or false. The problem is that we never know when we have made an error because we never know the real status of the null hypothesis.

In hypothesis testing, the investigator is in a better position than the juror because the investigator is able to control directly the probability of a Type I error. The alpha level set ahead of time by the researcher specifies the

Table 8.2 The decision grid depicting the four possible outcomes of hypothesis testing.

Our decision	True state of affairs	
	H_0 is true	H_0 is false
Fail to reject H_0	Correct	Type II error
Reject H_0	Type I error	Correct

probability of a Type I error. Suppose alpha is set at .05. *Given a true null hypothesis,* if the study were conducted numerous times, the null hypothesis would be mistakenly rejected about 5% of the time. Therefore, when the null hypothesis is true, the z_{obt} will rarely fall outside the critical values of ±1.96, and the null hypothesis will be correctly not rejected approximately 95% of the time. Likewise, if alpha is set at .01 and the null hypothesis is true, we can expect to draw a sample mean from the rejection region only about 1% of the time. A decision error, however, is still possible.

At first blush, this might suggest that we should make alpha as stringent as possible. Why not set alpha at .0001? We would only commit a Type I error 1 out of 10 000 times! However, remember the trade-off for the juror: If we are too afraid to convict an innocent person, we will make it easier for a guilty defendant to be set free. If alpha is made more stringent, it is true that the probability of a Type I error decreases, but the probability of a Type II error increases. Therefore, as we reduce the risk of rejecting a true null hypothesis, the risk of failing to reject a false null hypothesis increases.

In this way, adjustments to the Type I error rate reciprocally influence the likelihood of making a Type II error. The more permissive the Type I error rate selected, the less likely a Type II error will be made and vice versa. However, the relationship between the types of errors is a peculiar one. Since the Type I error rate is established from a known null distribution, actions designed to increase or decrease the type II error rate do not influence the risk of making a Type I error. The likelihood of making a Type I error is simply whatever alpha value is decided upon – end of story. It is a strange world in which we live!

A final word of caution: Do not make the mistake of thinking that if the probability of a Type I error is .05, the probability of making a Type II error must be .95. The reciprocal relationship between the two types of errors does not mean they are complementary. Calculating the probability of making a Type II error is not straightforward. Actions the researcher can take to influence the Type II error rate will be discussed in Chapter 11.

Deciding on Alpha

How do researchers decide on an alpha level? Just like jurors, they weigh the consequences. Suppose a researcher wants to investigate a new research area. There may be interesting effects to find there, but no one knows. Failing to reject a null hypothesis gets us nowhere. It is sort of like looking into a room for a missing item but with the lights off. We are not sure we have learned anything from the exercise. Researchers can become so disappointed after failing to reject the null they may decide against performing further research in the area. For this reason, in new research areas, the investigator may relax the alpha value and test at the 10% level of significance. It is easier to reject the null with an alpha level of .10 since the critical values are smaller (±1.65) than the critical values at the 5% alpha level (±1.96). Once a finding is identified, subsequent studies can use the more standard alpha level of .05. A permissive alpha level reflects an investigator's belief that a Type II error (failing to reject a false null) is more problematic than a Type I error (rejecting a true null) when initially exploring a new area of research.

There are instances in which the consequences of making a Type I error are so serious that a researcher would want to set a very stringent alpha level. Consider a study testing for serious drug side effects. Deciding that the medicine is safe, when, in fact, it produces side effects in a large number of people, is a potentially life-threatening error. If the null hypothesis is cast in such a way that rejecting it means that the drug is medically safe, the researcher might make alpha very conservative (for example, .001). This practice reflects the researcher's belief that it is much more serious to conclude that the drug is safe when it is not than to mistakenly conclude that the drug is not safe when, in fact, it is.

Another important issue to address is the timing of setting alpha. To measure properly the probability of making a Type I error, the alpha level must be established *ahead of time* and then followed. It is highly inappropriate to first run a study and find the probability associated with getting a particular z value and *then* decide where to set alpha so that the null hypothesis can be rejected. This is a form of statistical cheating!

One final point about making a Type I error: Suppose a researcher conducts a study and fails to reject the null hypothesis. Would the investigator be justified in performing the exact same study repeatedly until the null is rejected? No, because eventually even a true null hypothesis will be mistakenly rejected (see Box 7.2 and Box 8.1). However, it is acceptable for researchers to theoretically reflect on failed studies, revise experimental procedures, and then run the study again. However, conducting the same failed study over and over again is not only a waste of time, it is also scientifically dubious.

Box 8.1 Is the Scientific Method Broken? Type I Errors and the Ioannidis Critique

If there is a "ground zero" for the current reproducibility crisis in the social, behavioral, and medical sciences, it may be found in the personhood of John Ioannidis, Professor of Medicine and of Health Research and Policy at the Stanford University School of Medicine. In 2005, he published an article in PLoS Medicine entitled "Why Most Published Research Findings are False" (Ioannidis, 2005). As one might imagine, this article created a firestorm of controversy as well as an avalanche of articles reacting to this claim – some supporting (e.g. Freedman, 2010) and some critiquing (e.g. Leek and Jager, 2017). As a result, Ioannidis is currently one of the most cited scientists in the world.

Several of the points Ioannidis makes in the paper involve misunderstanding the perils of Type I errors. The Type I error rate is an accurate reflection of the risk involved in rejecting a singular null hypotheses. However, the testing of a null hypothesis does not take place within a vacuum, and other factors must be taken into account. These other factors include (i) how many questions are being asked in a given research project, (ii) how many other similar projects may be taking place elsewhere by other researchers, and, most importantly, (iii) what is the ratio of null relationships to actual relationships existing in a given area of inquiry.

To help illuminate the argument, Ioannidis (Wilson, 2016) asks readers to suppose there are 101 stones in a given field. Only one of them, however, contains a diamond (i.e. a true finding). Gratefully, we have at our disposal a diamond-detecting machine that advertises a 99% accuracy of detection (i.e. hypothesis testing using inferential statistics). That is, when the machine is placed overtop a stone without a diamond in it, 99% of the time it will not light up. Only 1% of the time will it give us a false positive (or Type I error). Further, imagine that after checking several stones and getting no reaction, the machine finally starts to flash with activity. What is the probability that this stone, if cracked open, will contain a diamond? We might initially suggest that there is a 99% chance. However, recall that there are 100 dud stones in this field. The machine, if functioning at a 1% false positive rate, will register, on average, 1 false positive if all stones are checked. This means, of the 101 stones in the field, two are likely to register as positive for the diamond (one false positive and one real positive). Therefore, there is only a 50% chance of finding a diamond when this particular stone is cracked open. This is a little disappointing.

Now, imagine a field that has several thousand stones in it – still only one of them containing a diamond. Do we see how in this situation even a false positive rate of 1% may lead persistent researchers to draw faulty conclusions far too frequently? One key factor here, which is impossible to answer, is the ratio of stones containing diamonds. As this ratio increases, the ratio of true positives to false positives will improve. However, how do researchers know ahead of time

in what sort of "field" they are working? Herein lies a big problem, the unknown ratio of real to null findings in a given area of investigation. Knowing the detection equipment has a 1% false positive rate does not solve this problem.

Further, do we see how the repeated testing of several stones changes the meaning of the 1% false positive rate? If we were to walk up to a field of stones and just test one, then the false positive rate of 1% makes sense. However, as we test stone after stone, the probability that at least one of the dud stones will register as significant grows as we work our way across the field. Herein lies a second problem, the additive nature of the Type I error rate.

One way to combat these problems, in addition to valuing replication (see Box 7.2), is to publically report nonsignificant findings. Only once researchers get a sense of how few "diamonds" there are in a field of inquiry can they begin to process what a supposed finding might mean. If the field of inquiry seems to be chock-full of effects and relationships, then a significant claim seems more likely to be an actual finding, but if the field has repeatedly been shown to be lacking meaningful findings, then a claim of significance should be interpreted with a great deal of suspicion. Unfortunately, despite the current reproducibility crisis, there seems to be little interest in creating publication opportunities for null findings. Until this happens, the Type I error problem is going to continue to bring a cloud of suspicion around claims of findings, especially those coming from new, previously unexplored, fields of inquiry.

8.4 Is a Significant Finding "Significant?"

When the decision rule directs the researcher to reject the null hypothesis, oftentimes the word "significant" is used in the interpretation. For instance, "the effect of the drug was found to be statistically significant." Even though this phrasing is quite common in the professional literature, it is easily misunderstood.

In ordinary parlance, significant means important. However, it is possible to achieve statistical "significance" even when the research finding is quite trivial. Consider the educational enrichment program as an example. Instead of using a sample size of 36, suppose we had used a sample of 1000 students. Further, suppose that the sample mean turned out to be 101 instead of 105. Using Formula 8.1,

$$z_{obt} = \frac{M - \mu}{\sigma_M}$$

$$z_{obt} = \frac{101 - 100}{10/\sqrt{1000}}$$

$$= \frac{1}{0.32}$$

$$z_{obt} = \mathbf{3.13}$$

Even if we set a conservative alpha level of .01, which has a critical value of ±2.58, the null hypothesis is to be rejected. However, it may not be clear to everyone looking at the finding that all of the time, effort, and expense used to implement the program would be worth a mere one point increase on the national exam. In this case, although the phrase "significant" is accurate in the statistical sense, it may not be accurate in a practical sense.

The other side of the issue arises when the null hypothesis is not rejected; the statistical test is deemed *nonsignificant*. Note that nonsignificant is not the same as *not important*. A statistically nonsignificant finding may be considered important. For example, new treatments are advertised all the time: treatments for arthritis, obesity, stress, or whatever ails us. If we contrast a no-treatment control group with a group receiving an ineffective treatment, we will likely end up failing to reject the null hypothesis, a *nonsignificant* finding. Since it is impossible to prove the null hypothesis true, we cannot conclude that the treatment has been *proved* ineffective. Nevertheless, in the absence of evidence that the treatment *is* effective, there is no compelling reason to use it. Assuming the study was well designed, the treatment had an opportunity to show its effect and failed. This scenario exemplifies a statistically nonsignificant finding that may be considered rather important, practically speaking, by a person contemplating the use of a particular treatment.

Before leaving this topic, it is important to mention a couple more often-used phrases regarding statistical significance that can be problematic. Sometimes we may read that a finding is described as *marginally significant*. This phrase is often used when a finding just fails to fall inside the critical scores when the alpha level is set at .05. The z_{obt} is close to the rejection region (for example, getting a z_{obt} of 1.93), and in the mind of the researcher, a Type II error is being made if the null is not rejected. The use of this term does highlight the arbitrary nature of the conventional 5% alpha level. However, it is important to realize that the burden established prior to performing the research for rejecting the null was not met. It may be true that a Type II error is occurring, and it is fine and acceptable for the researcher to bring this to the attention of others, but this possibility should not cause us to reclassify a "fail to reject" finding. The confusion around the interpretation of marginally significant findings constitutes another reason for valuing replication studies.

Finally, the phrase *highly significant* can be frequently found in the professional literature of the social and behavioral sciences. It is often used when the alpha value was initially set at .05, but the findings would have allowed for a rejection of the null even if the alpha value had been set at .01 or even .001. This phrasing implies that as the p value decreases, a researcher should be more impressed with the strength of the effect produced by the treatment (Bakan, 1966; Cohen, 1990). This phrase can be misleading. For one thing,

the *p* value reflects the degree of *certainty* for an effect, not necessarily the *size* of an effect. Again, think of the enrichment program example where a sample mean of just 1 point still allowed for the rejection of the null hypothesis and at $p < .01$. This finding could be described as "highly" significant. However, this description is misleading. It is true that the concepts of *certainty* and *size* are not completely independent of each other; after all, a very powerful effect is more likely to yield an extreme mean. However, factors other than effect size can influence our sense of certainty that an effect exists. (This topic will be discussed further in Chapter 11.) It is also simply inappropriate to redraw the lines of meaning *after* the findings have been analyzed. It may be true that a shot arrow did not only hit the bull's-eye but also hit it right in the middle. However, it is inappropriate to shrink the bull's-eye *after* the arrow has been shot. After all, we would probably not be willing to see the shot as a miss if the smaller bull's-eye would have been drawn ahead of time and then barely missed by the arrow.

Playing fair in hypothesis testing involves setting the standard ahead of time and then simply making a binary decision based on the evidence; either the criterion was reached for rejecting the null hypothesis or it was not. Recognizing that findings can be close but not quite in the rejection region *or* can register far into the rejection region can be theoretically important and can help direct future research, but should not cause us to go back and adjust the probabilistic structure of the basic decision to reject or fail to reject the null hypothesis.

For these reasons, this textbook will only use the terms "significant" and "significance" sparingly. Admittedly, these terms are simply unavoidable for a statistics book, but minimizing their use will help us think clearly about the proper meaning of statistical analyses. Furthermore, this textbook will refrain from using the terms "marginally significant" and "highly significant." Inferential decisions will only be described as either a rejection of or a failure to reject the null hypothesis.

A Measure of Effect Size: Cohen's *d*

As just noted, one problem with hypothesis testing concerns the impreciseness of the resulting statistic to reflect the size of the effect (assuming the null is to be rejected). To address this problem, it is recommended for researchers to provide a measure of effect size whenever a significant finding is reported. As a result, as different tests are presented in the textbook, a means of measuring effect size will also be included.

One of the simplest, direct, and most often-used measures of effect size is **Cohen's *d***. Cohen (1988) suggested that the size of an effect can be standardized by using the standard deviation of the population to quantify the difference between the two means (in this case, between the null and the sample mean).

Cohen's *d* for single-sample *z* test

$$d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{M - \mu}{\sigma} \quad \text{(Formula 8.2)}$$

The standard deviation is included in the formula to standardize the size of the mean difference in much the same way that it is used by the *z* score formula to standardize raw scores. A mean difference of 10 units might suggest a huge effect if the two distributions are tightly packed around their respective means, but it may mean much less if the scores are widely distributed around them. Imagine a mean difference of 10 points where all of the scores in the population are within 8 units of each other. A mean difference of 10 units would be dramatic. However, a mean difference of 10 units where the raw scores have a standard deviation of 180 would not be nearly as impressive.

In some texts, the reader is informed that certain effect sizes should be classified as small, medium, or large. These distinctions are rather arbitrary and should not be seen as authoritative. The value of an effect size measure is found in the quantification of this concept and the realization that larger values reflect greater effects, differences, or relationships – whatever the case may be. Additionally, since this statistic is concerned with the size of the effect and not the direction, it is typical to simply ignore the valence of the obtained value and always report it as a positive value.

Versions of Cohen's *d* can be used to measure the effect size of many different types of *t* tests. As a result, we will see this statistic presented in other places throughout Part 3.

8.5 The Statistical Test for the Mean of a Population When σ Is Unknown: The *t* Distributions

The *t* Distributions

Up to this point, we have used the *z* distribution to establish cutoffs for testing the null hypothesis at a given alpha level. However, when σ is unknown, transforming all the means of a sampling distribution by the *z* statistic cannot be accomplished (the standard error cannot be calculated). Instead, the sample standard deviation, *s*, must be used as an estimate of σ. This necessitates a change in the formula and defines the *t* statistic.

***t* Statistic**

$$t_{obt} = \frac{M - \mu}{s_M} \quad \text{(Formula 8.3)}$$

where

$$s_M = \frac{s}{\sqrt{n}}$$

A ***t* distribution** is theoretically established by transforming every mean of a sampling distribution into a *t* statistic. This transformation is achieved by applying Formula 8.3 to every sample of the distribution, using M, μ, s, and n. However, because we do not know the standard error and are merely estimating it, the *t* formula will not produce a standardized normal curve identical to the *z* distribution. This means that we can no longer use the critical values of ±1.96 and ±2.58 to test the null hypothesis at the 5 and 1% alpha levels. However, *t* distributions are normal when the population of raw scores being sampled are normal or the sample size used is not extremely small. A *t* distribution can be formed for any sampling distribution composed of means that are based on an *n* of 2 to infinity. However, as *n* decreases, the tails of the *t* distribution stretch out farther down each direction of the *X* axis (with the critical values marking the most extreme 5% also moving farther and farther away from 0; see Figure 8.4). Since a different *t* distribution can be established for each sampling distribution of size *n, t* distributions are collectively referred to as a *family of distributions*. Figure 8.4 illustrates four *t* distributions based on sampling distributions of decreasing *n's*. Note how the distributions approximate the standardized normal curve as *n* increases (with the critical scores of the *t* distribution eventually equaling the critical scores for the *z* distribution – ±1.96). The sample size (*n*) is *roughly* reflected by the symbol *df* – the degrees of freedom. This concept is discussed in the following section. For the time being, all we need to know is that the degrees of freedom for the single-sample *t* test are equal to $n - 1$.

As we examine Figure 8.4, remember that the *t* distribution is not a sampling distribution of means. The *t* distribution is a distribution of *t* values. Each *t* value is a sample mean transformed by the *t* formula. The mean of the *t* distribution is 0. However, unlike the *z* distribution, the standard deviation of the *t* distribution is not 1. The standard deviation of the *t* distribution changes, depending, in part, on the size of the samples used to establish the sampling distribution. Sample size influences the degree of accuracy of our estimate of σ; as it increases, our estimate of σ improves, and correspondingly the error in the resulting sampling distribution shrinks.

As *n* decreases and the *t* distribution elongates, the region for rejecting the null is moved farther away from 0. For example, if $\alpha = .05$, the critical value when *df = 2 is* ±4.30; but when *df* = 25, the critical value is ±2.06. The broader *t* distribution associated with *df = 2* requires that we move 4.30 standard error units in each direction from the mean, to bracket the middle 95% of the distribution. When *the df* = 25, we need to only move 2.06 standard error units from the mean to bracket the middle 95%.

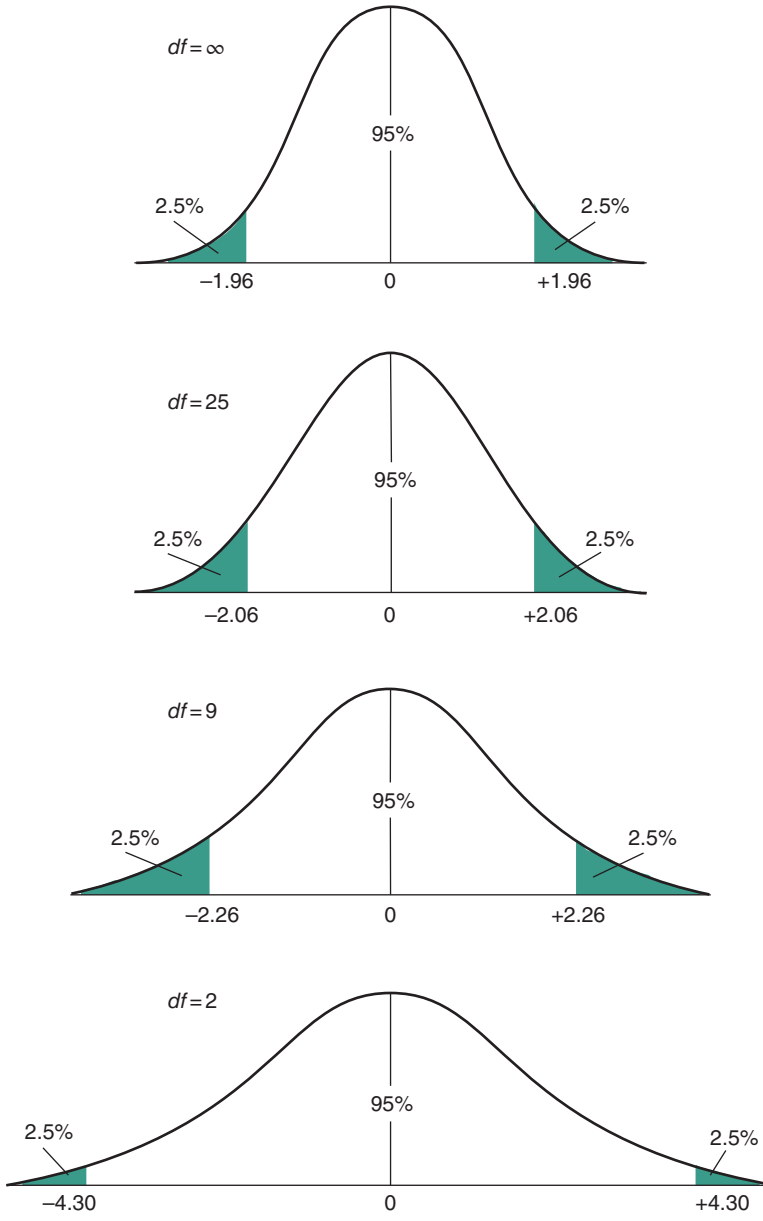


Figure 8.4 Distributions of *the t distribution* as the degrees of freedom change. As the sample size decreases, the tails of the distribution elongate. Note that the critical values that define the rejection region become more extreme as the sample size decreases.

Degrees of Freedom

Whenever a hypothesis test is conducted, an obtained value is computed and compared with a critical value. The critical value is found by referring to a table of critical values appropriate to the particular test statistic. We will get used to finding critical values for various kinds of inferential tests by looking up the **degrees of freedom** (***df***) associated with each test statistic. Moreover, degrees of freedom are also often used to compute obtained values. The reasons *why* mathematicians use degrees of freedom in developing inferential tests are complex and beyond an introductory statistics text (Walker, 1940). Therefore, instead of discussing why *df* are used, only the concept will be explained.

Suppose we are asked to pick four numbers. Since there is no restriction imposed, we are free to pick any four numbers. All four numbers are free to vary; therefore, we have four degrees of freedom. Now suppose a restriction is imposed: The four numbers must sum to equal 10. Now we are free to pick any numbers for the first three, but the fourth number must be determined based on the value of the other three so that the sum comes to 10. Since three numbers are free to vary, we have three degrees of freedom. In general, degrees of freedom refer to the number of values that are free to vary under some restriction.

The *df* concept applies only when making a statistical inference. Recall from Chapter 4 that there is a difference in the denominator when calculating the sample standard deviation as opposed to the population standard deviation. The sample standard deviation is used as an estimate of the population standard deviation.

Sample standard deviation

$$s = \sqrt{\frac{\sum (X - M)^2}{n - 1}}$$

Population standard deviation

$$\sigma = \sqrt{\frac{\Sigma (X - \mu)^2}{N}}$$

The restriction imposed when computing s is $\sum(X - M) = 0$. All of the numbers are free to vary, but the last number must result in $\sum(X - M) = 0$. The t test used to test the mean of one sample against a specified population mean (Formula 8.3) uses s in the process of determining the denominator (the estimate of the standard error), $s_M = s/\sqrt{n}$. Therefore, the *df* associated with this formula are $n - 1$. Whenever we use the t table to locate the critical value for the single-sample t test, $n - 1$ will be used as the *df* value.

Using the *t* Table to Find Critical Values

Table A.2 is the *t* table, a portion of which is reproduced below.

α Values for two-tailed test

df	.05	.01
⋮	⋮	⋮
8	2.306	3.355
9	2.262	3.250
10	2.228	3.169
⋮	⋮	⋮
∞	1.960	2.576

The first column of the *t* table specifies the degrees of freedom. Each *df* corresponds to a different *t* distribution. For *each df,* it would be possible to construct a table just like the *z* table, each allowing us to answer questions such as, "What percent of *t* values fall between a *t* of 0.34 and a *t* of 1.2?" However, no one is interested in asking these types of questions. The *t* table is only used to find cutoffs for different alpha levels used in hypothesis testing. Consequently, only the *t* values that correspond to various conventional rejection regions are presented.

The top two rows of the *t* table (Table A.2) state different alpha levels for one-tailed and two-tailed *t* tests. Only the two-tailed test is discussed in this chapter, and therefore only a couple typical alpha levels and their critical values for a two-tailed test are reproduced.[1] Suppose we set alpha at .05 and the sample size is 10. What are the critical values (or t_{crit})? Find $df = 9$, and then find the .05 column for a two-tailed test. The critical value is 2.262 (which should be understood as ±2.262 since we are splitting the rejection region into the two tails of the distribution). What are the critical values when $\alpha = .01$ and the sample size is 9? The answer is ±3.355 (remember that $df = n - 1$).

Now, follow the .05 column to the bottom row, ∞. The critical value is 1.96 – a number we should recognize. When *n* is very large, the *t* distribution assumes the shape of the standard normal curve. Therefore, the cutoffs for the 5% rejection region are the same as if we were using the *z* table. As we move from the bottom of the table to the top, the critical values increase. This reflects the fact that the *t* distribution becomes broader and the tails elongate as the sample size decreases. [Note: Table A.2 does not show every *df* value. If other t_{crit} values are needed, simply use an Internet search engine to find a more complete "*t* Table." Tables on the Internet can look different, but a careful reading should provide

1 The distinction between a one-tailed and two-tailed test is discussed in Chapter 9.

the proper value(s). Also, note that critical value differences between large *df*'s are negligible.]

A Note on Notation

The use of t_{obt} refers to the t value obtained from Formula 8.3; t_{crit} refers to the critical value to which t_{obt} is compared. The subscripts remind us which t value is being used.

Now we can put our knowledge to the test by using the t formula and the t table to test a hypothesis about the mean of a single population. To reject the null hypothesis, t_{obt} must equal or fall outside of t_{crit}.

■ **Question** *A health psychologist reports that the average high school student drinks six cups of coffee a week. The school board of Northside High would like to know if their students are drinking that much coffee. With confidentiality assured, 15 students are randomly selected and asked about their coffee habits. The mean number of cups of coffee consumed per week is found to be 4.2, and the sample standard deviation is 1.5. Test the hypothesis that the health psychologist's report is accurate for the students at this particular high school.*

Solution

Step 1. Identify the null and alternative hypotheses.

$$H_0: \mu = 6$$
$$H_1: \mu \neq 6$$

Step 2. Set alpha. Use an alpha of .05.

Step 3. Compute t_{obt} using Formula 8.3.

$$t_{obt} = \frac{M - \mu}{s_M}$$
$$s_M = \frac{s}{\sqrt{n}}$$
$$t_{obt} = \frac{4.2 - 6.0}{1.5/\sqrt{15}}$$
$$t_{obt} = \frac{-1.8}{0.39}$$
$$t_{obt} = \mathbf{-4.62}$$

Step 4. Using the t table, find the critical values for $df = 14$ and $\alpha = .05$. The critical values are ±2.145.

Step 5. Compare the t_{obt} of –4.62 with the critical values of ±2.145. Since t_{obt} falls outside ±2.145, the null hypothesis is rejected in favor of the alternative hypothesis ($H_1: \mu \neq 6$).

Step 6. Interpret the findings. Statistical evidence suggests the amount of coffee consumed per week among this school's students is less than the amount stated in the newspaper. ■

Interpreting Inferential Findings

Before going any farther, we need to make a few important comments about interpreting inferential tests. First, notice within the interpretation that the researcher is free to make a directional statement (i.e. the population mean in question is either less or more than the hypothesized mean value, as the case may be). Of course, the null can be rejected in either direction. Once the null is rejected, however, the researcher is free to look at the valence of the t_{obt} score and make a more specific interpretation. In this case, we conclude that students at this high school drink *less* coffee than high school students in general.

Second, notice that the interpretation is made using cautious language. It starts with the phrase, "Statistical evidence suggests…." Although this specific wording is not a requirement of any social science discipline, we will stick closely to it throughout the text, if, for no other reason, then it creates a good habit of thinking. This phrase allows us to underscore three important points about the outcome of any inferential test; it is *evidence* (not a guess and not an opinion) that is *statistical* in nature (not logical, legal, or of some other form) and merely *suggestive* (probabilistic, not proof). All three of these are important concepts.

Finally, if an analysis results in a failure to reject the null hypothesis, a good phrase to use is, "We do not have evidence to suggest…." We should *not* say, "We have evidence to suggest the null hypothesis is true." This would be to claim that we should accept the null and we have already clarified that inferential tests do not allow for this strong of an inference. The proper way to communicate a failure to reject the null is to state that we *do not have evidence* suggesting a difference, effect, or relationship, whatever the case may be. Either we *have* evidence (when we reject the null), or we *do not have* evidence (when we fail to reject the null).

■ **Question** *A bank president states that the average amount of money on deposit in savings accounts is $6500. To test the hypothesis that μ = $6500, a random sample of nine deposits is examined. The mean of the sample is $7500 and the standard deviation is $1500. Is there evidence to reject the president's claim?*

Solution

Step 1. Identify the null and alternative hypotheses.

$$H_0: \mu = 6500$$

$$H_1: \mu \neq 6500$$

Step 2. Set alpha. Use an alpha of .05.
Step 3. Compute t_{obt} using Formula 8.3.

$$t_{obt} = \frac{M - \mu}{s_M}$$

$$s_M = \frac{s}{\sqrt{n}}$$

$$t_{obt} = \frac{7500 - 6500}{1500/\sqrt{9}}$$

$$t_{obt} = \frac{1000}{500}$$

$$t_{obt} = \mathbf{2.00}$$

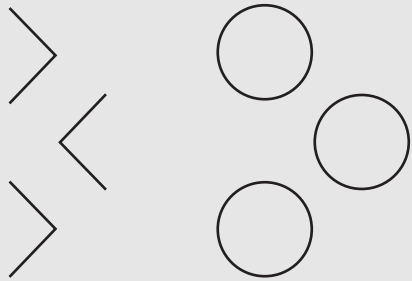
Step 4. Find the critical values *for df = 8* and $\alpha = .05$. The critical values are ±**2.306.**
Step 5. Compare the t_{obt} of 2.00 to the critical values of ±2.306. Since t_{obt} does not fall outside ±2.306, the null hypothesis is not rejected.
Step 6. Interpret the findings. We do not have evidence to suggest the null hypothesis is incorrect. Keep in mind that this interpretation does not mean that we know that $6500 *is* the average amount of money on deposit or even that we have evidence suggesting this to be true. Failing to reject the null hypothesis does not allow us to accept it. All we can say is that we have failed to find evidence rejecting the claim that $\mu = \$6500$. ■

Box 8.2 features the use of a single-sample *t* test to analyse the data for a published research study dealing with perceptual illusions.

Box 8.2 Visual Illusions and Immaculate Perception

Perceptual illusions reflect the fact that human perceptions are imperfect and are not just copies of images on the retinas. People actively construct sensory information, and in the case of illusions, their perceptions are hardly immaculate.

One visual illusion is called the Morinaga misalignment illusion. First, examine the alignment of the angles on the left. We will note that the apex of the middle angle does not appear to align perfectly with the apexes of the top and bottom angles. Now, take a ruler or the edge of a sheet of paper and see if the apexes of the angles fall along a straight line. They do. We experienced a perceptual illusion as we perceived the middle angle to be misaligned.



Day and Kasperczyk (1984) wondered if the same illusion is found with the circles on the right. (Notice that the left side of the middle circle does not appear to align with the right sides of the circles above and below.) Twelve participants were asked to move the middle circle (without the help of a straight edge) so that the left side of the middle circle aligned perfectly with the right sides of the top and bottom circle. Since the circles were already perfectly aligned, any adjustment by the participant defined an error of some magnitude, and reflected the workings of the Morinaga illusion. Errors were measured in millimeters (mm). If the illusion did *not* exist for circles, then there should not have been realignments by the participants, or the realignments across the participants should have been just as likely to be to the right as to the left, so the errors would sum to 0. The null hypothesis is a statement of no perceptual illusion: $H_0 : \mu = 0$, with $H_1 : \mu \neq 0$. The authors found an average error among the 12 participants to be 1.44 mm to the right with $s = 2.07$ mm.[2] The question is how unlikely is a sample mean of 1.44 mm when the population mean is hypothesized to be 0? The question can be answered with a single-sample *t* test:

$$t_{obt} = \frac{M - \mu}{s/\sqrt{n}}$$

$$t_{obt} = \frac{1.44 - 0}{2.07/\sqrt{12}}$$

$$t_{obt} = \frac{1.44}{0.60}$$

$$t_{obt} = 2.40$$

$$df = n - 1 = 12 - 1 = 11$$

$$\alpha = 0.05$$

$$t_{crit} = \pm 2.201$$

2 These values are reported by Kiess (1989, p. 214).

Since the obtained t of 2.40 falls outside the critical value of ±2.201, the null hypothesis is rejected. The statistical evidence suggests that the sample mean of 1.44 mm to the right is unlikely to come from a population with a mean of 0. The findings can be summarized as statistical evidence suggesting the Morinaga illusion is a general perceptual effect that is shown not only with angles but also with circles, $t(11) = 2.40$, $p < .05$.

Cohen's *d* for the Single-Sample *t* Test

As previously noted, one problem with hypothesis testing concerns the imprecise ness of the resulting statistic to reflect the size of the effect (assuming the null is to be rejected). Just as Cohen's d was used to determine the effect size for z test, it can also be used to *estimate* the effect size for a t test. For the single-sample t test, the formula is as follows:

Cohen's *d* for single-sample *t* test

$$d = \frac{mean\ difference}{sample\ standard\ deviation} = \frac{M - \mu}{s}. \quad \text{(Formula 8.4)}$$

8.6 Assumptions of the Single-Sample *z* and *t* Tests

Over the course of the remainder of the text, we will be exposed to several different inferential tests. The appropriate statistical test to use will depend on the research question of interest and the research method used. All statistical tests have assumptions. The ability to rely on the conclusions drawn from an inferential test is based on the degree to which the assumptions have been met. Some assumptions can be modestly violated without seriously compromising the interpretation of a statistical test. Other assumptions are critical. The assumptions for the single-sample t test, as well as the z test, are presented below.

1) **Representativeness.** It is assumed that the participants comprising the sample are *representative* of the population in question. The goal of inferential statistics is to generalize from a sample to a population. If the sample is not representative of the population, it is possible that an untrue statement about the population will be made from the nonrepresentative sample. The best way to ensure the representativeness of a sample is by *randomly sampling* from the population. Obviously not all studies can use random sampling; however, if other sampling methods are used, representativeness can become a concern. The representativeness assumption is *not* a mathematical assumption. Representativeness is an assumption of the research

methodology. If violated, the interpretive conclusions that follow from the *t* test may not be valid.

2) **Independent observations.** This assumption means that each score *within the sample* is *independent* of all other scores. In most applications, independent observations mean that each participant supplies only one score. However, it is possible to violate the independence assumption even when only one score is obtained per participant. If the behavior of one participant in the study is influenced by the behavior of another participant, then the scores from these two participants are *not* independent of each other. For example, earlier in this chapter we speculated about the effects of an educational enrichment program. Suppose two participants studied together, and, as a result, their performance was influenced by their contact. The scores from these participants would not be independent of one another.
3) **Interval or ratio scale of measurement**. The single-sample *z* and *t* tests utilize means as well as standard deviations. Both of these concepts only have meaning for data corresponding to a scale of measurement where the quantitative distance between integers is held constant, namely, an interval or ratio scale (see Chapter 2). Means and standard deviations should not be calculated for samples of scores measured using an ordinal or nominal counting system (see Chapters 3 and 4).
4) **Normality.** The fourth assumption states that the population from which the sample is taken is normally distributed. Recall that a normal sampling distribution is needed for inferential analysis. If the population is not normal, the tests may still lead to valid conclusions, provided the sampling distribution is normally distributed. For this reason, it is often claimed that *z* and *t* tests are robust to violations of normality, provided *n* is of sufficient size. A **robust statistic** is resistant to violations of certain assumptions; although the assumption was not met, the conclusions are still valid. Determining the sufficient sample size for a test to be robust to the assumption of normality involves an advanced discussion. For the purposes of this introductory textbook, we will assume that all data sets presented herein will be from normally distributed populations. For general use, a rule of thumb is if *n* is in double digits, and especially if it is approaching 30, we can safely assume the sampling distribution will be normal, and violations of normality in the raw population data are inconsequential.

8.7 Interval Estimation of the Population Mean

In the previous chapter, we learned that there are actually two kinds of inferential procedures, hypothesis testing and estimation. Up to this point, we have been focused on hypothesis testing. However, using a sample mean, the *t* distribution concept, and an estimate of the standard error, it is possible to

generate an interval estimation of the population mean and quantify the confidence that it falls within that interval. Since each potential sample mean drawn from a population has a corresponding t value, we can use the t distribution and our obtained sample mean (which is an unbiased estimate of the population mean) to generate a probability function for the value of the actual population mean. Choosing t_{crit} values corresponding to different probabilities within the t distribution allows us to create intervals with differing degrees of certainty. The formula for an interval in which we can have 95% confidence follows:

The 95% confidence interval for a population mean

$$LL = M - t_{.05}s_M$$
$$UL = M + t_{.05}s_M \quad \text{(Formula 8.5)}$$

where

LL = the lower limit of the confidence interval
UL = *the* upper limit of the confidence interval
$t_{.05}$ = the critical value for a t distribution of a given sample size

Since we are generating an interval, two values are calculated, one being the value at the lower end of the interval and the other at the upper end. As the interval widens and becomes less specific, the confidence grows that the actual mean falls within that window. A 95% confidence rate is typical, but the above formulas could easily be adjusted to find a 90 or 99% confidence interval simply by finding the corresponding t_{crit} values using the t table (Table A.2).

■ **Question** *Using the same data previously presented by the health psychologist investigating the coffee-drinking* habits *of Northside High School students, find the 95% confidence interval for the population mean (M = 4.2, s = 1.5, and n = 15).*

Solution

Step 1. Identify the null and alternative hypotheses.

$$H_0: \mu = 6$$
$$H_1: \mu \neq 6$$

Step 2. Set the confidence rate at 95%.

Step 3. Using the t table, find the cutoff values beyond which lie 2.5% in the right tail of the t distribution and 2.5% in the left tail of the distribution. The task is accomplished in the same way that we found t_{crit} for the t test. With df = 14 and α = .05 (two-tailed test), the cutoff points are ±2.145.

Step 4. Compute the confidence interval.

$$LL = M - t_{.05}s_M$$

$$UL = M + t_{.05}s_M$$

$$LL = 4.20 - 2.145\left(1.50/\sqrt{15}\right) = \mathbf{3.36}$$

$$UL = 4.20 + 2.145\left(1.50/\sqrt{15}\right) = \mathbf{5.04}$$

Step 5. Interpret the findings. Statistical evidence suggests the mean number of cups of coffee consumed per week among the *population* of Northside High students lies between 3.36 and 5.04. ■

8.8 How to Present Formally the Findings from a Single-Sample *t* Test

Proper reporting of inferential statistics can be challenging. Following are examples of how to report, in sentence form, a rejection of the null as well as a fail to reject the null. If rejecting the null, a sentence might read, "A single-sample *t* test found evidence that students from Northside High School consume less coffee than high school students in general, $t(14) = -4.62$, $p < .05$." Notice the following; *t* and *p* are italicized, the degrees of freedom are placed within parentheses right after the statistic is identified, a comma follows the actual statistical finding, the alpha value used is shown to have been eclipsed by the expression $p < .05$, t_{crit} is not reported, and a comma precedes the entire statistical expression that sits at the end of the sentence. Note also that a "0" is not typically placed to the left of the decimal in the probability statement. This accurately reflects what readers will find in most professional publications. If we also wanted to include a measure of effect size, the sentence could finish with, "… $t(14) = -4.62$, $p < .05$, $d = 1.2$."

If failing to reject the null, a sentence might read, "A single-sample *t* test did not find evidence that Northside High School students consume a different amount of coffee from high school students in general, $t(14) = -1.62$, *n.s.*" First, notice the wording near the front of the sentence. It did not accept the null by saying that evidence was found of no difference. Rather it said that no evidence was found of a difference. This may seem like an insignificant point, but it is not. Second, notice how the final part of the statistical expression used the letters *n.s.* (italicized). This stands for "not significant." Each inferential statement should end with either a description of the alpha value that was eclipsed, if the null hypothesis is rejected (e.g. $p < .05$), or an expression communicating that the null hypothesis was not rejected (*n.s.*).

Summary

Hypothesis testing involves either rejecting or not rejecting the null hypothesis. The decision is probabilistic in nature, and the ultimate truth value of the null hypotheses can never be known. The decision to reject or fail to reject the null hypothesis risks two types of errors. A Type I error is committed when a true null hypothesis is rejected. The probability of making a Type I error is directly controlled by alpha, the criterion of significance. A Type II error is committed when a false null hypothesis is not rejected.

The single-sample z test is used to decide if a population mean is not a specified value. The z_{obt} transforms a sample mean into a z value, which indicates the number of standard error units that the sample mean is from the mean of the sampling distribution. If the z_{obt} equals or falls outside of the critical values, then statistical evidence exists to reject the null hypothesis; otherwise, the null hypothesis should not be rejected.

The t statistic is used to test the null hypothesis when σ is unknown. In the t formula, s is used to estimate σ, and $s/\sqrt{n}$ is used to estimate $\sigma/\sqrt{n}$. The t statistic is used to transform a sampling distribution of means into a t distribution. The shape of the t distribution will approximate the standard normal curve as n increases. The critical value to which t_{obt} is compared is based on $n - 1$ degrees of freedom.

The assumptions for the single-sample z and t tests are representativeness, independent observations, interval-scaled or ratio-scaled data, and population distributions that are normally distributed. These tests are robust to violations of normality as n increases.

The standard error, sample mean, and t distribution can also be used to create a confidence interval for the actual value of an unknown population mean.

Statistical significance reflects the degree of certainty that the null hypothesis is false, but it does not necessarily reflect the size of the difference between the null mean and the sample evidence. To measure effect size, Cohen's d can be calculated.

Using Microsoft® Excel and SPSS® to Run Single-Sample *t* Tests

Excel

General instructions for data entry into Excel can be found in Appendix C.

Data Entry

Enter all of the scores from the sample in one column of the spreadsheet. Label the column appropriately.

Data Analysis

1) Excel has built-in programs for several types of *t* tests; however, it does not have one for the single-sample *t* test. As a result, we will need to figure the components of Formula 8.3 ourselves.
2) Determine the null hypothesis (μ), and record it in an open cell (label it appropriately).
3) Determine the sample mean by using the built-in Excel function "AVERAGE." Record it in an open cell (label it appropriately).
4) Determine the estimate of the standard error (s_M) by first determining the sample standard deviation (s) using the built-in function (either STDEV or STDEV.S; both calculate the standard deviation of a sample, which is what we want) and our sample size (n). Once we have these two values, we can determine the estimate of the standard error (Formula 7.3 – $s_M = s/\sqrt{n}$). Record this value in an open cell (label it appropriately).
5) The *t* value can now be determined using Formula 8.3 – $t_{obt} = (M - \mu)/s_M$.
6) Use n to find the appropriate *df*. The *df* for a single-sample *t* test is $n - 1$.
7) Go to Table A.2 (*t* Table) in the Appendix, and use the *df* value and the alpha value (usually .05) to find the appropriate t_{crit} value(s).
8) Compare the observed *t* with the critical *t*, and make the appropriate inference regarding the null hypothesis. (See Figure 8.5 for a worked example.)

Number of cousins		
5	$H_0 =$	6.5
7	$M =$	9.2
12	$s =$	4.131182
8	$s_m =$	1.307336
18	$t_{obt} =$	2.065268
3	$t_{crit} =$	±2.262
10		
8	Fail to reject the null hypothesis	
11		
10		

Figure 8.5 A worked example of using Microsoft Excel to calculate a single-sample *t* test value.

SPSS

General instructions for inputting data into SPSS can be found in Appendix C.

Data Entry

Label the variable appropriately in "Variable View." Enter all of the scores from the sample into the appropriate column in "Data View" – one score per row. (See Figure 8.6 for an example of data entry for a single-sample *t* test in SPSS.)

...	cousin_num
1	5
2	7
3	12
4	8
5	18
6	3
7	10
8	8
9	11
10	10

Figure 8.6 An example of entered data for a single-sample *t* test in SPSS.

Data Analysis

1) Click **Analyze** on the tool bar, select **Compare Means**, and then click **One-Sample T Test**.
2) Highlight the appropriate column label in the left box, and click the arrow to move it into the **Test Variable(s)** box.
3) In the **Test Value** box at the bottom of the right-hand side, enter the hypothesized value for the population mean given a true null hypothesis. This value is automatically set at zero unless it is changed.
4) Click **OK**.
5) The output will generate two boxes. The first box will identify how many scores were in the sample (N) as well as the mean, standard deviation, and standard error. The second box will identify the *t* value, degrees of freedom, significance level, mean difference, and the 95% confidence intervals of the actual mean difference. It will not generate t_{crit}. Either we can look up t_{crit} ourselves, or we can look at the significance level to see if that value is equal to or lower than .05. If it is, then we can reject the null. If it is not, then we need to fail to reject the null hypothesis. (See Figure 8.7 for a worked example.)

***T*-test**

One-sample statistics

	N	Mean	Std. deviation	Std. error mean
cousin_num	10	9.20	4.131	1.306

One-sample test

	Test value = 6.5					
					95% confidence interval of the difference	
	t	*df*	Sig. (2-tailed)	Mean difference	Lower	Upper
cousin_num	2.067	9	0.069	2.700	–0.26	5.66

Figure 8.7 A worked example using SPSS to calculate a single-sample *t* test.

Key Formulas

z Statistic

$$z_{obt} = \frac{M - \mu}{\sigma_M} \quad \text{(Formula 8.1)}$$

Cohen's *d* for single-sample *z* test

$$d = \frac{mean\ difference}{standard\ deviation} = \frac{M - \mu}{\sigma} \quad \text{(Formula 8.2)}$$

t Statistic

$$t_{obt} = \frac{M - \mu}{s_M} \quad \text{(Formula 8.3)}$$

Cohen's *d* for single-sample *t* test

$$d = \frac{mean\ difference}{sample\ standard\ deviation} = \frac{M - \mu}{s} \quad \text{(Formula 8.4)}$$

The 95% confidence interval for a population mean

$$LL = M - t_{.05}s_M$$
$$UL = M + t_{.05}s_M \quad \text{(Formula 8.5)}$$

Key terms

***z* Test**	**Type II error**
***t* Test**	**Cohen's *d***
Alpha level	***t* Distribution**
Critical values	**Degrees of freedom (*df*)**
Type I error	**Robust statistic**

Questions and Exercises

1 A statistical test aids a researcher in deciding whether an experimental effect is due to chance. What does this mean? Is it possible to know for sure that an effect was not due to chance? Explain.

2 Provide some examples of single-sample research projects. In each instance, provide a research hypothesis as well as a corresponding pair of statistical hypotheses.

3 For each of the following situations, specify the null and alternative hypotheses:
- **a** The average respiration rate per minute is 8. Do smokers have an average rate different from 8?
- **b** The average score on the Beck Depression Inventory is 12. Does the average depression score of mothers with young children deviate from the population mean?
- **c** The average miles per gallon (mpg) of cars used in the United States is 20. Is the observed mpg of a sample of cars used in Japan different?

4 When conducting an inferential test, when should we use the *t* distribution?

5 What if a researcher conducting a project using a single-sample design has access to both the population and the sample standard deviation, which test should they use?

6 On what basis does a researcher decide on a given alpha level?

7 What type of error corresponds to a "false positive?"

8 What type of error corresponds to a "false negative?"

9 Failing to reject the null when, in actuality, it is false is the making of what type of error?

10 Rejecting the null when, in actuality, it is true is the making of what type of error?

11 For which type of error can we set the precise risk rate, and which one can we only increase or decrease the chance of making?

12 If the difference between a sample mean from a treated sample and the known population mean is 5 and the standard deviation of the population is 10, what is Cohen's d?

13 If new basketball shoes are supposed to elevate the wearer's jump by 3 in. and the standard deviation of heights jumped by basketball players is 4 in., what is the supposed effect size of the shoes in terms of Cohen's d?

14 If we have been told the effect size, according to Cohen's d is 0.4, and we know the mean difference is 12, what is the standard deviation of the population?

15 A publisher of a new statistics textbook claims an effect size (Cohen's d) of 20% (or 0.2) regarding a nationally normed statistics knowledge test for its users. If we know the mean and standard deviation of this nationally normed test is 100 and 20, respectively, what is the publisher's hypothesized mean for the users of their textbook?

16 Among trained typists, suppose it is known that the average typing speed using a standard keyboard is 60 words per minute (wpm), with a standard deviation of 5 wpm. The manufacturer of an ergonomically designed keyboard claims their device will improve typing speed. A random sample of 50 typists is tested on the ergonomically designed keyboard, and the sample mean wpm is 65. Test the hypothesis that using the new device affects typing speed. Set alpha at .05.
- **a** Should we use the z distribution or t distribution? Why?
- **b** State H_0 and H_1.
- **c** What are the critical values?
- **d** What is the obtained statistic?
- **e** Reject the null hypothesis?
- **f** What type of decision error might have been made?
- **g** Is there sufficient evidence to support the manufacturer's claim?
- **h** If so, what is the effect size?

17 On one standardized measure of IQ, $\mu = 100$ and $\sigma = 15$. Imagine we want to test the hypothesis that children of parents with college degrees have an average IQ that is greater than the national average. A sample of 100 students who have college-educated parents is randomly selected, and the

mean is 110 with a standard deviation of 12. Conduct a test of the null hypothesis and set alpha at .05.

a Should we use the z distribution or t distribution? Why?
b State H_0 and H_1.
c What are the critical values?
d What is the obtained statistic?
e Reject the null hypothesis?
f What type of decision error might have been made?
g Interpret the finding.
h If the null is rejected, what is the effect size?

18 Suppose the mean weight of adult golden retrievers is 90 lb. A veterinarian claims to be able to double the size of golden retrievers by injecting a hormone into retriever pups when they are eight weeks old (why someone would want to produce a humongous golden retriever, who knows; it may have theoretical significance). A sample of 41 pups is injected with the hormone, and their average weight at maturity is found to be 110 lb, with a standard deviation of 30 lb. Conduct an inferential test with $\alpha = .05$. (Note that the inferential test will not be able to address the assertion that the hormone doubles the size of dogs. The statistical test will only be able to help us decide if the hormone affects the breed's weight, either increasing or decreasing.)

a Should we use the z distribution or t distribution? Why?
b State H_0 and H_1.
c What are the critical values?
d What is the obtained statistic?
e Reject the null hypothesis?
f What type of decision error might have been made?
g Interpret the finding.
h If the null is rejected, what is the effect size?

19 A researcher would like to determine whether the students at her university sleep more than most students. Suppose it is known that the amount of hours university students sleep is skewed to some degree with a $\mu = 7.5$ hours per night and $\sigma = 2.4$. The researcher takes a sample of $n = 200$ students at her university and finds they average 7.2 hours of sleep per night with a standard deviation of 1.8.

a What is the appropriate test statistic? Why?
b State H_0 and H_1.
c Does it matter that the population is skewed to some degree? Why or why not?
d What are the critical values at $\alpha = .05$?
e What is the obtained statistic?
f Reject the null hypothesis?

g What type of decision error might have been made?
h Interpret the finding.
i If the null is rejected, what is the effect size?

20 An industrial/organizational psychologist believes that people who work at home experience greater job satisfaction. Imagine that a job satisfaction rating scale exists. The publishers of this scale claim the population is normally distributed with a mean of 50. The psychologist samples 20 people who work at home finding $M = 63$ and $s = 17$.
a Should we use the z distribution or t distribution? Why?
b State H_0 and H_1.
c What are the critical values?
d What is the obtained statistic?
e Reject the null hypothesis?
f What type of decision error might have been made?
g Interpret the finding.
h If the null is rejected, what is the effect size?

21 How would we use the concept of sampling error in discussing hypothesis testing?

22 Subjective life expectancy is a person's belief in how long they will live. Several years ago, Robbins (1988) found that a sample of biological females estimated their life expectancy to be 77.2 years – close to the actual life expectancy for women at the time (79.2 years). Biological males, on the other hand, tend to overestimate their life expectancy. At the time of Robbins research, the actual life expectancy for biological males was 72.4 years. The following hypothetical data are consistent with Robbins' findings.

Subjective life expectancy for males (years)
77
74
80
72
82
76
78
75
79

a Should we use the z distribution or t distribution? Why?
b State H_0 and H_1.
c What is the critical value?
d What is the obtained statistic?
e Reject the null hypothesis?
f What type of decision error might have been made?
g Interpret the findings.
h If the null is rejected, what is the effect size?

23 An anthropologist hypothesizes that physical stress in childhood increases height (Landauer & Whiting, 1964). The researchers locate a tribe of people in which physical stress is a by-product of frequent tribal rituals (e.g. piercing and molding body parts, exposure to extreme temperatures, etc.). The mean height of the people in the region who do *not* use physically stressful rituals with their young is used as the population mean. The following raw data are for adult biological males and women of the tribe in question. Conduct a t test for men and a t test for women. The population mean height for men is 65 and 59 in. for women.

Men	Women
67	59
69	63
72	65
70	60
70	59
72	62
64	61
70	66

a What is t_{obt} for men?
b What is t_{obt} for women?
c What are the critical values for each test ($\alpha = .05$)?
d Compare each t_{obt} with its respective critical values and interpret the findings; present the findings in a professionally appropriate manner.

24 The chairperson of a sociology department at a major research institution claims that the mean number of publications by the department's faculty is higher than the mean for other sociology departments at comparable schools across the country. Suppose the mean number of publications in the

population is 16. A random sample of eight professors is taken from the sociology department in question. The sample mean is found to be 20, with a standard deviation of 2.8. Is there any evidence to support the chairperson's claim? Set alpha at .05 when conducting the *t* test. Properly present the finding.

25 The director of the Department of Mental Health has received conflicting reports about the frequency of patient assaults on inpatient units. The director would like an idea of the mean number of assaults in a one-month period. A random sample of 18 inpatient units is taken. The average number of assaults is found to be 24.50, per inpatient unit, with a standard deviation of 2.61. Compute the 95% confidence interval for the population mean.

26 A drug manufacturer is researching a new medication for high blood pressure. Early reports suggest that there are many negative side effects to the drug. A random sample of 13 patients taking the drug is selected, and the mean number of side effects is found to be 4.2, with a standard deviation of 0.86. Compute the 95% confidence interval for the population mean.

27 A tire company is interested in knowing the average number of highway miles its tires can tolerate before the treads wear out. A random sample of 60 tires is selected, and each is placed on a highway simulator wheel. The mean for the sample is found to be 56 000 miles, with a standard deviation of 4 300 miles. Compute the 95% confidence interval for the population mean.

Computer Work

28 A health psychologist is interested in educating high school students about the negative effects of smoking. Fifty students who smoke are randomly selected to participate in the program. To measure the success of the program, the average number of cigarettes smoked per day among the participants is obtained 10 weeks after the end of the program. Assume that previous research had shown that, among all smoking students, the average number of cigarettes smoked in a day was 17. Set alpha at .05, and conduct a *t test* on the following data. Interpret the findings.

Average number of cigarettes consumed per day among participants									
12	11	7	0	0	6	2	23	45	0
0	1	2	0	3	16	8	22	17	9
12	10	6	5	9	11	0	33	24	5
11	10	0	0	0	22	4	22	21	0
10	11	0	6	7	11	3	42	38	0

29 An insurance company states that it takes them an average of 15 days to process an auto accident claim. A random sample of 40 claims is drawn from processed claims over the past six months. Based on the following data, is there any evidence that the mean number of days to pay claims is not 15? Set $\alpha = .05$.

Number of days to process a claim									
22	11	7	9	9	8	7	23	45	9
23	21	8	8	5	16	9	22	17	6
12	29	6	5	9	23	7	33	24	5
15	14	9	7	3	17	8	19	15	8

30 Every day a commuter records the amount of time the train is late. Over a period of two months, the mean number of minutes that the train was late is 24.5. The train authorities state that the problem has been resolved by the addition of extra trains during rush hour. For the next 30 working days, the commuter records the amount of time that the train is late. Based on the following data, does there seem to be an improvement in service?

Number of minutes the train is late									
22	11	25	19	9	8	7	23	45	20
23	21	8	1	5	16	3	22	17	24
12	16	22	32	9	10	7	11	10	2

9

Testing the Difference Between Two Means: The Independent-Samples *t* Test

9.1 The Research Context

In Chapter 8, the *t* test was used to contrast a sample mean with a specified population value. Only one sample was drawn to infer the mean of a population; the question was whether the population mean was or was not a given value. A more common and interesting usage of the *t* test arises when means *from two different samples* are compared to infer whether there is a difference between the means of the two populations from which the samples came. If the two samples are scores coming from two different sets of participants, the appropriate *t* test is called an **independent-samples *t* test**. This is the topic for Chapter 9. If the two samples come from one set of participants measured under different circumstances, the appropriate *t* test is called a dependent-samples (or paired-samples) *t* test. (That test is the topic of Chapter 10.) Here are three research examples in which an independent-samples *t* test can be used to analyze the data.

► **Example 9.1** Pham, Hung, and Gorn (2011) found that the more relaxed a shopper is when they enter a store, the more money they will spend. The researchers induced two states of relaxation in participants: One group was induced to have a pleasant mind-set and become very relaxed, while the other group felt equally pleasant but was not as relaxed. Participants were then asked to assess the monetary value of a set of items. The findings showed that very relaxed participants bid significantly higher for a whole range of auctioned items than the less relaxed participants. ◄

► **Example 9.2** Zakahi and Duran (1988) hypothesized that the very lonely are less physically attractive than those who are not lonely. A loneliness questionnaire was administered. Participants who scored in the top 25% were considered very lonely, whereas participants who scored in the bottom 25% were defined as not lonely. All participants' photographs were rated by three

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

judges for attractiveness (1 [very unattractive] to 10 [very attractive]). Biological males rated biological females and vice versa. There was no statistical difference in attractiveness ratings between very lonely and not lonely females. However, lonely biological males were statistically rated as less physically attractive than biological males who were not lonely. ◄

► **Example 9.3** In 1972, Buffalo Creek, West Virginia, was the scene of a major flood. The flood was a consequence of corporate negligence. Coal waste that was dumped in a mountain stream created an artificial dam. After several days of rain, the dam gave way, and a black wall of water, over 30 ft high, descended on mining hamlets in the valley. In less than 1 hour, 125 people were dead and 5000 people lost their homes. Simpson-Housley and DeMan (1989) found that, 17 years later, the residents of Buffalo Creek scored higher on a measure of trait anxiety in comparison with the residents of Kopperston, a nearby mining town that did not experience the flood. ◄

The Between-Participants Design

Between-participants designs (also called *between-groups*, *independent-samples*, *or independent-groups* designs) are defined by the fact that each group of participants comes from a different population.[1] Moreover, no participant or group of participants are members of both populations. Oftentimes this design is used for experimental purposes. For instance, pulling a very simple but important study out of psychology's past, researchers Shelton and Mahoney (1978) asked one group of athletes to employ their customary "psyching-up" strategies and asked a control group to just count backward by sixes. Performance on a strength task served as the dependent variable. The averages of the two groups were then compared to evaluate the effect of using psyching-up strategies. Since each group received a different treatment, this study was a between-participants design. (The researchers found evidence suggesting "psyching-up" helped.)

The independent-samples *t* test can be used when a researcher uses a between-participants design, whether an experimental manipulation is involved or not. The impact of different data gathering techniques for the two means does not influence the statistical inference, but rather the interpretation. In experimental settings, if participants are randomly assigned to two conditions, it *might* be possible to make a causal statement about the relationship between the independent and dependent variables, depending on the presence or absence of confounding variables (see Chapter 1). However, the independent-samples

1 Older references may use the term *between-subjects* design; the term *subjects* was replaced with the term *participants* in most social and behavioral science professional literature in the 1990s.

t test can also be used to compare two means that are obtained from a study in which participants are not randomly assigned to groups (i.e. correlational designs or quasi-experimental designs). The second and third described research examples at the beginning of the chapter are cases in point. Since there was no randomization of participants, the research method was correlational in nature. Even though a *t* test can be used to compare means, no causal statement can be advanced about the relationship between variables. However, since Shelton and Mahoney's (1978) research on psyching-up strategies did randomly assign participants to experimental and control conditions, a causal relationship between the independent and dependent variables is possible. In general, the interpretation of any inferential test depends on the manner in which the study is designed. Methodology governs the interpretation of the statistical findings.

In Spotlight 9.1 we take a closer look at the person who is most responsible for creating *t* tests.

Spotlight 9.1 William Gosset

William Gosset (1876–1937) developed the *t* distribution as well as the independent- and dependent-samples *t* tests. After receiving a degree in chemistry and mathematics from Oxford, Gosset was hired by the Guinness brewery in Dublin in 1899. Around the turn of the century, many companies, especially in the agricultural industry, attempted to apply a scientific approach to product development. A typical research question would have been "Which fertilizer will produce the largest corn yield?" or "What is the best temperature to brew ale so as to maximize its shelf life?" Until Gosset's work, statisticians dealt with very large numbers of observations, in the hundreds and thousands. Traditional wisdom held that one should take a very large sample, compute the mean and standard deviation, and refer to the *z* table to make probability statements. The problem that confronted Gosset was how to make inferences about the difference between population means when sample sizes were small. For example, suppose 10 plots of barley are treated with one fertilizer and 10 plots are treated with another fertilizer. With such small samples (before Gosset), there was no way to determine if the difference in yield was due to sample fluctuation (chance) or the effect of the brand of fertilizer.

To test the mean of one sample against a specified population value or test the difference between two sample means, the *t* table (instead of the *z* table) is used to find critical values and make probability statements when σ is unknown. In his seminal 1908 article, "The Probable Error of a Mean," Gosset addressed the problem of small samples: "As we decrease the number of experiments, the value of the standard deviation found from the sample of experiments becomes itself subject to an increasing error, until judgments reached in this way become altogether misleading" (Student, 1908; p. 2). He realized that the standard

normal curve, on which the *z* table is based, leads to inaccurate judgments about the area under the curve of a sampling distribution when sample sizes are small and σ is unknown. In the following quote, Gosset expressed the purpose of his 1908 paper. "The aim of the present paper is to determine the point at which we may use the tables of the probability integral in judging of the significance of the mean of a series of experiments, and to furnish alternative tables for use when the number of experiments is too few" (p. 2). (His reference to the tables of the probability integral refers to the *z* table, and "alternative tables" refers to the newly developed *t* table.) Gosset's use of the term "significance" was prophetic since at this time the concept of significance testing had not been developed. The conventional use of the 5% level of significance emerged over the next 25 years.

Gosset's classic 1908 article is one of the most important publications in the history of inferential statistics. "With one stroke, he: (1) discovered a new statistical distribution; (2) invented a statistical test that became the prototype for a whole series of tests, including analysis of variance; and (3) extended statistical analysis to small samples…" (Tankard, 1984, p. 99). Although *t* tests are one of the cornerstones of modern statistics, Gosset's work was not greeted with enthusiasm. Fisher, the originator of the analysis of variance, described the reaction of colleagues as "weighty apathy" (Fisher, 1939, p. 5), and Cochran stated that "the *t* distribution did not spread like wildfire" (Cochran, 1976, p. 13). Even Gosset underestimated the impact that his discoveries would have, as he wrote to Fisher, "I am sending you a copy of Student's Tables as you are the only man that's ever likely to use them!" (Gosset, 1970; Letter 11).

An interesting aspect of Gosset's work is that he used a pseudonym when publishing; he took the name Student. Not wanting the competition to know of its scientific work, Guinness forbade their scientists from publishing. As a result, Gosset secretly published all his articles under the name of "Student." It is for this reason that the *t* test is also known as "Student's *t* test."

Gosset remained with Guinness until his death, assuming the position of head brewer a few months before he died in 1937.

9.2 The Independent-Samples *t* Test

The independent-samples *t* test is used when two samples of participants provide scores on a measure. The *t* test compares the means of the two samples. The ultimate goal, however, is not to determine whether the means of the two *samples* are different, but rather to make an inference about whether the *population* means from which the samples are taken are different. Figure 9.1 depicts this arrangement.

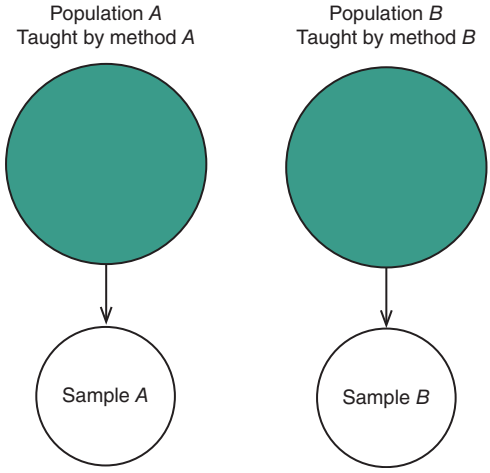


Figure 9.1 An independent-samples *t* test uses two samples. Do the samples come from the same or different populations?

Suppose we decide to conduct a study on the effects of stress. Based on previous research, we hypothesize involvement in an exercise regimen to have a beneficial effect. An experimental group receives 10 weeks of aerobic training, whereas a control group does not receive any aerobic training. After 10 weeks, both groups are asked to solve a series of simple mental arithmetic problems. Participants are told to work as quickly as possible and that an electric shock will be experienced if performance is inadequate. The measure of stress is the participants' heart rate during the task. The experimental hypothesis is that those participants who participated in the exercise program will show lower heart rates under stress in comparison with the control participants.

The two samples represent two hypothetical populations. One sample represents the hypothetical population of all of the participants who theoretically could have participated in the exercise program, the hypothetical population of treated participants. The control sample represents the hypothetical population of untreated participants.

The Null and Alternative Hypotheses

As with all inferential tests, when performing an independent-samples *t* test, the investigator specifies the null and alternative hypotheses beforehand. Recall that in single-sample research, the null hypothesis is stated as an equality, μ = some value known ahead of time, and the alternative hypothesis is stated as an inequality, $\mu \neq$ that same value. Similarly, in research involving two samples, the null hypothesis is stated as an equality, and the alternative hypothesis is expressed as an inequality. The null can be stated in two ways:

$$H_0 : \mu_1 = \mu_2 \text{ or } H_0 : \mu_1 - \mu_2 = 0$$

Note that these two statements are equivalent. One states that the two means are equal, and the other states that there is no difference between the two means. The alternative hypothesis can also be stated in two ways, depending on how we choose to state the null hypothesis:

$$H_1 : \mu_1 \neq \mu_2 \text{ or } H_1 : \mu_1 - \mu_2 \neq 0$$

In the exercise and stress study, the research hypothesis is that exercise will reduce participants' physiological reactions to a stressor. The statistical hypotheses, reflected in the null and alternative hypotheses, are statements that the samples are either taken from one population (no treatment effect) or come from two populations with different means (a treatment effect).

One final note: the subscripts of 1 and 2 are typically used to represent the two populations being compared. Researchers are free, however, to use other subscripts that may more specifically communicate the nature of the population in question, for example, $\mu_{control}$ and μ_{exp}, or μ_{drug} and $\mu_{placebo}$, or even letters such as μ_A and μ_B. These are all appropriate.

The Sampling Distribution for an Independent-Samples *t* Test

Theoretically Constructing the Sampling Distribution

The sampling distribution for an independent-samples *t* test is a sampling distribution of the *difference between independent sample means*. In Chapter 8, theoretical sampling distributions were constructed for single-sample *t* tests where one population was repeatedly sampled. In the present case, two populations will be used for sampling.

Returning to the aerobic exercise and stress study, the two hypothetical populations are "treated" and "untreated" participants. We sample from these two populations in the sense that one of our samples actually receives the exercise treatment and the second sample does not. Theoretically, to construct the sampling distribution, we would run the study, compute the mean heart rate under stress for each sample, and *take the difference between the group means*. This process would be repeated a near-infinite number of times. In each instance, the value $M_1 - M_2$ would be computed and included in the frequency distribution. (The subscripts refer to sample 1 and sample 2, respectively.) The result would be a sampling distribution that corresponds to the sample sizes used in the study. It is not a requirement that the sample sizes be the same; n_1 does not have to equal n_2 (though the test does assume they are at least similar). However, sampling distributions are built on situations where the repeated sampling of a given population is the same. Therefore, even though n_1 and n_2 do not need to be equal, n_1 needs to be the same size for each sample in the sampling distribution (as does n_2). This means there exists a large family of theoretical

sampling distributions, each distribution corresponding to a particular combination of sample sizes (degrees of freedom).

Characteristics of the Sampling Distribution

Because statisticians have worked out the characteristics of all sampling distributions, we have been spared the impossible task of constructing them. The following list clarifies the characteristics of sampling distributions of differences between independent means:

1) The mean of the sampling distribution of $M_1 - M_2$ is equal to the difference between the population means, $\mu_1 - \mu_2$. If the null hypothesis is true, that is, there is no difference between μ_1 and μ_2, then the mean of the sampling distribution of differences between means is 0. (M_1 would be just as likely to be larger than M_2 as it would be to be smaller than M_2 for all particular pairs of samples gathered.) If the means of the populations differ by, say, 10 units, then the mean of the sampling distribution of differences is 10.
2) The central limit theorem holds for sampling distributions of mean differences. If the populations of raw scores are normally distributed, the sampling distribution will likewise be normal. However, if the sample sizes are sufficiently large, the sampling distribution will be normal even if the populations of raw scores are not.
3) When two populations have the same variance (**homogeneity of variances**), and the independent-samples *t* test assumes that they do, then the standard deviation of the sampling distribution is given by Formula 9.1. The standard deviation is called the **standard error of the difference**, or simply the standard error.

The Standard Error of the Sampling Distribution of $M_1 - M_2$

Formula 9.1 is the standard error of the sampling distribution of differences. It describes the relationship between the amount of variability in the population and the variability of the sampling distribution of differences between means. Recall from Chapter 8 that the denominator of the single-sample *t* test is the estimated standard error of the sampling distribution of means, $t = (M - \mu)/s_M$, where $s_M = s/\sqrt{n}$.

Standard error of the difference, $\sigma_{M_1 - M_2}$

$$\sigma_{M_1 - M_2} = \sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad \text{(Formula 9.1)}$$

where

n_1, n_2 = the sample sizes of the two samples

σ^2 = the variance of either one of the population distributions. Since it is assumed that $\sigma_1^2 = \sigma_2^2$ (homogeneity of variances), it does not matter which variance is used

The Estimated Standard Error of the Difference

The formula for the independent-samples t test is presented in this section. The denominator of the independent-samples t test is the estimated standard error of the difference, symbolized as $s_{M_1 - M_2}$.

When σ is known, Formula 9.1 is the standard error of the difference. When σ is unknown, which is usually the case, then s is used to estimate σ. The estimated standard error of the difference is similar in form to the formula for the standard error when σ is known. In Formula 9.2, a new term is introduced, pooled variance, symbolized as s_p^2.

Definitional formula for the estimated standard error, $s_{M_1 - M_2}$

$$s_{M_1 - M_2} = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad \text{(Formula 9.2)}$$

where

s_p^2 = the pooled variance

Since it is assumed that $\sigma_1^2 = \sigma_2^2$, sample estimates of either population variance can be used to estimate σ^2. Since there are two sample variances, there are two estimates of σ^2. To generate the most accurate estimate of σ^2, a weighted average of the two sample variances will be used. This weighted average of variances is called the **pooled variance**.

Pooled variance

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2} \quad \text{(Formula 9.3)}$$

Each variance in Formula 9.3 is multiplied by its degrees of freedom. Variances from larger samples are weighted more than variances from smaller samples. This is how the "weighting" of the two samples is accomplished. Substituting the formula for the pooled variance into the formula for the estimated standard error gives Formula 9.4.

Variance formula for the estimated standard error, $s_{M_1 - M_2}$

$$s_{M_1 - M_2} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad \text{(Formula 9.4)}$$

Formula 9.4 shows that the estimated standard error of the difference combines the variances of both groups (samples) in the study. In addition, since Formula 9.4 is the denominator of the t ratio, it can be used, for instance, to analyze the results of a published study in which s^2 or s is reported in the article.

When working from raw data, Formula 9.5, the computational formula for $s_{M_1-M_2}$, is easier to use.

Computational formula for $s_{M_1-M_2}$

$$s_{M_1-M_2} = \sqrt{\frac{\left(\Sigma X_1^2 - (\Sigma X_1)^2/n_1\right) + \left(\Sigma X_2^2 - (\Sigma X_2)^2/n_2\right)}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \text{(Formula 9.5)}$$

Hypothesis Testing and the Sampling Distribution of Differences

Hypothesis testing operates statistically at the level of the sampling distribution. The sampling distribution is a theoretical tool that allows researchers to determine if the difference between two sample means is unlikely to be the result of sampling error and therefore more likely the result of sampling from two different populations. As the observed difference between the sample means increases, it becomes less and less likely that this difference is only due to sampling error, leaving us with the growing probability that there is a real difference between the populations.

Now assume that a sampling distribution of means from *each* population is theoretically established. Furthermore, assume that the population means are the same, both populations have identical standard deviations, and the two sampling distributions are based on repeated samples of the same size. Would there be a difference between the two sampling distributions? No, they would have the same means and standard errors. If we drew them on a graph, they would overlap so that it would look like one sampling distribution.

Now assume that the populations have different means. The sampling distributions would not show a perfect overlap. The mean of each sampling distribution would be the same as the mean of its respective population. As the difference between population means increases, the sampling distributions diverge. The sampling distribution of mean differences is a way of combining the two sampling distributions; it manages to take the difference between the two sampling distributions. If the means of the two populations are the same, the mean of the sampling distribution of differences will be 0. As the size of the difference between the population means increases, the mean of the sampling distribution of differences departs from 0. The null hypothesis assumes there is no difference between the population means; therefore, the mean of the sampling distribution of differences is 0.

In Figure 9.2a, the sampling distributions from identical populations overlap perfectly. The mean of each sampling distribution is the same as the mean of each population, 25. The sampling distribution of differences has a mean of 0, the difference between μ_1 and μ_2. In Figure 9.2b, the two sampling distributions are taken from different populations, one with $\mu_1 = 25$ and a second population with $\mu_2 = 27$. The sampling distribution of differences has a mean of 2, the difference between 27 and 25. (The sign of the difference can be ignored

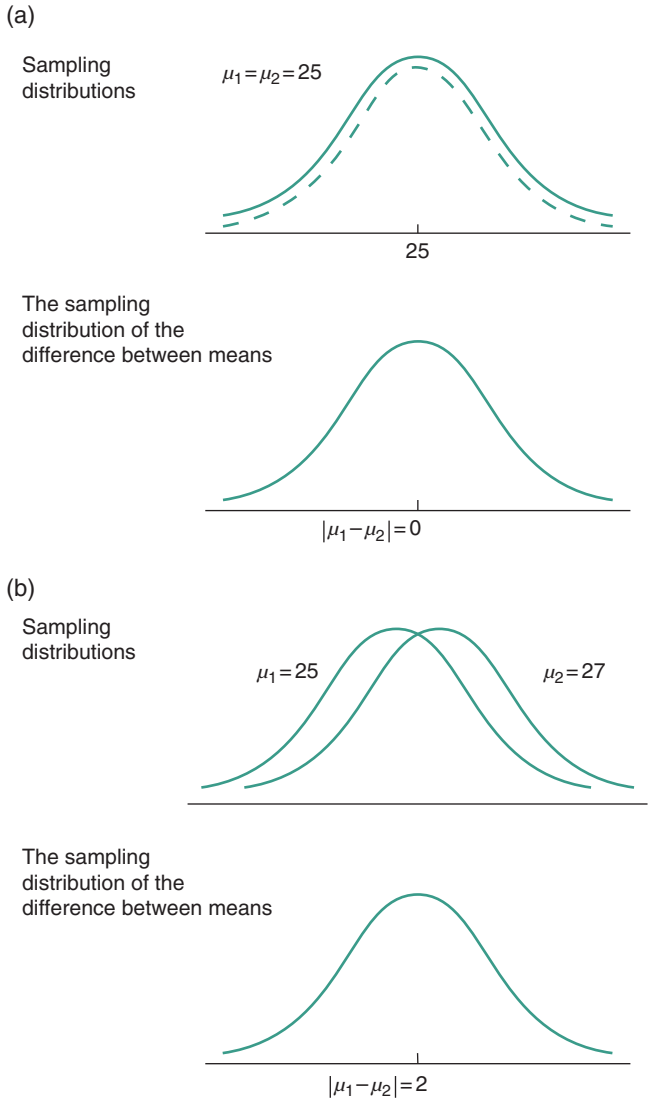


Figure 9.2 The sampling distribution of differences has the same mean as the difference between the means of the sampling distributions of the two populations.

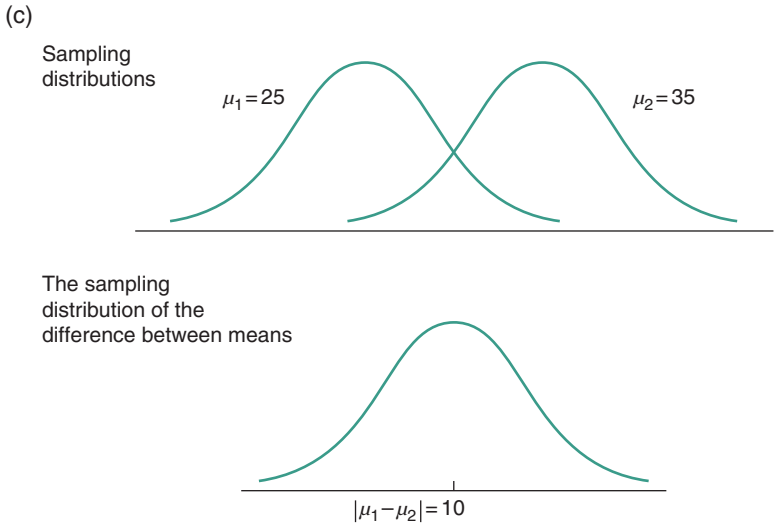


Figure 9.2 (Continued)

since, at this point, the size of the difference between μ_1 and μ_2 is what is important.) In Figure 9.2c the sampling distributions are generated from populations with $\mu_1 = 25$ and $\mu_2 = 35$. As a result, the sampling distribution of differences has a mean of 10.

From the Sampling Distribution of Differences to the *t* Distribution for Independent Samples

The sampling distribution of mean differences, although theoretically of fundamental importance, is not directly used to conduct the inferential test. What the researcher would like to determine is if the difference between the sample means is so unusual as to suggest they came from a distribution of sample mean differences that does not have a mean of 0. If the obtained mean difference, when plotted on the relevant *t* distribution, falls in one of the tails, then the validity of the null hypothesis is questioned. The *t* ratio is a formula that indicates the distance the difference between sample means is from 0 within a sampling distribution. The logic of the independent-samples *t* test follows from the discussion of *z* scores and the *z* test.

In Chapter 5, the *z* score formula was used to indicate the distance a score is from the mean of the raw score distribution. In Chapter 8, we learned that the *z* statistic provides a measure of how far a sample mean, in standard error units, is from the mean of the sampling distribution.

The *z* statistic requires that we know σ. If σ is unknown, the *t* statistic is used to transform all the scores of the sampling distribution into a *t* distribution. By using the single-sample formula for *t,* we were able to determine how many

standard errors a sample mean was from the hypothesized mean of the sampling distribution.

Whether we perform a z test or a single-sample t test, a ratio of the difference between two means and the standard error is obtained:

$$z_{obt} = \frac{M - \mu}{\sigma_M}$$

$$t_{obt} = \frac{M - \mu}{s_M}$$

With the (estimate of the) standard error in the denominator, the ratio indicates the number of (estimated) standard error units the sample mean is from the hypothesized mean. The independent-samples t statistic is also a ratio that specifies the distance (in estimated standard error units) between the sample mean and the hypothesized mean of the sampling distribution of differences.

The *t* Ratio

Formula 9.6 is the formula for the independent-samples t test.

t statistic for independent samples

$$t_{obt} = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{M_1 - M_2}} \quad \text{(Formula 9.6)}$$

First, consider the numerator of the t ratio in Formula 9.6. To understand why it looks as it does, it is helpful to recall the t ratio for a single-sample t test found in Chapter 8:

$$t_{obt} = \frac{M - \mu}{s_M}$$

In the numerator of this formula, there is an obtained statistic, M, which is contrasted with a hypothesized parameter, μ. The value of μ is the mean of the sampling distribution of means if the null hypothesis is true. When the null hypothesis is true, the difference $M - \mu$ will be close to 0. The formula for an independent-samples t test also contrasts a hypothesized population parameter with an obtained statistic. When the null hypothesis is true, the population means are the same ($\mu_1 - \mu_2 = 0$). The sampling distribution of differences will then have a mean of 0. The obtained sample statistic is $M_1 - M_2$. Therefore, we are contrasting an obtained difference of sample means with a hypothesized difference of population means. As the obtained difference between sample means departs from the hypothesized difference between population means, we begin to question the null hypothesis of no difference. Since the hypothesized difference between population means is almost always 0, the expression $\mu_1 - \mu_2$ can be dropped. As a result, the form of the t ratio used for an independent-samples t test is given in Formula 9.7.

The *t* ratio

$$t_{obt} = \frac{M_1 - M_2}{s_{M_1 - M_2}} \quad \text{(Formula 9.7)}$$

The *t* Distributions for the Independent-Samples *t* Test and Degrees of Freedom

The *t* distribution for the independent-samples *t* test is a transformation of a sampling distribution of differences between means. It is symmetric and has a mean of 0. The *t* ratio specifies the number of standard errors the obtained difference between sample means is from 0, the null hypothesized difference between population means.

There is a different sampling distribution for every combined sample size. One study may have two samples with 10 participants each, another study may have one sample with 20 and the other with 15, and so on. Each situation yields a different theoretical sampling distribution that can be transformed into a *t* distribution. Therefore, there is a family of *t* distributions, each with its own degrees of freedom. For an independent-samples *t* test, the degrees of freedom associated with the *t* distribution are $n - 1$ from the first sample and $n - 1$ from the second sample, or $n_1 + n_2 - 2$.

Table 9.1 summarizes the differences between the formulas for a *z* score, *z* statistic, single-sample *t* statistic, and independent-samples *t* statistic. The purposes

Table 9.1 A summary comparison of the transformation formulas for a *z* score, *z* statistic, single-sample *t* statistic, and independent-samples *t* statistic.

***z* score**

$$z = \frac{X - \mu}{\sigma}$$

Purpose
Transforms raw scores into a *z* distribution. A *z* score indicates the number of standard deviations a raw score is from the mean of the raw score distribution.

Distribution characteristics
The *z* distribution has a mean of 0 and a standard deviation of 1. It is distributed normally if the population is normally distributed.

***z* statistic**

$$z_{obt} = \frac{M - \mu}{\sigma_M}$$

Purpose
Transforms a sampling distribution of means into a distribution of *z* values. The z_{obt} indicates the number of standard error units a sample mean is from the hypothesized mean of a sampling distribution.

(*Continued*)

Table 9.1 (Continued)

Distribution characteristics
If the sampling distribution is normal, and since σ is known, the distribution of z statistics will be normally distributed, with a mean of 0 and a standard deviation of 1.

Single-sample *t* statistic

$$t_{obt} = \frac{M - \mu}{s_M}$$

Purpose
Transforms a sampling distribution of means into a t distribution. The t_{obt} indicates the number of standard error units a sample mean is from the hypothesized mean of the sampling distribution.

Distribution characteristics
All t distributions are symmetrical with a mean of 0. Each one has $n - 1$ degrees of freedom. As the sample size of the sampling distribution increases, the t distribution approximates a standard normal distribution.

Independent-samples *t* statistic

$$t_{obt} = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{M_1 - M_2}} \quad \text{or} \quad t_{obt} = \frac{M_1 - M_2}{s_{M_1 - M_2}}$$

Purpose
Transforms a sampling distribution of differences between means into a t distribution. The t_{obt} indicates the number of standard errors the difference between sample means is from the hypothesized difference between population means, typically 0.

Distribution characteristics
All t distributions are symmetrical with a mean of 0. Each t distribution is distributed with $n_1 + n_2 - 2$ degrees of freedom. As the sample size of the sampling distribution increases, the t distribution approximates a standard normal curve.

of the formulas and the underlying distributions of the formulas are presented. In addition, the degrees of freedom are specified for the z test and the t tests.

An Example of Hypothesis Testing Using the Independent-Samples *t* Test

With the logic of the t statistic in place and the requisite formulas provided, we are now ready to work through a problem and decide whether to reject the null hypothesis. Keep in mind that we are making an inference about whether the means of two populations are unequal. In the context of an experiment, deciding that the population means are *not* equal is tantamount to claiming the independent variable has an effect on the dependent variable, assuming no confounds. Consider once again the study about the effectiveness of aerobic conditioning on participants' ability to tolerate stress.

Worked Example

Twenty participants are randomly assigned to an experimental condition ($n_1 = 10$) and a control condition ($n_2 = 10$). The experimental participants exercise three times a week for 10 weeks. During each workout, they walk a

treadmill for 20 minutes while their heart rate is maintained between 160 and 180 beats per minute. The control participants do not exercise during the 10-week period. After 10 weeks, all participants are brought into the lab and asked to solve mental arithmetic problems under the threat of electric shock for poor performance. The measure of stress is the participants' heart rate during the task. The experimental hypothesis is that, during stress, the aerobic group will have a lower heart rate than the control group. The null hypothesis is that the population means are the same. The data, along with the computation of t_{obt}, are presented in Table 9.2.

Six Steps for Testing the Null Hypothesis Using the Independent-Samples *t* Test

Step 1. Define the null and alternative hypotheses:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Step 2. Set alpha. Alpha is set at .05.

Step 3. Compute t_{obt} (see Table 9.2).

Step 4. Locate t_{crit} in Table A.2. Recall, the degrees of freedom formula for this problem is $n_1 + n_2 - 2$. For this problem, this means 10 + 10 – 2 = 18. Enter the left column of the *t* table and locate the number 18. Now move to the column under alpha of .05 for a two-tailed test. The critical values are ±2.10.

Step 5. Compare the t_{obt} of –2.23 with the critical value of ±2.10. Since t_{obt} falls outside of ±2.10, the null hypothesis is rejected.

Step 6. Interpret the findings. Statistical evidence suggests that aerobic training leads to a reduction in the participants' heart rate when solving mental arithmetic problems under threat of electric shock for poor performance, $t(18) = -2.23$, $p < .05$. Additional research would be required to determine if other kinds of stressors could be managed as well through aerobic training.

Table 9.2 Computing t_{obt} for an independent-samples *t* test.

Aerobic training	Control
84	88
78	97
67	74
87	80
80	87
78	90

(*Continued*)

Table 9.2 (Continued)

Aerobic training	Control
78	90
79	86
82	84
81	78
$M_1 = 79.40$	$M_2 = 85.40$
$s_1 = 5.25$	$s_2 = 6.69$
$\Sigma X_1 = 794$	$\Sigma X_2 = 854$
$\Sigma X_1^2 = 63\,292$	$\Sigma X_2^2 = 73\,334$
$n_1 = 10$	$n_2 = 10$

$$t_{obt} = \frac{M_1 - M_2}{\sqrt{\frac{\left(\Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1}\right) + \left(\Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n_2}\right)}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$t_{obt} = \frac{79.4 - 85.4}{\sqrt{\frac{\left(63\,292 - \frac{(794)^2}{10}\right) + \left(73\,334 - \frac{(854)^2}{10}\right)}{10 + 10 - 2}\left(\frac{1}{10} + \frac{1}{10}\right)}}$$

$$t_{obt} = \frac{79.4 - 85.4}{\sqrt{\frac{(63\,292 - 63\,043.6) + (73\,334 - 72\,931.6)}{10 + 10 - 2}(0.2)}}$$

$$t_{obt} = \frac{-6}{\sqrt{(650.8/18)(0.2)}}$$

$$t_{obt} = \frac{-6}{\sqrt{7.23}}$$

$$t_{obt} = \frac{-6}{2.69}$$

$$t_{obt} = \mathbf{-2.23}$$

Raw scores are heart rates under stress. Since t_{obt} falls outside of $\pm t_{crit}$, the null hypothesis is rejected.

Box 9.1 presents a study of the effects of relaxation training on the frequency of epileptic seizures. An independent-samples *t* test is used to determine if the change in the number of seizures between groups is due to the experimental manipulation of relaxation.

Box 9.1 Can Epileptic Seizures Be Controlled By Relaxation Training?

In the last 50 years, researchers have begun to discover an association between emotionality and frequency of seizures among people suffering from epilepsy (e.g. Baslet, 2011). For instance, daily hassles, fear, anger, and anxiety have all been found to correlate with seizure activity (Feldman & Paul, 1976; Standage, 1972; Symonds, 1970; Temkin & Davis, 1984). At least one study has shown that symptoms of anxiety are twice as high in an epileptic sample compared with those in persons with other kinds of physical problems (Standage & Fenton, 1975).

The standard medical treatment for epileptic seizures is the administration of an antiseizure medication, like Dilantin or phenobarbital. Although the *causal* role of emotionality in producing seizures is debatable (seizures could cause negative emotions), the association between emotions and seizures leads researchers to wonder if perhaps a stress-reduction treatment could reduce the frequency of epileptic seizures.

Puskarich (1988) compared the effects of relaxation training versus a placebo on the frequency of seizures among epileptics. After an eight-week baseline period in which all patients recorded their frequency of seizures, 13 patients received six weeks of relaxation training. Eleven placebo-control patients were seen the same number of times as the experimental participants but were placed in a room alone and told that sitting quietly would induce relaxation, which would help reduce their seizures. All patients took their usual medications throughout the study. For eight weeks post-treatment, all patients recorded their frequency of seizures. The dependent variable was the change in the number of seizures from baseline through post-treatment.

The raw data are presented in the following table. A negative number means a decrease in seizures; a positive number indicates an increase in seizures. Because this was a relatively new area of research, and because it is important to minimize the probability of failing to reject a false null hypothesis (a Type II error), the investigator sets alpha at .10 when conducting the statistical test. By not using the traditional alpha of .05, the investigator doubles the probability of a Type I error.[2] An independent-samples *t* test is used to compare the means of the relaxation and placebo conditions.

2 If Puskarich submits the article for publication, it will have to include a convincing case for "relaxing" the alpha level. Scientists are a conservative group. In most cases, they would prefer to miss something that is there (Type II error) than think that they have found something that is not there (Type I error).

Relaxation	Placebo
+7	−1
0	−5
−5	+16
−1	−7
−14	+3
−5	+4
−6	−6
−7	−2
−1	+2
−4	−4
−7	−3
−13	
−8	
$M_1 = -4.92$	$M_2 = -0.27$
$s_1 = 5.51$	$s_2 = 6.51$
$n_1 = 13$	$n_2 = 11$

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

$\alpha = .10$

$df = n_1 + n_2 - 2 = 22$

$$t_{obt} = \frac{M_1 - M_2}{s_{M_1 - M_2}}$$

Since the s_1 and s_2 are provided in the summary statistics, Formula 9.4 can be used to compute the standard error:

$$s_{M_1 - M_2} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$s_{M_1 - M_2} = \sqrt{\frac{(5.51)^2(12) + (6.51)^2(10)}{13 + 11 - 2}\left(\frac{1}{13} + \frac{1}{11}\right)}$$

$$s_{M_1 - M_2} = \sqrt{(35.82)(0.17)}$$

$$s_{M_1 - M_2} = \sqrt{6.09}$$

$$t_{obt} = \frac{-4.92 - (-0.27)}{\sqrt{6.09}} = \frac{-4.65}{2.47}$$

$$t_{obt} = \mathbf{-1.88}.$$

The critical values for t are $t_{.10}(22) = \pm 1.72$. Since t_{obt} falls outside of $\pm t_{crit}$, the null hypothesis is rejected. The author concluded that relaxation training is more effective than a placebo treatment in reducing the frequency of seizures, $t(22) = -1.88$, $p < .10$.

A Measure of Effect Size: Cohen's *d*

One secondary question that can be asked when a null is rejected is the size of the treatment effect. The t_{obt} value is not designed to measure effect size, but rather the likelihood that an effect exists, that is, the *certainty* of an effect, not the *size* of an effect. These are related concepts, but not identical. As noted in the previous chapter, a simple, direct, and often-used measure of effect size is Cohen's *d*. The formula is as follows.

Cohen's *d* for independent-samples *t* test

$$d = \frac{estimated\ mean\ difference}{estimated\ standard\ deviation} = \frac{M_1 - M_2}{\sqrt{s_p^2}} \quad \text{(Formula 9.8)}$$

where

$\sqrt{s_p^2}$ = the pooled standard deviation

$M_1 - M_2$ is used as the best estimate of the mean difference between the two populations, and the pooled standard deviation is used as the best estimate of the population standard deviation (assuming equal variances; see assumptions for the independent-samples *t* test). Notice this only allows us to *estimate* the effect size. Nonetheless, this statistic is typically symbolized simply as *d*. As with previous version of the statistic, Cohen's *d* reflects the difference between the means in standard deviation terms. Larger *d* values reflect larger effect sizes.

9.3 The Appropriateness of Unidirectional Tests

Up to this point, hypothesis testing has been discussed from a nondirectional perspective, which means that our interest is in detecting differences between population means, irrespective of whether $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$. The null hypothesis has always been stated as an equality (i.e. $\mu_1 = \mu_2$ or μ=some known value), and the alternative hypothesis has always been stated as an inequality (i.e. $\mu_1 \neq \mu_2$ or μ= some known value).

Research hypotheses (scientific hypotheses) are usually stated as predictions about the expected direction of an effect or relationship. For example, persuasion technique *A* will induce *greater* attitude change than persuasion technique *B*; participants' perceptions of control over a stressor will *decrease* stress reactions; or higher levels of physiological arousal will create *stronger* emotions. Researchers, however, typically frame their *statistical* hypotheses in a nondirectional form. In other words, even though the research hypothesis makes a prediction about which of two means will be larger, the null and alternative hypotheses allow the investigator to discover if a treatment effect or relationship

between variables is opposite to the predicted effect. For instance, if a researcher hypothesizes that an advertisement will *increase* the sales of a product, a nondirectional test of this hypothesis allows for the discovery that the advertisement actually *decreases* sales. However, there are times when researchers have chosen to use a directional test of the research hypothesis. A directional test is capable of detecting *only* a difference between means in one direction. If a researcher uses a directional test to see if an advertisement increases sales, then it will be blind to the possibility that the advertisement decreases sales. As we will soon learn, the decision to adopt a directional versus nondirectional test has implications for how the statistical hypotheses are stated and how to use the *t* table. In addition, we will learn why directional tests of hypotheses are very controversial.

We begin the discussion of directional tests by specifying how they affect the way the null and alternative hypotheses are stated; then the implications that a directional test has for the critical value that is used are addressed; and finally issues associated with *if* and *when* to use a directional test are discussed. As a vehicle for presenting the concepts of directional and nondirectional tests, we will use the research context in which the independent-samples *t* test is appropriate. Note: A nondirectional test is called a two-tailed test and a directional test is called a one-tailed test. The reason for the two-tailed/one-tailed terminology will become clear as we read the next section.

One-Tailed and Two-Tailed Tests

The discussion of hypothesis testing up to this point has addressed only two-tailed tests. The null and alternative hypotheses for a two-tailed test have been presented as follows:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

When the null and alternative hypotheses are stated in this fashion, the rejection region is divided equally between both tails of the *t* distribution. For example, when testing at the 5% alpha level, 2.5% of the rejection region is in both the right tail and the left tail of the *t* distribution. With the rejection region in both tails, it is possible to detect a difference when $\mu_1 > \mu_2$ *as well as* when $\mu_1 < \mu_2$. Since tests with a rejection region in both tails can detect a difference between population means in either direction, they are called **two-tailed** (or **nondirectional**) **tests**.

Two things change when conducting a **one-tailed** (or **directional**) **test**: the manner in which the null and alternative hypotheses are stated and the placement of the rejection region. (There are no new formulas associated with a one-tailed test.) Consider the following research situation. Suppose a standard drug for treating some illness exists, but this drug has a certain number of side effects. A new, more expensive drug is being tested; despite its cost, if it has fewer side effects, it will become the treatment of choice. Since the new drug

is more expensive than the standard (old) drug, the old drug will remain the treatment of choice if they both have the same number of side effects. The same conclusion would be reached if the new drug had *more* side effects than the old drug. In this way, whether the new drug has the same or more side effects makes no difference; either way it will not be marketed. The only finding of interest is if the new drug has fewer side effects.

Recall that when setting up the null and alternative hypotheses, two related rules must be followed. Together, the null and alternative hypotheses must be mutually exclusive and collectively exhaustive. The null and alternative hypotheses for a directional test can be stated as

$$H_0: \mu_{new} \geq \mu_{old}$$
$$H_1: \mu_{new} < \mu_{old}.$$

These hypotheses are collectively exhaustive. The null hypothesis, H_0, includes the case in which the population means are the same *and* the case in which the mean of the old drug is lower. The alternative hypothesis specifies that the mean number of side effects for the new drug is less than the mean of the old drug. The null and alternative hypotheses cover all of the possible outcomes; therefore, they are collectively exhaustive. They are also mutually exclusive because they both cannot be true at the same time.

In this example, note that the difference $\mu_{new} > \mu_{old}$ is embodied in the *null hypothesis*. This is a bit of a misnomer since "null" implies no difference. (This will be discussed further later on.) Right now, it is important to note that even if it is true that $\mu_{new} > \mu_{old}$, there is no way a directional test can come to this conclusion; remember, we can only reject the null or fail to reject the null.

Setting up the null and alternative hypotheses as a directional test determines the placement of the rejection region in the t distribution. Since only a difference between means in one direction can be detected, many researchers feel justified in placing the entire rejection region in this one tail of the distribution. If alpha is set at .05, the entire 5% of the rejection region is placed in one tail. Whether the region is in the left tail or the right tail depends on the direction of interest. In the drug example, we are only interested in detecting if the new drug has fewer side effects, that is, when the mean of the new drug group is smaller than the mean of the old drug group, $\mu_{new} < \mu_{old}$. As a result, the rejection region is to the left side of the t distribution. Figure 9.3 uses the t distribution to illustrate the rejection region for a two-tailed and one-tailed test when alpha is .05.

The Sign of t_{obt} is Important in a Directional Test

When conducting a nondirectional test, the researcher allows that $\mu_{new} > \mu_{old}$ or $\mu_{new} < \mu_{old}$. As a result, it does not matter if the t_{obt} is positive or negative in value. For this reason, we always state t_{crit} as ± when conducting a nondirectional test. The situation is different with a directional test, whether t_{obt} is

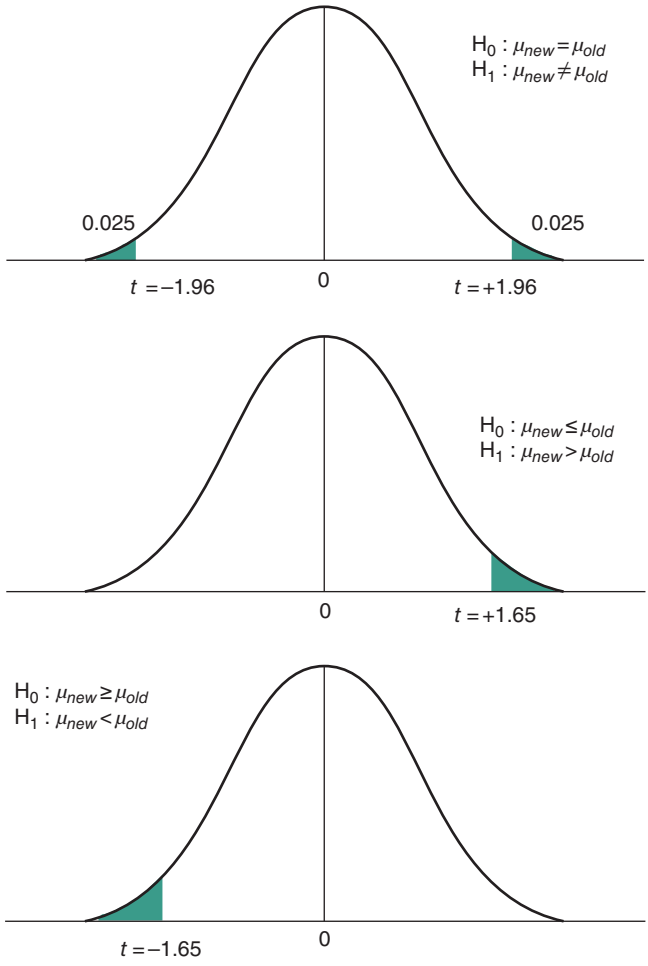


Figure 9.3 The rejection region of a *t* distribution for a one-tailed and two-tailed test, $\alpha = .05$.

positive or negative matters. If the alternative hypothesis states that $\mu_{new} < \mu_{old}$, the null hypothesis is only rejected if t_{obt} is a *negative* number equal to or greater than t_{crit} (in absolute value terms). This is because the rejection region is entirely to the left of the mean of the *t* distribution; all *t* values to the left of the mean are negative. If our alternative hypothesis is stated as $\mu_{new} > \mu_{old}$, then t_{obt} must be a *positive* number equal to or greater than t_{crit} for the null to be rejected. Recall that when we are conducting a two-tailed test, it does not matter how we arrange the means in the numerator. However, the sign of t_{obt} is important in a one-tailed test, so we must be careful to arrange and subtract means in the numerator properly.

Using the *t* Table for a Directional Test

We already have experience in finding t_{crit} for a two-tailed *t* test: Enter the *t* table, find the appropriate *df*, and look under the alpha heading for a two-tailed test. As we look at the *t* table (Table A.2), note that it has two rows of headings, one for alpha levels for a one-tailed test and one heading for alpha levels used for a two-tailed test. Notice for a given level of alpha, the critical value used in a one-tailed test is closer to zero than the value used in a two-tailed test. Therefore, a decision to reject the null is more likely to occur, if the mean difference is in the predicted direction.

	α levels for two-tailed test	
df	*.10*	*.05*
	α levels for one-tailed test	
	.05	.025
⋮	⋮	⋮
6	1.943	2.447
7	**1.895**	2.365
8	1.860	2.306
9	1.833	2.262
⋮	⋮	⋮

■ **Question** *Using alpha of .05, what is the critical value for an independent-samples one-tailed t test if there are four participants in one group and five participants in the other group?*

Solution The *df* for an independent-samples *t* test is $n_1 + n_2 - 2$, or $4 + 5 - 2 = 7$. The critical value is either +1.895 or −1.895, depending on the predicted direction. ■

Use Caution in Deciding to Conduct a One-Tailed Test

Research hypotheses are almost always framed as a directional prediction. Therefore, why not always use a directional *t* test? Well, if we use a one-tailed test, we will be unable to detect differences between population means opposite of the predicted direction. Unexpected findings, however, may be theoretically important. Contrary results may lead to a revision in or rejection of the theory being explored. Performing a one-tailed test implies that the researcher believes findings in one direction are meaningless. However, just because the results of a

study may not fit current theory does not mean they are meaningless. After all, the history of science is littered with stories of unexpected yet truly meaningful findings being uncovered without prior intention. Unexpected findings should stimulate future research. For this reason, a researcher should never use a one-tailed test simply because the research hypothesis is stated directionally.

The strongest justification for using a one-tailed test occurs when a new course of *action* is to be taken *only* if the result of the test is in one direction. We might think of this as a "theory versus application" issue. If the research is theoretically important, there is simply no debate; use a two-tailed test. Many feel, however, that if the main purpose behind the research project is one of application *and* if that application will only be taken if one type of finding is uncovered, then a direction test is justified. However, even in these situations the use of a one-tailed test is controversial (e.g. Lombardi & Hurlbert, 2009). If the purpose of the t test is to measure the likelihood that sampling error can explain the difference between means, then moving half of the rejection region from one side over to the other, thus correspondingly reducing the ability for sampling error to explain a mean discrepancy in the selected direction, misrepresents what we are trying to measure – the explanatory power of sampling error. If the null is actually true, getting a t value around 1.6 or 1.7 is simply much more likely than getting one around 2 or 2.1. Many argue the use of a one-tailed test results in a Type I error rate that is effectively greater than .05. If the critics are right, the use of directional tests constitute a form of statistical cheating, whether or not the primary purpose of the research project is one of application (for more on the debate over directional and nondirectional tests, see Kirk, 1972; Liberman, 1971).

One final point concerns *when* a decision is made to use a one-tailed test. If a researcher believes a one-tailed test is warranted, the decision *must* be made and, of course, justified before the data is gathered and analyzed. Suppose we conduct a two-tailed test and find that the t_{obt} is close, but does not fall in the rejection region. If we then move the entire rejection region over to one tail so that the t_{obt} value falls within it, we will be engaging in professional misconduct. Although we have claimed to use an alpha of .05, in actuality our alpha value is higher.

Despite these serious concerns, one-tailed tests continue to be used on occasion. For this reason, the tables in Appendix A include references to one-tailed critical values.

9.4 Assumptions of the Independent-Samples *t* Test

There are five assumptions that should be met when conducting an independent-samples t test. These are identical to the four assumptions for the single-sample t test with one additional assumption. The first two are methodological and the final three are statistical:

1) **Representativeness.** It is assumed that the samples are representative of the populations from which they are drawn. Random sampling is the best data gathering method to meet this assumption; however, other sampling methods might be sufficient. Meeting this assumption allows us to generalize from samples to populations.
2) **Independent observations.** Independent observations mean that each participant is contributing only one score and that those scores are not influenced by other participants. If a score from one participant in the study is influenced by the behavior of another participant, then the scores from these two participants are *not* independent.
3) **Interval or ratio scale of measurement**. The independent-samples t test utilizes means and standard deviations. These concepts only have meaning for data measured on a scale where the quantitative distance between integers is held constant, namely, an interval or ratio scale (see Chapter 2).
4) **The populations from which the samples are taken are normally distributed.** This assumption assures that the sampling distribution will be normally distributed. If a population is not normally distributed, the assumption of normality is violated *unless* the sample size is sufficiently large. (Recall that sampling distributions approximate a normal curve as n increases.) When sufficiently large, the resulting sampling distribution is normal despite the nonnormality of the raw scores.
5) **Homogeneity of variances.** The independent-samples t test assumes the variances of the two populations sampled are equal. An inferential test can be conducted on the sample variances to see if the population variances are unequal (see Kirk, 1989). In lieu of this test, a rule of thumb to judge homogeneity is to see if one of the sample variances is four times larger than the other. If so, the assumption of homogeneity is probably violated. The t test may also be robust to violations of this assumption, particularly if $n_1 = n_2$. However, when sample sizes are unequal and variances are quite discrepant, the t test should not be run. Thankfully, other tests can be used (see Chapter 18).

9.5 Interval Estimation of the Population Mean Difference

Recall that there are two kinds of inferential procedures, hypothesis testing and estimation. This chapter is focused on hypothesis testing, but we can also use sample means, the t distribution concept, and an estimate of the standard error to generate an interval estimation of the population mean difference. Further, we can quantify the confidence we have that this difference falls within that interval. Since each potential sample mean difference drawn has a corresponding t value, we can use the t distribution and our obtained sample mean difference (which is an unbiased estimate of the population mean difference) to generate a probability function for the value of the actual population mean

difference. Choosing t_{crit} values corresponding to different probabilities within the t distribution allows us to create intervals with differing degrees of certainty. The formula for an interval in which we can have 95% confident follows.

Confidence interval for a population mean difference for independent samples

$$LL = (M_1 - M_2) - t_{.05}s_{M_1 - M_2}$$
$$UL = (M_1 - M_2) + t_{.05}s_{M_1 - M_2} \quad \text{(Formula 9.9)}$$

where

LL = the lower limit of the confidence interval
UL = *the* upper limit of the confidence interval
$t_{.05}$ = the critical value for a t distribution of a given sample size

Since we are generating an interval, two values are calculated, one being the value at the lower end of the interval and the other at the upper end. As the interval widens and becomes less specific, the confidence grows that the actual mean difference falls within that window. A 95% confidence rate is typical, but the above formulas could easily be adjusted to find a 90 or 99% confidence interval simply by finding the corresponding t_{crit} values using the t table (Table A.2).

■ **Question** *Using the same data from the exercise and stress study explored earlier in the chapter, find the 95% confidence interval for the population mean difference in heart rates due to aerobic training. ($M_1 = 79.4$; $M_2 = 85.4$; $s_1 = 5.25$; $s_2 = 6.69$; $n_1 = 10$; $n_2 = 10$).*

Solution

Step 1. Identify the null and alternative hypotheses:

$$H_0: \ \mu_1 = \mu_2$$
$$H_1: \ \mu_1 \neq \mu_2$$

Step 2. Set the confidence rate at 95% (equivalent to $\alpha = .05$).

Step 3. Using the t table, find the cutoff values beyond which lie 2.5% in the right tail of the t distribution and 2.5% in the left tail of the distribution. The task is accomplished in the same way that we found t_{crit} for the t test. With $df = 9$ and $\alpha = .05$ (two-tailed test), the cutoff points are ±2.262.

Step 4. Compute the confidence interval.

$$LL = (M_1 - M_2) - t_{.05}s_{M_1 - M_2}$$
$$UL = (M_1 - M_2) + t_{.05}s_{M_1 - M_2}$$

From having worked the problem previously, we know

$$M_1 - M_2 = -6$$
$$s_{M_1 - M_2} = 2.69$$

$$LL = -6 - 2.262(2.69) = \mathbf{-0.08}$$

$$UL = -6 + 2.262(2.69) = \mathbf{12.08}$$

Step 5. Interpret the findings. Statistical evidence suggests that we can be 95% confident that aerobic training leads to a reduction in the participants heart rate when solving mental arithmetic problems under threat of electric shock for poor performance by somewhere between –0.08 and 12.08 beats per minute. ■

9.6 How to Present Formally the Conclusions for an Independent-Samples *t* Test

Proper reporting of inferential statistics can be challenging. Following are examples of how to report, in sentence form, a rejection of the null as well as a fail to reject the null. If rejecting the null, a sentence might read, "An independent-samples t test found evidence suggesting aerobic training leads to a reduction in participants' physiological reaction to stress, $t(18) = -2.23$, $p < .05$." If we also wanted to include a measure of effect size, the sentence could finish with "…$t(18) = -2.23$, $p < .05$, $d = 1.05$." If failing to reject the null, a sentence might read, "An independent-samples t test did not find evidence suggesting aerobic training leads to a change in participants' physiological reaction to stress, $t(18) = -1.62$, *n.s.*" For a more detailed analysis of the style, symbols, and punctuation used in these sentences, please see Section 8.8.

Summary

The independent-samples t test is used for between-participants designs when two separate samples of participants provide scores on a measure. The purpose of the t test is to help us decide whether the two samples come from the same or different populations.

The sampling distribution for an independent-samples t test is a distribution of the differences between independent sample means. Sampling distributions of differences between means have several characteristics. First, the mean of the sampling distribution of $M_1 - M_2$ is equal to the difference between the population means, $\mu_1 - \mu_2$; this is usually zero. Second, the central limit theorem holds for sampling distributions of the differences between means. If the populations are normally distributed, the sampling distribution will likewise be normal. However, if the sample sizes are sufficiently large, a sampling distribution of differences will be normal, whether or not the populations are normally distributed. Third, the standard deviation of a sampling distribution of differences is called the standard error of the difference or the standard error. The t test uses the weighted average of two sample variances (the pooled variance) to estimate the population variance, which is then used to estimate the standard error.

The independent-samples t test relies on the t distribution – a family of normal distributions that vary based on sample size. The t distribution transforms a sampling distribution of differences into a standardized curve by applying the t formula to each mean of the sampling distribution. The t distribution, represented in the t table (Table A.2), is used to identify the critical values that t_{obt} must equal or be greater than (in absolute value) to reject the null hypothesis. The t_{obt} indicates the number of standard errors the difference between sample means is from the hypothesized mean of the sampling distribution of differences. If t_{obt} is small, the proper decision is to fail to reject the null. Sampling error is a viable explanation for the differences between the two sample means. This does not mean the null is true, merely that the sample data is not allowing us to reject it. If t_{obt} is large, the proper decision is to reject the null, although there will be a chance (determined by the value of alpha) that a Type I error will be made. If a null hypothesis is rejected, Cohen's d can be calculated and used as a measure of effect size.

The assumptions of the independent-samples t test are representativeness, independent observations, interval or ratio measures, normally distributed populations, and the homogeneity of population variances.

Research hypotheses are usually stated as predictions about the expected direction of an effect or relationship. Statistical hypotheses, however, are typically presented in such a way as to detect a difference in either direction. Some researchers argue that there are times when it is appropriate to use a one-tailed or directional test of the research hypothesis, typically when dealing with a question of application. However, even in these situations the use of one-tailed test is problematic, controversial, and generally discouraged by statisticians. When conducting a one-tailed (or directional) test, the placement of the rejection region changes. The entire rejection region is placed in one tail. As a result, the critical value moves closer to zero, making a decision to reject the null more likely to occur if the mean difference is in the predicted direction. Many theorists believe that this procedure inflates the alpha value beyond the stated level.

The standard error, sample means, and t distribution can also be used to create a confidence interval for the actual value of the difference between the means of the two populations.

Using Microsoft® Excel and SPSS® to Run an Independent-Samples *t* Test

Excel

General instructions for data entry into Excel can be found in Appendix C.

Data Entry

Enter all of the scores from the samples into two adjacent columns, one sample in each column. Label the columns appropriately.

Data Analysis

1) Excel has built-in programs for many inferential tests, including the independent-samples *t* test. To access it, click on the **Data** tab on the top menu, and then click **Data Analysis**. If this option is not found, the Data Analysis ToolPak needs to be installed. See Excel instruction materials for how to install this feature.
2) With the Data Analysis box open, select **t-Test: Two-Sample Assuming Equal Variances**. (Do not select a similar option, **t-Test: Two-Sample Assuming Unequal Variances**. This function does not create a pooled variance.)
3) Input the data range for one variable in box **Variable 1 Range**. Input the data range for the other variable in box **Variable 2 Range**. (If the labels were included in the range, make sure to click the **Labels** box to exclude those cells.)
4) Decide on an Output option. The default is to place it on a separate worksheet.
5) Click **OK**.
6) The output box will present the means, variances, and observations (sample sizes). Additionally presented will be the pooled variance, hypothesized mean difference (0, unless otherwise specified), degrees of freedom (labeled df), observed *t* value (labeled "t stat"), and the critical scores and probabilities for both one- and two-tailed versions of the test. Compare t stat with either the probability value (labeled **P(T<=t) two-tail**) or the critical score (labeled **t Critical two-tail**) associated with the two-tailed test to make a decision regarding the null hypothesis. (See Figure 9.4 for a worked example.)

Control	Drug
2	16
0	20
10	2
7	22
2	3
5	17
8	23
5	4
6	23
0	14
5	22
11	9
2	22
10	20
2	5

t-Test: Two-Sample Assuming Equal Variances

	Control	*Drug*
Mean	5	14.8
Variance	13.28 571 429	64.31 428 571
Observations	15	15
Pooled Variance	38.8	
Hypothesized Mean Difference	0	
df	28	
t Stat	–4.308 646 306	
P(*T*<=*t*) one-tail	9.13838E-05	
t Critical one-tail	1.701 130 934	
P(*T*<=*t*) two-tail	0.000 182 768	
t Critical two-tail	2.048 407 142	

Figure 9.4 A worked example of using Microsoft Excel to calculate an independent-samples *t* test value.

SPSS

General instructions for inputting data into SPSS can be found in Appendix C.

Data Entry

In SPSS, each row of the data file represents a participant. Since both samples in an independent-samples *t* test have different participants, all of the dependent variable data from both samples will need to be placed in one column. Within **Variable View**, label this variable appropriately. However, also create a second variable that will allow us to identify which data goes with which group. A typical label for this variable might be "condition." Then, go to **Data View**. Input the sample data to the appropriate column, and use a nominal variable in the "condition" column to distinguish the two samples (either "0" and "1" or "1" and "2" are typical). See Figure 9.5 for a worked example.

	heart_rate	condition
1	84	1
2	78	1
3	67	1
4	87	1
5	80	1
6	78	1
7	78	1
8	79	1
9	82	1
10	81	1
11	88	2
12	97	2
13	74	2
14	80	2
15	87	2
16	90	2
17	90	2
18	86	2
19	84	2
20	78	2

Figure 9.5 An example of entered data for an independent-samples *t* test in SPSS.

Data Analysis

1) Click **Analyze** on the tool bar, select **Compare Means**, and then click **Independent-Samples T Test**.
2) Highlight the dependent variable column label in the left box, and click the arrow to move it into the **Test Variable(s)** box. Move the "condition" variable to the **Grouping Variable** box.

3) Because there may be more than two conditions identified under the grouping variable, click **Define Groups** to identify which two groups we want to compare. Place the nominal values used to distinguish the groups into the two group boxes, one in each. Click **Continue**.
4) Click **OK**.
5) The output will generate two boxes. The first box will identify how many scores were in the sample (*N*) as well as the mean, standard deviation, and an estimate of the standard error from the perspective of each variable (neither of these is used in the *t* test). The second box will identify, among other things we are not currently interested in, the *t* value, the degrees of freedom, the significance level, mean difference, and the estimate of the standard error (labeled as "Std. Error Difference"). The output will not generate the t_{crit}. We can find t_{crit} ourselves, or we can look at the given significance level to see if that value is equal to or lower than .05. If it is, we can reject the null. If it is not, we need to fail to reject the null hypothesis. (Note: SPSS does not compare the t_{obt} value with a directional or one-tailed critical score.) See Figure 9.6 for a worked example.

T-test

Group statistics

	Condition	*N*	Mean	Std. deviation	Std. error mean
heart_rate	1	10	79.40	5.254	1.661
	2	10	85.40	6.687	2.115

Independent samples test

		Levene's test for equality of variances		t-test for equality of means			
		F	Sig.	*t*	*df*	Sig. (2-tailed)	Mean difference
heart_rate	Equal variances assumed	0.975	0.337	–2.231	18	0.039	–6.000
	Equal variances not assumed			–2.231	17.046	0.039	–6.000

Independent samples test

		t-test for equality of means		
			95% confidence interval of the difference	
		Std. error difference	Lower	Upper
heart_rate	Equal variances assumed	2.689	–11.650	–0.350
	Equal variances not assumed	2.689	–11.672	–0.328

Figure 9.6 A worked example using SPSS to calculate an independent-samples *t* test.

Key Formulas

Standard error of the difference, $\sigma_{M_1-M_2}$

$$\sigma_{M_1-M_2} = \sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \quad \text{(Formula 9.1)}$$

Definitional formula for the estimated standard error, s_{M_1-M2}

$$s_{M_1-M_2} = \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \quad \text{(Formula 9.2)}$$

Pooled variance

$$s_p^2 = \frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2} \quad \text{(Formula 9.3)}$$

Variance formula for the estimated standard error, s_{M_1-M2}

$$s_{M_1-M_2} = \sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \quad \text{(Formula 9.4)}$$

Computational formula for s_{M_1-M2}

$$s_{M_1-M2} = \sqrt{\frac{\left(\Sigma X_1^2-(\Sigma X_1)^2/n_1\right)+\left(\Sigma X_2^2-(\Sigma X_2)^2/n_2\right)}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \quad \text{(Formula 9.5)}$$

***t* statistic for independent samples**

$$t_{obt} = \frac{(M_1-M_2)-(\mu_1-\mu_2)}{s_{M_1-M_2}} \quad \text{(Formula 9.6)}$$

The *t* ratio

$$t_{obt} = \frac{M_1-M_2}{s_{M_1-M_2}} \quad \text{(Formula 9.7)}$$

Cohen's *d* for independent-samples *t* test

$$d = \frac{\textit{estimated mean difference}}{\textit{estimated standard deviation}} = \frac{M_1-M_2}{\sqrt{s_p^2}} \quad \text{(Formula 9.8)}$$

Confidence interval for a population mean difference for independent samples

$$LL = (M_1-M_2)-t_{.05}s_{M_1-M_2}$$
$$UL = (M_1-M_2)+t_{.05}s_{M_1-M_2} \quad \text{(Formula 9.9)}$$

Key Terms

Independent-samples *t* test
Between-participants designs
Homogeneity of variances
Standard error of the difference
Pooled variance
Two-tailed (or nondirectional) test
One-tailed (or directional) test

Questions and Exercises

For the following problems that involve a *t* test, assume a two-tailed test is called for unless otherwise specified.

1 Which of the following describes a situation in which an independent-samples *t* test would be the proper inferential test to use?
- **a** Two separate groups of participants are sampled from two separate populations that will then be compared to see if there is a difference.
- **b** One sample of participants will be measured under two different conditions: one a control condition and one an experimental condition.
- **c** One group of participants is used to obtain one sample and compared with a known population mean.
- **d** None of the above.

2 Independent-samples *t* tests use _____ means to draw inferences about _____ means.

3 Which of the following is an accurate way to represent the null hypothesis for an independent-samples *t* test?
- **a** $M_1 - M_2 = 0$
- **b** $M_1 - M_2 \neq 0$
- **c** $\mu_1 - \mu_2 = 0$
- **d** $\mu_1 - \mu_2 \neq 0$
- **e** $\mu_1 = M_1$ and $\mu_2 = M_2$
- **f** $\mu_1 = M_2$ and $\mu_2 = M_1$

4 Which of the following is an accurate way to represent the alternative hypothesis for an independent-samples *t* test?
- **a** $M_1 - M_2 = 0$
- **b** $M_1 - M_2 \neq 0$
- **c** $\mu_1 - \mu_2 = 0$
- **d** $\mu_1 - \mu_2 \neq 0$
- **e** $\mu_1 \neq M_1$ and $\mu_2 \neq M_2$
- **f** $\mu_1 \neq M_2$ and $\mu_2 \neq M_1$

5 Suppose one sample has an $n = 12$ with an $s^2 = 14$ and another sample has an $n = 10$ with an s^2 of 8. If we wanted to find a t_{crit} score, what df would we use?
a 22
b 44
c 9
d 20

6 If the null hypothesis for an independent-samples t test is ultimately true, then the observed difference between the sample means is due to _____.

7 Must researchers only use the subscripts 1 and 2 when writing out null and alternative hypotheses? Why or why not?

8 Which of the following statistical expressions reflects the proper formal presentation of an independent-samples t test analysis?
a $t(24) = 2.21, d = .29, p > .05$
b $t(24) = 2.21, p < .05, d = .29$
c $t(24) = 2.21, d = .29, p < .05$
d $t(24) = 2.21, p > .05, d = .29$

9 Imagine a study that is looking to see if voice recognition software in cell phones decreases the likelihood of car accidents among drivers. What sort of concerns might come up regarding the assumptions of an independent-samples t test?

10 What is the difference between a pooled variance and a pooled standard deviation? How would one get one from the other?

11 State the critical values and the proper decision regarding the null hypothesis for each of the following independent-samples t test results (all tests are two-tailed).
a $n_1 = 8, n_2 = 7, \alpha = .05, t_{obt} = 1.90$
b $n_1 = 30, n_2 = 30, \alpha = .01, t_{obt} = 7.82$
c $n_1 = 9, n_2 = 9, \alpha = .10, t_{obt} = -4.55$
d $n_1 = 14, n_2 = 12, \alpha = .05, t_{obt} = -.63$

12 A hypothetical study is conducted to evaluate the hypothesis that two-year-old children with no siblings will show more fear around unfamiliar children than two-year-old children with one or more siblings. Each child

is put in a playroom with an unfamiliar child, and fear ratings are obtained through behavioral observations made by an evaluator who is blind to the hypothesis. Fear ratings can range from 1 (no fear) to 10 (a great deal of fear). For the following data:

a Specify the null and alternative hypotheses.
b Perform the appropriate inferential test.
c Identify the critical values for $\alpha = .05$, two-tailed test.
d Reject the null hypothesis?
e If so, what is the size of the effect?
f What type of decision error might have been made?
g Properly present the findings.

No siblings	Siblings
10	7
6	3
8	2
4	4
9	1
7	2

13 A social psychologist is interested in evaluating the hypothesis that anxiety increases biological males' attraction to biological females. The scientific hypothesis is that anxiety will increase interpersonal attraction (see Dutton & Aron, 1974). All participants are told that they are to hear several random bursts of noise in a learning experiment. Half of the participants are led to expect a loud noise (high anxiety condition), and the other half are led to expect a soft noise (low anxiety condition). While waiting for the experiment to begin, participants are placed in a room with a biological female confederate. Later, participants are asked to rate the attractiveness of this female, from 1 (unattractive) to 5 (very attractive). Hypothetical summary statistics are presented in the following tabular list.

a Specify the null and alternative hypotheses.
b Perform the appropriate inferential test.
c Identify the critical values for $\alpha = .05$, two-tailed test.
d Reject the null hypothesis?
e If so, what is the estimated effect size?
f What type of decision error might have been made?
g Properly present the findings.

High anxiety	Low anxiety
$M_1 = 4.2$	$M_2 = 2.2$
$s_1^2 = 0.5$	$s_2^2 = 0.7$
$n_1 = 10$	$n_2 = 10$

14 For Problem 13, suppose that the variances of attraction ratings are increased so that $s_1^2 = 5.2$ and $s_2^2 = 5.4$.
- **a** Perform an independent-samples t test with $\alpha = .05$.
- **b** Interpret the findings.
- **c** What effect has increasing the variability had on t_{obt} and the conclusion about the null hypothesis?

15 Using the variance in Problem 14, suppose the number of participants in each group is increased to 30.

High anxiety	Low anxiety
$M_1 = 4.2$	$M_2 = 2.2$
$s_1^2 = 5.2$	$s_2^2 = 5.4$
$n_1 = 30$	$n_2 = 30$

- **a** Perform an independent-samples t test with $\alpha = .05$.
- **b** Interpret the findings.
- **c** What effect has the increase in sample sizes had on t_{obt} and our decision to reject or fail to reject the null hypothesis?

16 For an independent-samples t test, to what does the standard error of the difference refer?

17 How does the sample size influence t_{crit} and why?

18 In what way does sample size affect the probability of rejecting the null hypothesis?

19 Biaggio (1989) administered a Personal Incidents Record to biological male and female college students to assess the frequency of anger reactions. The author found statistical evidence that biological males reported more anger reactions in comparison with biological females. The following hypothetical data presented are consistent with Biaggio's findings.

a Specify the null and alternative hypotheses.
b Perform the appropriate inferential test.
c Identify the critical values for $\alpha = .05$, two-tailed test.
d Reject the null hypothesis?
e If so, what is the estimated effect size?
f What type of decision error might have been made?
g Properly present the findings.

Frequency of anger reactions	
Males	**Females**
16	9
18	10
15	8
20	4
9	14

20 Burke and Greenglass (1989) have concluded that, "It may be lonely at the top but it's less stressful." These authors found evidence of a statistical difference between teachers and principals on a measure of burnout, with teachers exhibiting higher levels of stress than principals. The following hypothetical data are consistent with their findings.
a Specify the null and alternative hypotheses.
b Perform the appropriate inferential test.
c Identify the critical values for $\alpha = .05$, two-tailed test.
d Reject the null hypothesis?
e If so, what is the estimated effect size?
f What type of decision error might have been made?
g Properly present the findings.

Burnout scores	
Teachers	**Principals**
42	28
38	35
44	40
33	38
49	30
42	24

21 Zakahi and Duran (1988) hypothesized that the very lonely are less physically attractive than those who are not lonely. A loneliness questionnaire was administered. Participants who scored in the top 25% were considered very lonely, while those participants who scored in the bottom 25% were deemed not lonely. Three judges rated all participants' photographs for attractiveness (1 [very unattractive] to 10 [very attractive]). Biological males rated biological females and vice versa. There was no evidence of a statistical difference in attractiveness ratings between very lonely and not lonely biological females. However, there was statistical evidence suggesting lonely biological males were rated as less physically attractive than biological males who were not lonely. The following hypothetical data are in line with the findings of the authors.

a Specify the null and alternative hypotheses.
b Perform the appropriate inferential test.
c Identify the critical values for $\alpha = .05$, two-tailed test.
d Reject the null hypothesis?
e If so, what is the estimated effect size?
f What type of decision error might have been made?
g Properly present the findings.

Attractiveness rating for biological males	
Lonely	**Not lonely**
3	8
6	9
5	7
4	5
7	8

22 In 1972, Buffalo Creek, West Virginia, was the scene of a major flood. The flood was a consequence of corporate negligence. Coal waste that was dumped in a mountain stream created an artificial dam. After several days of rain, the dam gave way, and a black wall of water, over 30 ft high, descended on mining hamlets in the valley. In less than 1 hour, 125 people were dead and 5000 others lost their homes. Simpson-Housley and DeMan (1989) found that, 17 years later, the residents of Buffalo Creek scored higher on a measure of trait anxiety in comparison with the residents of Kopperston, a nearby mining town that did not experience the flood. The following data are hypothetical but are consistent with the findings of the researchers.

a Specify the null and alternative hypotheses.
b Perform the appropriate inferential test.
c Identify the critical values for $\alpha = .05$, two-tailed test.
d Reject the null hypothesis?
e If so, what is the estimated effect size?
f What type of decision error might have been made?
g Properly present the findings.
h Are there any assumptions that should give the researcher cause for concern?

Anxiety scores	
Buffalo Creek	**Kopperston**
50	35
45	37
48	36
40	39
42	40
38	38

23 Narcissism is characterized by self-centeredness, feelings of being "special," and possessing a sense of entitlement. When comparing college students who were firstborns with students who were born later, Jourbert (1989) found that firstborns were more narcissistic than those participants who were born later in the family. The effect was the same for biological males and biological females. Hypothetical summary data are presented for biological males and females. Conduct an independent-samples t test between firstborns and later-borns (a) for males and (b) for females. Set alpha at .05.

a Biological males

Firstborns:	M_1:	23	s_1:	7.84	n_1:	10
Later borns:	M_2:	16	s_2:	6.43	n_2:	15

b Biological females

Firstborns:	M_1:	17	s_1:	6.52	n_1:	19
Later borns:	M_2:	12	s_2:	6.57	n_2:	28

24 For the same data as in Problem 23, find the 95% confidence intervals for the actual narcissism difference between firstborn and later-born males and then firstborn and later-born females.

25 Seery, Holman, and Silver (2010) seem to have found statistical support for the old adage, "that which doesn't kill us makes us stronger." In their study, they found that individuals possessing a history marked by some adversity reported better mental health and higher measures of subjective well-being compared with participants with little or no history of adversity. Following are some data consistent with their findings. Fifteen individuals with 2 or fewer negative life experiences in the past 5 years possessed a mean subjective well-being score of 41 with a s of 4, while 14 individuals more than 5 negative life events in the past 5 years possessed a mean well-being score of 47.2 with a s of 5.

a Specify the null and alternative hypotheses.
b Perform the appropriate inferential test.
c Identify the critical values for $\alpha = .05$, two-tailed test.
d Reject the null hypothesis?
e If so, what is the estimated effect size?
f What type of decision error might have been made?
g Properly present the findings.

26 For the same data as in Problem 25, find the 95% confidence intervals for the actual difference in subjective life expectancy between those who have had 2 or fewer negative life experiences within the past 5 years compared with those who have had 5 or more within the past 5 years.

27 A child psychologist has reason to believe that children who do not spend much time with peers during recess have a problem starting conversations. Sixty children are randomly assigned to 2 treatment conditions (30 children per condition). In the Experimental condition, children learn how to begin a conversation. Children in the Control condition are given talks about the importance of having friends. For a 30-day period after treatment, the children are observed during recess. Each child's average amount of time spent with peers is recorded. The mean playtime for the Experimental condition is 17.0 minutes; the mean for the Control condition is 13.5 minutes. The standard error of the difference is 2.0.

a State the null and alternative hypotheses.
b Conduct an independent-samples t test.
c What is t_{crit}?
d Reject the null hypothesis?
e Properly present the findings.

28 Using the summary data from Problem 27, conduct a one-tailed test that will detect a mean difference only if the Experimental condition produces a higher level of peer interaction.
a State the null and alternative hypotheses.
b Conduct the t test.
c Interpret the findings.
d Is the researcher justified in performing a directional test? Why or why not?

29 Think of a study where a defense could be made for it to be analyzed with a one-tailed t test. Explain why this test would be appropriate.

30 Where students study may be as important as how much they study. Students who have one setting in which they regularly study may perform differently than students who have no regular study location. To test this, a random sample of 15 Introductory Psychology students is asked to study their class material for one hour every day in a special quiet room in the university library. A second sample of 15 students from the same class is also asked to study their class material one hour every day but rotating among various settings (dorm room, cafeteria, and library). At the end of the semester, point totals for the psychology class are obtained and compared, with the following summary statistics:

One setting	Various settings
$M_1 = 80$	$M_2 = 69$
$s_1^2 = 106.09$	$s_2^2 = 156.25$
$n_1 = 15$	$n_2 = 15$

a Specify the null and alternative hypotheses.
b Perform the appropriate inferential test.
c Identify the critical values for $\alpha = .05$, two-tailed test.
d Reject the null hypothesis?
e If so, what is the estimated effect size?
f What type of decision error might have been made?
g Properly present the findings.
h Should the researcher be concerned about a potential confound in this study?

31 Which of the methodological assumptions of the independent-samples t test is effectively dealt with by randomly sampling from the populations in question?

32 Which of the statistical assumptions of the independent-samples t test is most benefited by having $n_1 = n_2$?

Computer Work

33 We observe that people seem to be happier when they are wearing a new article of clothing. To test this, we provide a small random sample of students with new t-shirts and instruct them to wear the shirts all day. At the end of the day, we ask these participants to rate, on a 10-point Likert scale, their happiness (or what social psychologists call "subjective well-being"). A control group of students is also asked for this self-rating at the end of the day, but without the experimental manipulation. Ratings for each participant are reported below. Higher scores indicate greater happiness. Perform an independent-samples t test on the following data set ($\alpha = .05$) and interpret the findings. Are there any assumptions that should give the researcher cause for concern?

Happiness	
New t-shirt	**Control**
8	4
7	6
9	5
6	4
8	6
4	4
5	6
3	8
8	3
5	5
7	3
9	7

34 Social psychologists hypothesize that when a person believes they have unintentionally harmed someone, the person will be motivated to compensate the victim. However, if compensation is not possible, the "harm doer" will be more likely to act generously to some other person (e.g. Carlsmith & Gross, 1969). Fifty participants are randomly assigned to two treatment conditions (25 participants per condition). Participants are told that they will be in a problem-solving study. When a participant arrives, they are asked to wait in the hallway while the apparatus is being set up. Soon

afterward, two confederates, one carrying a camera, approach the participant and ask if they could have their picture taken together. The camera is handed to the participant, and the experimental manipulation begins. In the Compensation condition, the camera is rigged so that it breaks when the participant adjusts the focus. The confederates appear mildly distraught over the mishap but, if offered, refuse any compensation. In the Control condition, the camera does not break.

During the problem-solving phase of the study, participants are asked to write down their answers to 75 arithmetic questions. A stack of answer sheets that has been ostensibly completed by other participants is next to the participant. At the end of the experiment, the researcher asks if the participant would mind staying and scoring a few of the answer sheets.

In order for the data to be in a form appropriate for a *t* test, let us assume that all participants agree to help. The dependent variable is how many answer sheets the participant scores before leaving. (If the dependent variable were the percentage of people in each condition who agreed to help, a different statistical analysis would have to be used [see Chapter 17].) The research hypothesis is that participants in the Compensation condition will score more answer sheets than participants in the Control condition. Perform an independent-samples *t* test on the following data set ($\alpha = .05$) and interpret the findings. Would our conclusion be different if we had performed a one-tailed test at the same alpha level?

Number of answer sheets scored

Compensation condition							
40	34	48	22	35	16	67	84
33	22	50	54	60	68	59	22
30	29	32	33	19	55	54	49
40							
Control condition							
22	30	40	22	35	16	40	79
13	22	25	40	54	61	59	22
28	24	20	18	19	42	34	40
29							

35 A social psychologist hypothesizes that snake-phobic individuals would be more likely to approach a snake if they believe that they are not experiencing anxiety (see Valins & Ray, 1967). Sixty college students who reported

on a "questionnaire of fears" that they were very frightened of snakes served as participants. Upon arriving at the laboratory, all participants were asked to walk over to a large, nonpoisonous snake and pick it up. Heart rate sensors were attached to them, and a speaker on the recording device sounded heartbeats that the participants believed were accurate recordings of their own heart rate. In the High-Arousal condition, participants heard a heart rate that was 120 beats per minute. In the Low-Arousal condition, participants heard a heart rate of 75 beats per minute. In truth, the heart rates were prerecorded and did not reflect the true heart rates of the participants. The scientific hypothesis was that participants in the Low-Arousal condition, believing that they were not experiencing anxiety, would walk closer to the snake than those who believed that their hearts were beating fast. Participants were told that they could stop approaching the snake whenever they felt too uncomfortable to continue. Markings on the floor allowed the investigator to determine how close participants were to the snake when they stopped their approach. Hypothetical data are presented. Lower numbers reflect *greater* approach behavior. Perform an independent-samples *t* test and interpret the findings. Set alpha at .05.

Approach behavior (in ft)

Low arousal					
6.0	5.8	3.2	4.5	6.8	8.2
7.4	6.9	4.3	5.5	6.2	7.0
5.2	6.1	5.9	4.4	3.2	1.3
8.7	5.2	4.8	3.2	1.6	4.8
6.6	7.6	8.0	8.5	4.5	7.9
High arousal					
2.0	3.8	2.2	4.5	5.8	6.1
1.2	2.2	2.3	4.5	3.2	4.0
2.2	1.1	5.9	3.4	3.2	1.3
6.7	4.2	8.8	3.2	1.6	4.8
6.6	5.6	4.0	8.5	4.5	7.9

36 One of the more exciting areas of recent research concerns the disparity between our perceptions of personal honesty and our practices of deceit and cheating. Gino and Ariely (2012) examined the relationship between creativity and honesty by having participants answer general knowledge questions and then transfer their answers to a proper coding form. These forms, the participants were told, mistakenly had faint

impressions of the right answers on them from a previous project. By secretly gaining access to both the participants' original answers and the answers they transferred to the coding forms, the researchers were able to gain a measure of cheating. The participants had been premeasured on a task of creativity and placed into a high creativity group and a low creativity group. The following data are representative of what the researchers found. Higher numbers represent more occurrences of cheating. Perform an independent-samples *t* test and interpret the findings. Set alpha at .05.

High creativity					
8	6	5	6	9	14
9	9	0	14	7	6
15	2	7	0	0	1
11	14	1	2	3	13
8	8	9	13	15	8
Low creativity					
3	1	3	0	0	2
8	11	0	8	0	0
16	0	1	7	7	7
12	15	3	2	3	2
2	1	7	0	6	18

10

Testing the Difference Between Two Means: The Dependent-Samples *t* Test

10.1 The Research Context

The independent-samples *t* test is used to contrast means computed from two unrelated samples of scores. In *dependent sampling*, each score in one sample is related to another score in a second sample. Pairs of scores are formed by either *the way conditions are presented* or *the manner in which participants are assigned to conditions*. First, a discussion of the case in which pairs of scores are formed by the way conditions are presented will be discussed. Then the situation in which pairs of scores are formed by the manner in which participants are assigned to conditions will be considered. The appropriate *t* test for dependent sampling is called a **dependent-samples *t* test** (also known as a *paired-samples t test*, *related-samples t test*, or *correlated-samples t test*).

Repeated-Measures Designs

Most designs utilizing a dependent-samples *t* test for analysis are experimental in nature. For that reason, the language of experimentation will be used in the following sections. However, dependent sampling designs can be nonexperimental. A **repeated-measures design** (or *within-participants design*) assesses one group of participants under two or more treatment conditions; however, in this chapter, only the case in which *two* treatment conditions are evaluated is addressed.[1] The following research example is presented as a between-participants design and then as a repeated-measures design.

Suppose a cognitive psychologist is interested in assessing the relative merits of two kinds of memory strategies that can be used to remember the names of people. The first strategy involves imagery. A sample of participants is taught to associate unfamiliar names of people with a related mental image of the person's

1 Chapter 14 addresses the analysis of repeated-measures designs in which more than two treatment conditions are used.

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

face. For instance, when they are presented with a snapshot of the Hall-of-Fame baseball player Ty Cobb, the participants could picture him eating corn on the cob. The next time they see an image of the famous outfielder, these participants will presumably think of corn on the cob and remember "Cobb." The second method uses repetition and employs a *separate* group of participants. When viewing snapshots of people, these participants repeat the person's name three times before moving on. Because two separate samples of participants are used, this is a between-participants design; each participant is only in one of the two groups. In this case, an independent-samples t test would be used to compare the means of the two groups.

This same study might also be conducted using a repeated-measures design. Unlike the between-participants design in which participants are assigned to *one* of the two experimental conditions (imagery vs. repetition), in the repeated-measures design, *all* participants receive *both* experimental treatments. Each participant supplies a pair of scores, one score for each condition. In this way, the imagery method would be taught and assessed, and then the repetition method would be taught and assessed. The means of the two experimental groups are compared with a dependent-samples t test. The repeated-measures design is an efficient design because the researcher can use fewer participants in comparison with the between-participants design. More importantly, the repeated-measures design also increases the *power* when testing the null hypothesis.[2] The power of a statistical test is the probability of correctly rejecting a false null hypothesis, the probability of avoiding a Type II error. That is, if the null is actually false, how likely are we to find evidence to show it is false? As likelihood increases, power increases. The power of a dependent-samples t test is greater than the power of the independent-samples t test. This is due to the ability of the dependent-samples test to remove some of the variability associated with using different participants. (This concept will be explored in greater detail in Chapter 14.) As a result, the denominator in the dependent-samples test will be smaller than the denominator in the independent-samples test. Smaller denominators yield larger t values, and larger t values are more likely to fall into the rejection region of a distribution.

Two research examples from the psychological literature are presented in the following sections. A repeated-measures design is used in each instance, and a dependent-samples t test can be used to analyze the data from these studies.

► **Example 10.1** Lehman et al. (2013) tested a variety of hypothesis regarding the features of music listened to and walking gait of the participants. Eighteen individuals walked around while listening to a variety of different playlists. One hypothesis concerned the influence of strong tempos. The research team found

2 See Chapter 11 for an extensive discussion of the concept of power.

that stronger tempo songs influenced participants' walking gait more so than softer tempo songs. ◄

► **Example 10.2** Addison (1989) tested the hypothesis that people perceive bearded men differently from nonbearded men. Addison used a repeated-measures design in which one group of participants rated pictures of bearded (Condition 1) and nonbearded men (Condition 2) on a number of dimensions. The results indicated that, compared with nonbearded men, males with beards are rated as more masculine, more aggressive, stronger, and more dominant. However, they are not viewed as more intelligent. ◄

Another example of repeated-measures design occurs when participants supply scores both before and after a treatment. This type of repeated-measures design is called a pretest/posttest repeated-measures design.[3] It is useful when evaluating behavior change that is caused by the introduction of an independent variable. Studies that examine learning, performance, or therapy effects frequently employ pretest/posttest designs.

Before leaving repeated-measures designs, it is important to note the methodological challenges that are introduced when participants are measured more than once. The primary issue concerns whether the second measurement of a participant is influenced by the first measurement. If so, then the difference between the participant's two scores may be influenced by practice effects or fatigue effects. This explanation competes with the independent variable and creates a confound. There are potential methodological solutions for some of these situations, but they are not always applicable. In the memory task described above, it would be important to have two separate lists of names that have been judged ahead of time to be of equal difficulty to memorize. Furthermore, it might be necessary to have half of the participants associate a particular list of names with a particular mnemonic technique while having the other half switch these associations; this is called **counterbalancing**. Furthermore, it might be necessary to counterbalance the order of exposure – having half of the participants experience the imagery condition first, while the other half experience the repetition condition first. The statistical advantages of repeated-measures designs are attractive, but these are often more than offset by the corresponding methodological challenges that accompany them.

A full exploration of the reasons researchers decide to use one design or the other are beyond the scope of this text. Please consult a research methodology resource for more information.

3 More accurately, this is an example of a quasi-experimental design. There are some inherent weaknesses with this type of design regarding the adequate control of all extraneous variables.

Matched-Samples Design

Another type of design in which a dependent-samples *t* test is used involves two independent but related samples of participants; the **matched-samples** (or *matched-participants*) **design**. A matched-samples design derives its name from the way participants are assigned to conditions. In this design, participants are assigned based on prior information about the participants.

Imagine comparing two educational programs designed for a grade school. One program uses behavior modification, emphasizing individually tailored performance goals, positive reinforcement for increasing mastery of material, and the provision of teaching machines. The other program employs the Montessori method, a program in which students are free to learn in an unstructured environment. In this pedagogical system, the teacher makes learning opportunities available and assumes that students will learn different topics as their interests guide them to new content areas. When comparing the impact of the two programs, the experimenter wants to ensure that the two groups are the same for any important extraneous variable, like intelligence. (If the children in one program had higher IQs than children in the other program, then IQ could explain group differences instead of the educational program.) Random assignment is an adequate technique to control participant variables (see Chapter 1). However, it is unlikely that parents will allow their children's grade school educational experience to be determined by the random assignment of a researcher.[4] Another way to ensure that groups do not differ on an important extraneous variable is to employ a matched-samples design. To do this, we would obtain IQ scores for each student *before* beginning the programs. Next, we would pair up students with identical IQ scores, one from each program. Only students who could be paired with another from the other educational system would be included in the study. This matching procedure controls IQ between the treatment conditions.

Yet another way to match participants, but which would not work in the above example, is to premeasure participants on some important variable, create identical pairs based on this premeasure, and then randomly split the pairs placing one in each condition. (See methodology texts for more detailed information about forming matched samples and the methodological challenges that accompany this procedure.)

This chapter will use the repeated-measures design to illustrate the dependent-samples *t* test. However, the formulas and basic concepts of the test are the same for any dependent sampling design that compares two means, including matched samples.

---

4 The inability to assign randomly students to the type of educational experience makes this a quasi-experimental study as well.

10.2 The Sampling Distribution for the Dependent-Samples *t* Test

The sampling distribution for the dependent-samples t test is based on differences between means of dependent samples. Consider a repeated-measures design in which one group of participants is exposed to two treatment conditions. Since there are two measurements taken from each participant, there are two samples of scores. Each sample represents a hypothetical population, the population of potential participants operating under each condition.

Using Dependent Sampling to Form a Theoretical Sampling Distribution

The sampling distribution formed by dependent sampling is theoretically constructed in the following manner.

Step 1. Take one group of participants, and administer the two treatment conditions, measuring the effects of each treatment and with each participant supplying one pair of scores.

Step 2. For each participant, subtract the second score from the first. This gives us one distribution of *difference scores*.

Step 3. Compute the mean of this distribution of difference scores, symbolized as $\bar{D}$.

Step 4. Using the same sample size, repeat steps 1–3 an infinite number of times.

Step 5. Plot the relative frequencies of $\bar{D}$ to obtain the sampling distribution of differences.

The sampling distribution is a distribution of mean differences. The mean of the sampling distribution, symbolized as $\mu_{\bar{D}}$, will be the same as the difference between the means of the two hypothetical populations. If there is a difference of five units between the two populations, then the mean of the sampling distribution will be 5. If there is no difference between the populations, then the mean of the sampling distribution will be 0. The null hypothesis is that the mean of the sampling distribution is 0. This is the same as saying that there is no treatment effect. Under the null hypothesis, $\mu_X - \mu_Y = \mu_D = 0$. (The X and Y subscripts refer to the first and second set of scores, respectively.) Recall that the mean of a sampling distribution is the same as the mean of the population from which it is taken. Therefore, $\mu_{M_X} = \mu_X$, $\mu_{M_Y} = \mu_Y$, and $\mu_{\bar{D}} = \mu_D$. Figure 10.1a shows two populations with different means. Figure 10.1b is the sampling distribution of differences derived from the populations. (Ignore Figure 10.1c for now.)

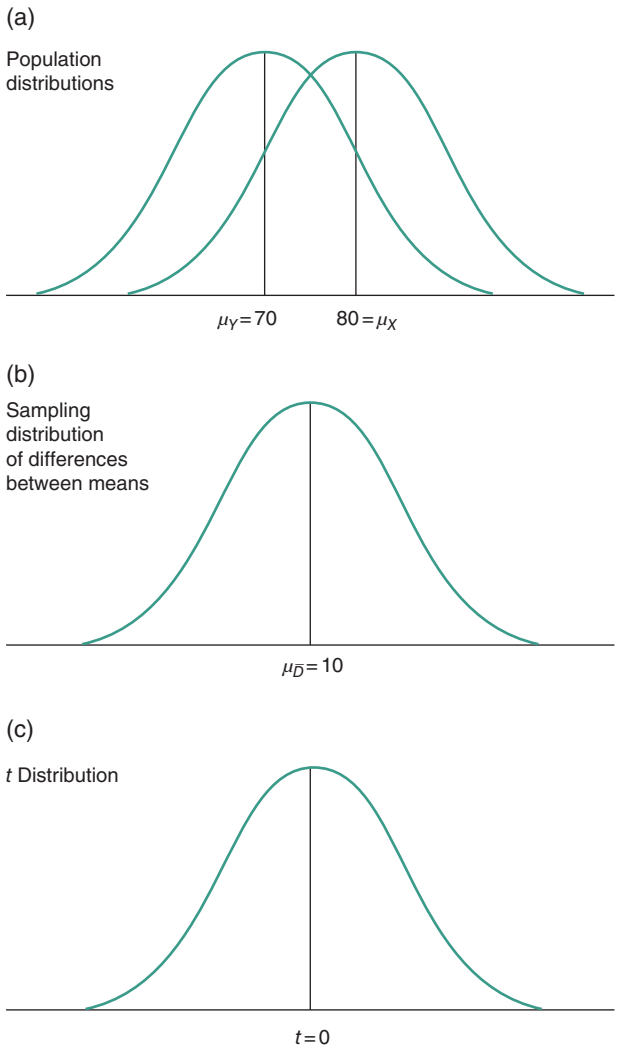


Figure 10.1 In (b), the mean of the sampling distribution of differences, $\mu_{\bar{D}}$, equals the difference between population means, $\mu_X - \mu_Y$, found in (a). In (c), the *t* statistic transforms the sampling distribution to a *t* distribution with a mean of 0.

The Standard Error of the Sampling Distribution for Dependent Samples

The standard deviation of the sampling distribution is called the standard error of the difference, or standard error. The estimate of the population standard error of the difference is symbolized $s_{\bar{D}}$. The formula uses the standard deviation

of the difference scores, s_D, divided by the square root of the number of pairs of scores. Formula 10.1 computes the estimated standard error once s_D is known.

The estimate of the standard error of the difference, $s_{\bar{D}}$

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n_p}} \quad \text{(Formula 10.1)}$$

where

s_D = the standard deviation of the difference scores
n_p = the number of *pairs* of scores

The computational formula for s_D is given in Formula 10.2.

Computational formula for the standard deviation of the difference scores, s_D

$$s_D = \sqrt{\frac{\Sigma D^2 - \left[(\Sigma D)^2 / n_p\right]}{n_p - 1}} \quad \text{(Formula 10.2)}$$

where

$D = X - Y$, a participant's score in treatment 1 minus the same participant's score under treatment 2
n_p = the number of *pairs* of scores

A word of caution: Remember to calculate the standard error ($s_{\bar{D}}$) by dividing s_D by the *square root* of n_p. The standard deviation of the difference scores s_D is not equal to the estimate of the standard error of the difference $s_{\bar{D}}$.

Before discussing the formula for a dependent-samples t test, let us use raw data to compute $s_{\bar{D}}$. Since $s_{\bar{D}}$ is the estimate of the standard error of the sampling distribution, it will be placed in the denominator of the t statistic.

■ **Question** *What is $s_{\bar{D}}$ for the following data set?*

Solution

Scores			
X	**Y**	**D**	**D^2**
5	3	2	4
6	4	2	4
7	7	0	0
6	7	−1	1
		$\Sigma D = 3$	$\Sigma D^2 = 9$

Using Formula 10.2 to compute s_D,

$$s_D = \sqrt{\frac{\Sigma D^2 - \left[(\Sigma D)^2 / n_p\right]}{n_p - 1}}$$

$$s_D = \sqrt{\frac{9 - \left[(3)^2/4\right]}{4-1}} = \sqrt{\frac{6.75}{3}}$$

$$s_D = \sqrt{2.25} = \mathbf{1.50}$$

The standard deviation of the difference scores is 1.50. Now place 1.50 in the $s_{\bar{D}}$ formula.

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n_p}}$$

$$s_{\bar{D}} = \frac{1.50}{\sqrt{4}}$$

$$s_{\bar{D}} = \mathbf{0.75}$$

The estimate of the standard error of the difference is 0.75. Remember that this estimate is the standard deviation for a distribution of means; for dependent-samples *t* tests, the standard error is the standard deviation for a distribution of mean differences, given a particular sample size (n_p). ■

10.3 The *t* Distribution for Dependent Samples

To see if there is a difference between the means of two dependent samples, the sampling distribution must be transformed into a *t* distribution, and the sample mean difference must be positioned on it. Recall that the mean of the sampling distribution will equal the difference between the means of the populations. The null hypothesis typically assumes this difference to be 0, and the *t* test is set up accordingly. The difference between the obtained sample means is divided by the estimate of the standard error. This step registers how likely it is to get that sample mean difference *if* the null hypothesis is true. In this way, a sample mean difference is transformed and placed onto a *t* distribution. Similar to the independent-samples *t* test, the proper formula for the dependent-samples *t* test also acknowledges the contrast between sample means ($M_X - M_Y$) and the hypothesized difference between the populations ($\mu_X - \mu_Y$) in the numerator (see Formula 10.3). However, since $\mu_X - \mu_Y$ is normally hypothesized to be 0, the formula typically used for this *t* test simplifies the formula by omitting this term (see Formula 10.4).

Dependent-samples *t* statistic

$$t_{obt} = \frac{(M_X - M_Y) - (\mu_X - \mu_Y)}{s_{\bar{D}}} \quad \text{(Formula 10.3)}$$

Dependent-samples *t* test

$$t_{obt} = \frac{(M_X - M_Y)}{s_{\bar{D}}} \quad \text{(Formula 10.4)}$$

The t_{obt} in these formulas is the number of estimated standard error units; the sample mean difference is from the mean of the t distribution. Therefore, if t = 1.45, the mean difference ($M_X - M_Y$) is 1.45 estimated standard error units above the mean of the t distribution.

The degrees of freedom for the dependent-samples t test is $n_p - 1$, the number of *pairs* of scores minus one. There is a different t distribution for every sample size (i.e. every number of pairs of scores). Similar to the independent-samples t test, the number of degrees of freedom is used to find t_{crit} in the t table (Table A.2).

The Null and Alternative Hypotheses for the Dependent-Samples *t* Test

The null hypothesis states that the population means are the same. The alternative hypothesis states that the means are not the same.

$$H_0 : \mu_1 = \mu_2 \text{ or } \mu_D = 0$$
$$H_1 : \mu_1 \neq \mu_2 \text{ or } \mu_D \neq 0$$

The requisite formulas are now in place to conduct a dependent-samples t test.

Worked Example

A cognitive psychologist is interested in the effects of alcohol intoxication on learning and recall. A repeated-measures design is used in which one group of participants is exposed to both an experimental as well as a control condition. In the experimental condition, participants consume an amount of alcohol sufficient to raise their blood alcohol level to 0.10%, the legal criterion for intoxication in many states. While participants are intoxicated, slides of geometric designs and a nonsense word printed below each design (e.g. "geostatic" or "gravoserv") are projected one at a time. After 20 presentations, the slides are presented again, in random order, without the associated nonsense word. The participants are to provide the word associated with each design. The dependent variable is the number of incorrect responses.

The control condition is administered one week later. Different designs and nonsense words are presented, but now the participants are given drinks that taste like alcohol but contain no alcohol (placebo). Since improvement in

Table 10.1 A worked problem using the dependent-samples *t* test.

Alcohol (*X*)	Placebo (*Y*)	*D*	D^2
8	4	4	16
12	8	4	16
6	2	4	16
4	6	–2	4
11	8	3	9
15	9	6	36
8	5	3	9
7	4	3	9
$M_x = 8.88$	$M_y = 5.75$	$\Sigma D = 25$	$\Sigma D^2 = 115$

$$t_{obt} = \frac{M_X - M_Y}{s_{\bar{D}}}$$
$$s_{\bar{D}} = \frac{s_D}{\sqrt{n_p}}$$
$$s_D = \sqrt{\frac{\sum D^2 - \left[(\sum D)^2 / n_p\right]}{n_p - 1}}$$
$$s_D = \sqrt{\frac{115 - \left[(25)^2/8\right]}{8-1}}$$
$$s_D = \sqrt{\frac{115 - 78.3}{7}}$$
$$s_D = \sqrt{5.27}$$
$$s_D = 2.30$$
$$s_{\bar{D}} = \frac{2.30}{\sqrt{8}}$$
$$s_{\bar{D}} = 0.81$$
$$t_{obt} = \frac{8.88 - 5.75}{0.81}$$
$$t_{obt} = \frac{3.13}{0.81}$$
$$t_{obt} = \mathbf{3.86}$$
$$df = n_p - 1 = 7$$
$$t_{crit} = \pm 2.365$$

Since t_{obt} falls outside of the t_{crit} values, reject the null hypothesis that claims $\mu_X = \mu_Y$.

The raw scores are number of errors in recall.

performance between the experimental and control conditions may be because the control condition comes *after* the experimental condition, one-half of the participants are run through the control condition first, followed one week later by the experimental condition. Table 10.1 presents the raw data (number of errors) for each condition, as well as the computation of t_{obt}.

A dependent-samples *t* test is used to compare the effects of alcohol on learning and recall. The procedural steps in the analysis and interpretation of the data are presented in the following list:

Step 1. Define the null and alternative hypotheses:

$$H_0: \mu_X = \mu_Y$$
$$H_1: \mu_X \neq \mu_Y$$

Step 2. Set alpha. Alpha is set at .05.

Step 3. Compute t_{obt} (see computations in Table 10.1).

Step 4. Locate the critical *t* value in Table A.2 in the Appendix. The degrees of freedom is $n_p - 1$, that is, 8 – 1 = 7. Enter the left column of the *t* table, and locate the number 7. Move to the column under alpha of .05 for a *two-tailed* test. The critical value is 2.365 (understood as ±2.365 for a two-tailed test).

Step 5. Compare the t_{obt} of 3.86 with the critical value of ±2.365. Since t_{obt} falls outside of the critical values, the null hypothesis is rejected in favor of the alternative hypothesis.

Step 6. Interpret the findings. Statistical evidence suggest learning and/or recall of names is negatively affected by the ingestion of an intoxicating amount of alcohol, $t(7) = 3.86$, $p < .05$. Additional research is needed to see if other forms of learning are likewise affected.

A Measure of Effect Size: Cohen's *d*

One secondary question that can be asked when a null hypothesis is rejected concerns the size of the treatment effect. As noted in previous chapters, t_{obt} is not designed to measure effect size. A simple, direct, and often used effect size measure is Cohen's *d*. Here is the formula:

Cohen's *d* for dependent-samples *t* test

$$d = \frac{\textit{sample mean difference}}{\textit{standard deviation of difference scores}} = \frac{M_X - M_Y}{s_D} \quad \text{(Formula 10.5)}$$

$M_X - M_Y$ is used as the best estimate of the mean difference between the two populations, and s_D is the best estimate of σ_D, the standard deviation of the population of difference scores. Notice, this only allows us to *estimate* the effect size. As with previous versions of the statistic, Cohen's *d* reflects the difference between the means in standard deviation terms. Larger *d* values reflect larger effect sizes. A word of caution, however, is needed; Cohen's *d* is susceptible to overestimating effect size in this particular test.

10.4 Comparing the Independent- and Dependent-Samples *t* Tests

As stated earlier, the dependent-samples t test is more sensitive to detect an experimental effect than the independent-samples t test. In other words, the probability of rejecting an incorrect null hypothesis is greater when using the dependent-samples t test compared with the independent-samples t test, all other things being equal. A good way to demonstrate this point is to imagine that the alcohol and learning study (in the Worked Example) was conducted using independent samples; that is, two unrelated groups of participants. Table 10.2 presents the summary statistics of the data in Table 10.1 and the computation of t_{obt} for an independent-samples t test. The obtained t value is 2.06, which just misses the critical value of ±2.145. In this case, the difference in designs and type of t test makes the difference between rejecting and failing to reject the null hypothesis. This is largely because the use of a repeated-measures design and

Table 10.2 An independent-samples *t* test on the raw data of Table 10.1.

Alcohol	Placebo
$M_1 = 8.88$	$M_2 = 5.75$
$s_1^2 = 12.67$	$s_2^2 = 5.90$
$n_1 = 8$	$n_2 = 8$

Formula for independent-samples t test

$$t_{obt} = \frac{M_1 - M_2}{\sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$t_{obt} = \frac{8.88 - 5.75}{\sqrt{\frac{12.67(8-1) + 5.90(8-1)}{8+8-2}\left(\frac{1}{8}+\frac{1}{8}\right)}}$$

$$t_{obt} = \frac{3.13}{\sqrt{\frac{88.69 + 41.30}{14}(0.25)}}$$

$$t_{obt} = \frac{3.13}{\sqrt{9.29(0.25)}}$$

$$t_{obt} = \frac{3.13}{1.52}$$

$$t_{obt} = \mathbf{2.06}$$

$df = n_1 + n_2 - 2 = 8 + 8 - 2 = 14$

$\alpha = .05$

$t_{crit} = \pm 2.145$

Since t_{obt} does not fall outside of the t_{crit} values, fail to reject the null hypothesis that claims $\mu_1 = \mu_2$.

the corresponding dependent-samples t test reduces the variability of the data, making the standard error smaller (0.81 compared with 1.52) and generating a larger t_{obt} value (in terms of absolute value). If we can manage the methodological problems associated with a repeated-measures design and use a dependent-samples t test to analyze the data, it will be statistically advantageous.

10.5 The One-Tailed *t* Test Revisited

The difference between a one-tailed and two-tailed t test was discussed in Chapter 9. A one-tailed, or directional, test places the entire rejection region in one tail of the t distribution. Consequently, for a given alpha level, the critical value to which t_{obt} is compared is smaller when using a one-tailed test. Therefore, it is more likely for t values in the predicted direction to fall into the rejection region. However, when using a one-tailed test, we will be unable to detect mean differences opposite of the predicted direction. We have learned that one-tailed tests are very problematic. Never should a directional test be used when testing a theory, since findings that oppose a theoretical prediction may be important to detect. The strongest justification for the use of a one-tailed test is when a course of action will only be taken if the mean of a particular group is larger than the mean of the other group. Box 10.1 presents a study in which a one-tailed dependent-samples t test is used. Were the authors' justified in using a one-tailed test?

10.6 Assumptions of the Dependent-Samples *t* Test

The assumptions for the dependent-samples t test are nearly identical to the assumptions for the independent-samples t test (see Section 9.4). However, two observations coming from the same participant (as in a repeated-measures design) are clearly not independent. For the dependent-samples t test, it is assumed that scores *within* a given treatment condition will be independent of each other. In addition, the normality assumption for the dependent-samples t test refers to the population of *difference* scores. Finally, just as in previous tests, as n_p increases a test becomes robust to violations of this assumption.

The first published use of the t test utilized the dependent-samples version. Box 10.2 provides a closer look at the study which employed it.

10.7 Interval Estimation of the Population Mean Difference

Just as in previous chapters, we can use statistical information from the dependent samples to generate an interval estimate for the mean difference between the populations. Since each potential sample mean difference drawn has a

corresponding t value, we can use the t distribution and our obtained sample mean difference (which is an unbiased estimate of the population mean difference) to generate a probability function for the value of the actual population mean difference. Choosing t_{crit} values corresponding to different probabilities within the t distribution allows us to create intervals with differing degrees of

Box 10.1 Is the Scientific Method Broken? The Questionable Use of One-Tailed *t* Tests

This is another box in the series exploring the various reasons for the current reproducibility crisis in the social, behavioral, and medical sciences. Fellow researchers sometimes wonder if the use of one-tailed tests in the literature occurs because it is the only way to reject the null hypothesis. The following study may be a case in point. Buttery and White (1978) were interested in the relationship between affective states (feelings) and biorhythms. According to biorhythm theory, people experience a 28-day emotional cycle. At the peak of the cycle, people are expected to be cheerful and optimistic. At the bottom of the cycle, people are prone to be irritable and negative.

Twenty participants were asked to provide ratings of 11 emotionally related concepts. Ratings were obtained from participants at both the high and low points of their emotional biorhythm. Since each participant supplied two scores, a dependent-samples t test was conducted. The number of paired scores is 20; therefore, the df is 19. The critical value for a two-tailed test when $df = 19$ is ±2.09. The critical value for a one-tailed test, with the same df, is 1.73. The authors' t_{obt} was 1.76, which they reported as evidence to reject the null hypothesis given the use of a one-tailed test. This raises at least a couple issues. First, how can readers know that the decision to run a one-tailed test was made ahead of time, for this study or any other one that uses a one-tailed test? Second, even if the authors decided on a one-tailed t test before the data were collected, this decision fails to allow for the possibility that the direction of the effect could have been contrary to their prediction. Perhaps people at the bottom of their cycle are *more* cheerful than when they are at the top of their biorhythm cycle. In an area that does not have strong theoretical or empirical reasons for expecting a directional finding, the use of a one-tailed t test is highly questionable. In this situation, the chance of a Type I error may be greater than 5%. The libertarian view taken by many researchers toward the use of one-tailed tests may be another reason for the current reproducibility crisis in the social, behavioral, and medical sciences.

Box 10.2 The First Application of the *t* Test

The first application of a *t* test is found in Gosset's classic 1908 paper, published in the biostatistical journal *Biometrica*. A dependent-samples *t* test was applied to data from a previously published 1904 study by Cushny and Peebles on the effects of sleep medication.

The authors had used a repeated-measures design to contrast the effects of dextro-hyoscyamine hydrobromide and laevohyoscyamine hydrobromide on sleep duration. The participants used were ten residents from the "Michigan Asylum for the Insane at Kalamazoo." The following data are the changes in hours of sleep from a baseline (no drug) period to the period under one or the other drug. A positive score reflects an increase in sleep duration, and a negative score reflects a decrease in sleep duration, relative to baseline.

Since the *t* test had not been invented when Cushny and Peebles conducted their study, the authors merely "eyeballed" the raw data and concluded that one compound was more effective than the other. It turned out that their conclusions were supported by Gosset's subsequent statistical analysis. As we follow the data analysis and the null hypothesis test, keep in mind that the convention of using a 5% level of significance was not yet standard. The data using Gosset's dependent-samples *t* test is analyzed first, and then a test of the null hypothesis is conducted following modern practices.

Patient	*Dextro-X*	*Laevo-Y*	*D*	D^2
1	0.7	1.9	−1.2	1.44
2	−1.6	0.8	−2.4	5.76
3	−0.2	1.1	−1.3	1.69
4	−1.2	0.1	−1.3	1.69
5	−0.1	−0.1	0	0
6	3.4	4.4	−1.0	1
7	3.7	5.5	−1.8	3.24
8	0.8	1.6	−0.8	0.64
9	0	4.6	−4.6	21.16
10	2.0	3.4	−1.4	1.96
	$M_X = 0.75$	$M_Y = 2.33$	$\Sigma D = -15.80$	$\Sigma D^2 = 38.58$

$$t_{obt} = \frac{M_X - M_Y}{s_{\bar{D}}}$$

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n_p}}$$

$$s_D = \sqrt{\frac{\sum D^2 - \left[(\sum D)^2 / n_p\right]}{n_p - 1}}$$

$$s_D = \sqrt{\frac{38.58 - \left[(-15.80)^2 / 10\right]}{10 - 1}}$$

$$s_D = \sqrt{\frac{38.58 - 24.96}{9}}$$

$$s_D = \sqrt{1.51} = 1.23$$

$$s_{\bar{D}} = \frac{1.23}{\sqrt{10}}$$

$$s_{\bar{D}} = 0.39$$

$$t_{obt} = \frac{0.75 - 2.33}{0.39}$$

$$t_{obt} = \mathbf{-4.05}$$

$df = n_p - 1 = 10 - 1 = 9$

$t_{crit} = \pm 2.262$

Since −4.05 falls outside of ±2.262, reject the null hypothesis that claims $\mu_X = \mu_Y$.

certainty. The formula for an interval in which we can have 95% confident follows:

Confidence interval for a population mean difference for dependent samples

$$LL = (M_X - M_Y) - t_{.05}s_{\bar{D}}$$
$$UL = (M_X - M_Y) + t_{.05}s_{\bar{D}} \quad \text{(Formula 10.6)}$$

where

LL = the lower limit of the confidence interval
UL = *the* upper limit of the confidence interval
$t_{.05}$ = the critical value for a t distribution of a given sample size

Since we are generating an interval, two values are calculated, one being the value at the lower end of the interval and the other at the upper end. As the interval widens and becomes less specific, the confidence grows that the actual mean difference falls within that window. A 95% confidence rate is typical, but the above formulas could easily be adjusted to find a 90 or 99% confidence interval simply by finding the corresponding t_{crit} values using the t table (Table A.2).

■ **Question** *Using the same data from the alcohol and learning study explored earlier in the chapter, find the 95% confidence interval for the population mean difference in performance between the two conditions. (M_X = 8.88; M_Y = 5.75; ΣD = 25;ΣD^2 = 115; n_p = 8)*

Solution

Step 1. Identify the null and alternative hypotheses.

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y$$

Step 2. Set the confidence rate at 95% (equivalent to α = .05).

Step 3. Using the t table, find the cutoff values beyond which lie 2.5% in the right tail of the t distribution and 2.5% in the left tail of the distribution. With df = 7 and α = .05 (two-tailed test), the cutoff points are ±2.365.

Step 4. Compute the confidence interval.

$$LL = \left(M_x - M_y\right) - t_{.05}s_{\bar{D}}$$
$$UL = \left(M_x - M_y\right) + t_{.05}s_{\bar{D}}$$

From having worked the problem previously, we know

$$M_X - M_Y = 3.13$$
$$s_{\bar{D}} = 0.81$$

$$LL = 3.13 - 2.365(0.81) = \mathbf{1.21}$$

$$UL = 3.13 + 2.365(0.81) = \mathbf{5.05}$$

Step 5. Interpret the findings. Statistical evidence suggests we can be 95% confident that ingesting an intoxication amount of alcohol negatively affects learning and/or recall of names in this particular experimental task by somewhere between 1.21 and 5.05 names. ■

10.8 How to Present Formally the Conclusions for a Dependent-Samples *t* Test

Proper reporting of inferential statistics can be challenging. Following are examples of how to report, in sentence form, a rejection of the null as well as a fail to reject the null. If rejecting the null, a sentence might read, "A dependent-samples *t* test found statistical evidence suggesting learning and/or recall of names is negatively affected by the ingestion of an intoxicating amount of alcohol, $t(7) = 3.86$, $p < .05$." If we also wanted to include a measure of effect size, the sentence could finish with, "…$t(7) = 3.86$, $p < .05$, $d = 1.36$." If failing to reject the null, a sentence might read, "A dependent-samples *t* test did not find statistical evidence suggesting learning and/or recall of names to be affected by the ingestion of an intoxicating amount of alcohol, $t(7) = -1.62$, *n.s.*" For a more detailed analysis of the style, symbols, and punctuation used in these sentences, please see Section 8.8.

Summary

In *dependent sampling*, each score in one sample is related to another score in a second sample. Pairs of scores are formed either by matching or by the use of a repeated-measures design. Matching is a method used to assign participants to groups so that a particular important variable cannot account for the results of the study. A repeated-measures design exposes each participant to every condition in the study. The sampling distribution for the dependent-samples *t* test is based on differences between means of matched or paired samples. The mean of the sampling distribution of differences equals the difference between means of the populations. The null hypothesis for the dependent-samples *t* test is $\mu_X = \mu_Y$ or $\mu_D = 0$. The alternative hypothesis is $\mu_X \neq \mu_Y$ or $\mu_D \neq 0$.

A dependent-samples *t* test can be performed on pairs of scores. This *t* test is more powerful than an independent-samples *t* test because, all other things being equal, there is less variability associated with paired scores. This means the resulting *t* values tend to be larger and, as a result, more likely to fall into

the rejection region. If the null hypothesis can be rejected, a version of Cohen's *d* can be calculated and used as a measure of effect size.

The assumptions of the dependent-samples *t* test are very similar to the assumptions for an independent-samples *t* test. However, dependently sampled scores are obviously not independent between conditions. For this test, the assumption pertains only to scores *within* a given condition.

As with the independent-samples *t* test, the standard error, sample means, and the *t* distribution can be used to create a confidence interval for the actual value of the difference between the means of the two populations.

Using Microsoft® Excel and SPSS® to Run a Dependent-Samples *t* Test

Excel

General instructions for data entry into Excel can be found in Appendix C.

Data Entry

Enter all of the scores from the samples into two adjacent columns, one sample in each column. Make sure the data from each participant is placed into the same row across the two columns. (See Figure 10.2 for an example.) This is crucial; it is needed to establish a difference score for each participant. Label the columns appropriately.

Alcohol	Placebo
8	4
12	8
6	2
4	6
11	8
15	9
8	5
7	4

t-Test: paired two sample for means

	Alcohol	*Placebo*
Mean	8.875	5.75
Variance	12.69 643	5.928571
Observations	8	8
Pearson correlation	0.769 782	
Hypothesized mean difference	0	
df	7	
t Stat	3.85 104	
P(*T*<=*t*) one-tail	0.003 142	
t Critical one-tail	1.894 579	
P(*T*<=*t*) two-tail	0.006 284	
t Critical two-tail	2.364 624	

Figure 10.2 A worked example of using Microsoft Excel to calculate a dependent-samples *t* test value.

Data Analysis

1) Excel has built-in programs for many inferential tests, including the dependent-samples *t* test. To access it, click on the Data tab on the top menu and then click **Data Analysis**. If this option is not found, the Data Analysis ToolPak needs to be installed. See Excel instruction materials for how to install this feature.
2) With the Data Analysis box open, select **t-Test: Paired Two-Sample for Means**.
3) Input the data range for one variable in box **Variable 1 Range** and the data range for the other variable in box **Variable 2 Range**. (If the labels were included in the range, make sure to click the **Labels** box to exclude those cells.)
4) Decide on an Output option. The default is to place it on a separate worksheet.
5) Click **OK**.
6) The output box will present the means, variances, and observations (sample sizes). Additionally presented will be the Pearson correlation (covered in Chapter 15), hypothesized mean difference (0, unless otherwise specified), degrees of freedom (labeled df), and observed *t* value (labeled "t stat"), as well as the critical scores and probabilities for both one- and two-tailed versions of the test. Compare "t stat" with either the probability value (labeled **P(T<=t) two-tail**) or the critical score (labeled **t Critical two-tail**) associated with the two-tailed test to make a decision regarding the null hypothesis. (See Figure 10.2 for a worked example.)

SPSS

General instructions for inputting data into SPSS can be found in Appendix C.

Data Entry

In SPSS, each row of the data file represents a participant. Since each participant is being measured twice, we will need two columns to hold the raw data. Within **Variable View**, label the two column headings using terms that will distinguish between the two conditions of the study (e.g. Exp/Control, TechA/TechB, Cond1/Cond2, etc.). Then, go to **Data View**. Input the sample data to the appropriate column, being careful to keep the data from each participant within the same row; this will be essential for creating the proper difference score. See Figure 10.3 for a worked example.

	Alcohol	Placebo
1	8	4
2	12	8
3	6	2
4	4	6
5	11	8
6	15	9
7	8	5
8	7	4

Figure 10.3 An example of entered data for a dependent-samples *t* test in SPSS.

Data Analysis

1) Click **Analyze** on the tool bar, select **Compare Means**, and then click **Paired-Samples T Test**.
2) Highlight one variable, and use the right arrow key to move it into the **Variable1** box. Move the other variable to the **Variable2** box in the same manner. (Disregard the new row of boxes that are added underneath. These are for running more than one dependent-samples *t* test at once.)

T-test

Paired samples statistics

		Mean	*N*	Std. deviation	Std. error mean
Pair 1	Alcohol	8.88	8	3.563	1.260
	Placebo	5.75	8	2.435	.861

Paired samples correlations

	N	Correlation	Sig.
Pair 1 Alcohol and Placebo	8	.770	.025

Paired samples test

	Paired differences					*t*	*df*
	Mean	Std. deviation	Std. error mean	95% Confidence interval of the difference Lower	Upper		
Pair 1 Alcohol and Placebo	3.125	2.295	.811	1.206	5.044	3.851	7

Paired samples test

	Sig. (2-tailed)
Pair 1 Alcohol and Placebo	.006

Figure 10.4 A worked example using SPSS to calculate a dependent-samples *t* test.

3) Click **OK**.
4) The output will generate three boxes. The first box will identify some descriptive statistics, including the sample means. The second box runs a correlation analysis that can be disregarded. The third box will identify, among other things we are not currently interested in, the mean difference (Mean), the estimate of the standard error (Std. Error Mean), the obtained *t* value (*t*), the degrees of freedom (*df*), and the probability of obtaining a *t* value of that size given a true null and a two-tailed test (Sig. [two-tailed]). It will not generate t_{crit}. We can find t_{crit} ourselves, or we can look at the given significance level to see if it is equal to or lower than .05. If so, we can reject the null. If it is not, we need to fail to reject the null. (Note: SPSS does not compare the t_{obt} value to a directional or one-tailed critical score.) See Figure 10.4 for a worked example.

Key Formulas

The estimate of the standard error of the difference, $s_{\bar{D}}$

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n_p}} \quad \text{(Formula 10.1)}$$

Computational formula for the standard deviation of the difference scores, s_D

$$s_D = \sqrt{\frac{\Sigma D^2 - \left[(\Sigma D)^2 / n_p\right]}{n_p - 1}} \quad \text{(Formula 10.2)}$$

Dependent-samples *t* statistic

$$t_{obt} = \frac{(M_X - M_Y) - (\mu_X - \mu_Y)}{s_{\bar{D}}} \quad \text{(Formula 10.3)}$$

Dependent-samples *t* test

$$t_{obt} = \frac{(M_X - M_Y)}{s_{\bar{D}}} \quad \text{(Formula 10.4)}$$

Cohen's *d* for dependent-samples *t* test

$$d = \frac{\textit{sample mean difference}}{\textit{standard deviation of difference scores}} = \frac{M_X - M_Y}{s_D} \quad \text{(Formula 10.5)}$$

Confidence interval for a population mean difference for dependent samples

$$LL = (M_X - M_Y) - t_{.05} s_{\bar{D}}$$
$$UL = (M_X - M_Y) + t_{.05} s_{\bar{D}} \quad \text{(Formula 10.6)}$$

Key Terms

Dependent-samples *t* Test **Counterbalancing**
Repeated-measures Design **Matched-samples Design**

Questions and Exercises

1 Which of the following describes a situation in which a dependent-samples *t* test would be the proper inferential test to use?
- **a** Two separate groups of participants are sampled from two separate populations; the samples will then be compared to see if there is a difference.
- **b** One sample of participants will be measured under two different conditions, one a control condition and one an experimental condition.
- **c** One group of participants is used to obtain one sample and compared with a known population mean.
- **d** None of the above.

2 A _____ design measures each participant under each condition of the study, while a _____ design premeasures participants, pairs them up based on this measure, and then randomly assigns them to conditions.
- **a** Repeated-measures; matched-samples.
- **b** Matched-samples; repeated-measures.
- **c** Repeated-measures; within-participants.
- **d** Within-participants; repeated-measures.
- **e** None of the above.

3 Why is it important when calculating a dependent-samples *t* test to keep each participant's data aligned across the conditions?

4 Think of a study where it would be necessary to employ counterbalancing.

5 Think of a study where it would be beneficial to use a matched-samples design.

6 The sampling distribution for a dependent-samples *t* test is a distribution of:
- **a** Mean differences gathered from two independent samples.
- **b** Mean differences gathered from two dependent samples.

c Means gathered from one population of raw scores.
d None of the above.

7 Explain the difference between s_D and $s_{\bar{D}}$. Why is this difference important?

8 Why is Formula 10.4 much more frequently used than Formula 10.3?

9 For a repeated-measures study comparing two conditions with 12 scores in each condition, what is the *df* value for the *t* statistic?
a 11
b 12
c 22
d 24

10 A dependent-samples *t* result was as follows: $t(21) = 2.31$, $p < .05$. Which of the following analyses is consistent with this statement?
a The study used a total of 23 participants and the null was not rejected.
b The study used a total of 23 participants and the null was rejected.
c The study used a total of 22 participants and the null was not rejected.
d The study used a total of 22 participants and the null was rejected.

11 Which of the following, if increased, would have little to no influence on the effect size as measured by Cohen's *d*?
a The sample size.
b The size of the sample mean difference.
c The sample variance.
d All three would clearly influence Cohen's d if increased.

12 Which of the following can be seen as a clear advantage of an independent-groups design over a repeated-measures design?
a It usually requires fewer participants to get stronger power.
b It eliminates the methodological problems associated with measuring participants multiple times.
c It eliminates the need for a two-tailed test.
d All other things being equal, there tends to be a smaller error terms, therefore larger t values.

13 Addison (1989) found that we perceive bearded men differently from non-bearded men. Compared with nonbearded men, males with beards are

rated as more masculine, more aggressive, stronger, and more dominant. The researcher used a repeated-measures design in which one group of participants rated pictures of both bearded and nonbearded men. The following data are hypothetical but were generated to reflect Addison's results regarding beardedness and masculinity. A score can range from 1 (feminine) to 10 (masculine).

a State the null and alternative hypotheses.
b Perform the appropriate inferential test.
c Identify the critical values for $\alpha = .05$, two-tailed test.
d Reject the null hypothesis?
e If so, what is the effect size?
f What type of decision error might have been made?
g Properly present the findings.
h Are there any assumptions that should give the researcher cause for concern?

Participant	Bearded	Nonbearded
P_1	10	6
P_2	8	8
P_3	5	4
P_4	7	3
P_5	10	5
P_6	6	6
P_7	5	5
P_8	10	8

14 For the same data as in Problem 13, find the 95% confidence intervals for the actual difference in masculine ratings between those images of men with beards compared with those without beards.

15 Ruth, Mosatche, and Kramer (1989) tested the hypothesis that people would state a preference for purchasing a liquor product if the product were advertised with sexual symbolism. One group of participants was shown advertisements both with and without sexual symbolism. In each condition, the participant was asked to indicate their likelihood of purchasing the product. The following hypothetical data are generated to yield results consistent with those found by the researchers. Higher scores indicate a greater willingness to purchase the product.

a State the null and alternative hypotheses.
b Perform the appropriate inferential test.
c Identify the critical values for α = .05, two-tailed test.
d Reject the null hypothesis?
e If so, what is the effect size?
f What type of decision error might have been made?
g Properly present the findings.

Participant	Sexual symbolism	No sexual symbolism
P_1	6	3
P_2	5	5
P_3	4	2
P_4	5	3
P_5	4	1
P_6	6	3

16 An industrial psychologist working for a marketing firm wants to know which of two cheeses are preferred by college students, Gouda or Swiss. After tasting both, ratings are obtained that can range from 1 (lousy) to 7 (fantastic).
a State the null and alternative hypotheses.
b Perform the appropriate inferential test.
c Identify the critical values for α = .05, two-tailed test.
d Reject the null hypothesis?
e If so, what is the effect size?
f What type of decision error might have been made?
g Properly present the findings.
h Are there any assumptions that should give the researcher cause for concern?
i Are there any methodological issues that need to be cleared up?

Participant	Gouda cheese	Swiss cheese
P_1	5	3
P_2	7	6
P_3	9	4
P_4	8	7
P_5	6	8

17 For the same data as in Problem 16, find the 95% confidence intervals for the actual difference in students' ratings between the two types of cheeses.

18 What is the advantage of being able to use a dependent-samples t test instead of an independent-samples t test? (Hint: The answer is not that the dependent-samples t test is easier to compute.)

19 A psychologist tests a new drug for insomnia. The average amount of time (in minutes) it takes participants to fall asleep is assessed before treatment, over a one-week period. These data are presented in the Pretest column of the following table. Posttest scores indicate the average time to fall asleep during the following week in which the medication is administered.

a State the null and alternative hypotheses.
b Perform the appropriate inferential test.
c Identify the critical values for $\alpha = .05$, two-tailed test.
d Reject the null hypothesis?
e If so, what is the effect size?
f What type of decision error might have been made?
g Properly present the findings.

Participant	Pretest	Posttest
P_1	120	30
P_2	60	40
P_3	90	30
P_4	100	80

20 A preschool teacher would like to make sure the students rest during quiet time. The teacher wonders if the children will relax more quickly if a story is read to them or soft music is played. For one week, a story is read every day at quiet time, and the average number of minutes it takes each child to fall asleep is recorded. For the second week, soft music is played every day at quiet time, and the same measures are taken. Assume all methodological challenges have been addressed. Data for each child are shown below.

a State the null and alternative hypotheses.
b Perform the appropriate inferential test.

c Identify the critical values for α = .05, two-tailed test.
d Reject the null hypothesis?
e If so, what is the effect size?
f What type of decision error might have been made?
g Properly present the findings.

Participant	Story	Music
P_1	6	4
P_2	6	8
P_3	9	7
P_4	8	6
P_5	8	10
P_6	10	6
P_7	12	5

21 Which of the methodological assumptions of the dependent-samples *t* test is different from the assumptions for the independent-samples *t* test?

Computer Work

For all of the following inferential tests, set $\alpha = .05$ and use a two-tailed test unless otherwise specified. Also, assume all methodological issues are properly addressed.

22 A now antiquated study referenced in the *Chronicle of Higher Education* (1990) showed students wrote papers of higher quality when they used a PC instead of a Macintosh (Mac) computer. One explanation of the finding was that the Mac was so user-friendly that students tended to write very casually. The following hypothetical data are based on a repeated-measures design. Each participant writes a paper using a PC *and* a Mac. We will find that the interpretation of the dependent-samples *t* test is consistent with the findings of the original study. Each raw score in the table represents a composite measure of the length of the paper and its quality. Higher numbers reflect greater quantity and quality. One wonders what might be found today if the study were rerun using current models of both types of computers.

PC	Mac	PC	Mac
95	80	29	32
88	70	88	66
99	88	42	42
79	54	55	39
80	80	71	65
77	87	97	84
92	75	75	72
55	34	45	65
79	72	84	77
65	70	73	56

23 A manufacturer of sunglasses wants to know if vision is affected by the color of the lens. A test of vision is administered when participants wear glasses with a blue lens and glasses with a yellow lens. Test the null hypothesis that there is no difference in vision between the two sets of glasses. Higher numbers indicate better vision.

Blue	Yellow	Blue	Yellow
12	15	14	12
7	4	9	15
22	16	8	4
16	12	21	21
14	14	12	14
10	8	11	10
17	17	11	4
16	22	10	19

24 An orthopedic surgeon is interested in whether the firmness of a mattress influences the amount of back pain experienced by patients. For one week, all participants sleep on a firm mattress and provide pain ratings each morning. A month later, the same participants sleep on a soft mattress, again providing pain ratings each morning for a week. In the following table, higher ratings indicate more back pain, with each rating an average for the week.

Test the null hypothesis that there is no difference in pain ratings between the firm-mattress condition and the soft-mattress condition. Since the surgeon desperately wants to avoid a Type I error, set alpha at .01.

Firm mattress	Soft mattress
4	8
2	2
1	7
8	10
6	6
3	6
1	4
5	7
7	9
2	5
4	7
7	10

25 Imagine an internet provider wanted to know which of two package features were more satisfying to their customers, a free music streaming account or free basic TV. The company randomly surveyed customers whose plan included both features. Satisfaction was determined by customers' numeric responses to a question for each feature. The response scale ranged from 1 (not at all satisfied) to 7 (completely satisfied). Test the null hypothesis that each feature is equally liked. Properly present the findings. Is there a potential problem with a statistical assumption? If so, which one and why?

Customer	Free music streaming	Free basic TV
C_1	6	4
C_2	5	6
C_3	7	4
C_4	5	5
C_5	6	5

(*Continued*)

(Continued)

Customer	Free music streaming	Free basic TV
C_6	5	7
C_7	7	5
C_8	7	5
C_9	6	2
C_{10}	4	6
C_{11}	7	1
C_{12}	6	7
C_{13}	7	6
C_{14}	6	3
C_{15}	5	3

26 An academic psychologist notices that some students prefer to sit in the same seat during every lecture, while others prefer to switch. The psychologist wonders if seating behavior has an effect on performance in the class. For the third and fourth weeks of the semester, all students are assigned a permanent seat for the two-week period. At the beginning of the fifth week, students are assigned to different seats, and again are reassigned to different seats at the beginning of the sixth week. Quizzes are given at the end of weeks 4 and 6. Below are the scores for 15 students on each quiz. Test the null hypothesis that seating behavior does not influence test performance. Properly present the findings.

Participant	Quiz 1 (same seat)	Quiz 2 (different seats)
P_1	2	5
P_2	9	5
P_3	2	8
P_4	7	3
P_5	4	5
P_6	8	2
P_7	10	6
P_8	9	7
P_9	6	8

(Continued)

Participant	Quiz 1 (same seat)	Quiz 2 (different seats)
P_{10}	5	4
P_{11}	8	3
P_{12}	4	6
P_{13}	6	5
P_{14}	6	7
P_{15}	7	1
P_{16}	8	7
P_{17}	9	3
P_{18}	6	5
P_{19}	5	2

11

Power Analysis and Hypothesis Testing

11.1 Decision-Making While Hypothesis Testing

In the experimental context, a hypothesis test is used to determine if there is a treatment effect. The researcher attempts to reject the null hypothesis, which always states that there is no effect due to treatment. (Even though hypothesis testing is not limited to experiments, we will use the language of experimentation as we discuss statistical power.) Hypothesis testing using inferential statistical tools is a decision-making process; as a result, there is always the possibility of making a decision error. A Type I error is made when a true null hypothesis is rejected; that is, the investigator states that there is a treatment effect when, in fact, there is not. The probability of making a Type I error is controlled directly when setting an alpha level. A Type II error is committed when a false null hypothesis is not rejected. In other words, there is a treatment effect, but the researcher does not find evidence of it and must decide that the null is still a viable explanation. Analyzing the probability of making Type I and II errors focuses attention on the *mistakes* that can be made. In this chapter, the topic of hypothesis testing is approached from another angle. Instead of asking questions about the probability of making an error, we will explore ways of improving our chances of arriving at a *correct decision*. More specifically, if there is a treatment effect, what is the probability that our test will detect it? The **power** of a statistical test is the ability it has to reject a false null hypothesis. We can think of power as *test sensitivity*. How sensitive is our inferential test to detecting an effect, if one exists? It may also be helpful to recollect the concept of power in visual enhancement systems like microscopes. The more powerful microscope, the more likely small objects will be seen when inspected. A microbe that is detected by a high-powered microscope may be unseen when inspected by one possessing lower power.

In terms of probability, there is *not* a complementary relationship between Type I and Type II errors; they do not sum to 1. If the probability of a Type I error is .05, it does *not* mean that the probability of a Type II error is .95.

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

Table 11.1 The four outcomes when rejecting or failing to reject a true or false null hypothesis.

		True state of affairs	
		H_0 is true	H_0 is false
Our decision	Fail to reject H_0	Correct $1-\alpha$ (95%)	Type II error β (20%)
	Reject H_0	Type I error α (5%)	Correct $1-\beta$ (Power)(80%)
		100%	100%

The values given to the Type I and Type II error are just examples.

These errors inhabit different "worlds." In one world the null is true, and we are measuring the chance of improperly inferring it is not (Type I error). In the other world the null is false, and we are measuring the likelihood we will not find evidence to show it to be false. (Although the errors are not complementary, there is a relationship between them; to be described in Section 11.4.) Within each given world, there is a trade-off as indicated in Table 11.1. For example, if the probability of making a Type I error is .05, the probability of *not rejecting* a true null hypothesis is .95. In other words, if there is a 5% chance of saying that there is an effect when, in fact, there is no effect, it stands to reason that there is a 95% chance of saying that there is no effect when, in reality, there is no effect.

The probability of committing a Type II error is symbolized as β (beta). If the probability of making a Type II error is β, then the probability of *correctly rejecting* a false null hypothesis is $1 - \beta$. In other words, if there is a 20% chance of failing to find an existing effect, there must be an 80% chance of correctly identifying an effect. *The probability of correctly detecting an effect is the power of the statistical test.* Therefore, power equals $1 - \beta$.

11.2 Why Study Power?

Understanding the concept of power will increase our "research IQ." For example, now that we know how *t* tests work, we have probably adopted the attitude that large samples are preferred to small samples. Large samples give us better estimates of population parameters and, therefore, make it easier to reject a false null hypothesis. However, consider this situation: A friend is going to conduct a study that evaluates the effectiveness of a new treatment for depression. Suppose it is determined, prior to the treatment, that the patient population of depressed individuals has a mean depression score that is two standard deviations above the

mean of a typical population. The goal of the treatment is to bring the depression scores back into the normal range. Our friend asks us, "How many participants should I use to see if the treatment works?" If we are only interested in determining if the treatment improves the scores to any degree, we might suggest that big samples are better than small samples. Our friend might respond by telling us that their resourcing (time, energy, finances, and access to patients) allows them to get as many as 500 participants. A sample of this size will allow our friend to detect even a very small effect. However, would finding evidence of a rather small effect be worth the time, energy, and financial resources? A better question for a researcher to ask might be, "Does the treatment *meaningfully* reduce depression?" We might decide that a meaningful reduction for the depression scale we are using is a drop in the average score of depressed individuals by more than one standard deviation. If this is the case, the chances are that we will not need 500 participants to detect such a large effect. Understanding the concept of power and the procedures of a power analysis can lead us to offer intelligent advice to our friend regarding the number of participants needed to detect an effect of a given size. In this situation, after a power analysis, we might be able to tell our friend, "If you create a sample size of 40 patients, there is an 80% chance you will be able to detect an effect size that is one standard deviation or more. Reduce your sample size; save your money, time, and resources." That would be helpful information to know ahead of time, would it not?

Here is another example of how an understanding of power can raise our research IQ. Suppose other friends are distraught over the failure of their study to reject the null hypothesis. They had tested the hypothesis that viewing aggressive pornography leads to negative attitudes toward women (which, by the way, is the case; for example, see Donnerstein, 1980; Hald, Malamuth, & Yuen, 2009; Malamuth, Heim, & Feshback, 1980). We examine the details of their research method, including the sample size, and because of our grasp of statistical power, we are able to tell them, "Given the way you conducted your study, there was very little chance of rejecting the null hypothesis, unless the effect of viewing aggressive pornography is exceptionally strong. It would be theoretically interesting, if not pragmatically important, to be able to identify even a small effect. Do not give up on the hypothesis; here is how you can modify your study…" A power analysis is valuable in planning a study as well as evaluating a study after the fact. However, clearly much more is gained by considering the issue of power before a study is conducted. It is hard to justify the time and expense that go into running a study if it has such little power that it is doomed to fail from the start.

11.3 The Five Factors that Influence Power

There are several factors influencing power, and some of these factors can be easily adjusted by the researcher. The factors are: (i) the magnitude of the difference between the means of the two populations being studied, (ii) the size of

the standard deviations of these population distributions (or sampling error), (iii) the sample size used by the study, (iv) the alpha value selected, and (v) the nature of the inferential test (one-tailed or two-tailed). The last two are decisions made by the researcher; these are decision-driven factors. The first three are data-driven factors. These can be understood by examining the mathematical mechanics of an inferential test. Think of a generic *t* test; it is composed of a numerator comparing two means and a denominator composed of a measure of the variability in the distributions as well as the sample sizes being used. Below is the formula for the single-sample *t* test. The denominator s_M has been replaced with an equivalent expression $s/\sqrt{n}$, so that each of the data-driven factors that influence power can be identified. Other inferential tests cannot be segmented this cleanly, but all inferential tests are composed of the same three elements.

$$t_{obt} = \frac{M - \mu}{\frac{s}{\sqrt{n}}} \begin{array}{l} \rightarrow (1 \\ \rightarrow (2 \\ \rightarrow (3 \end{array}$$

We will look at each of these three factors separately. As a working example to help us explore them, let us suppose we want to evaluate the effectiveness of a study program designed to increase scores on the Scholastic Achievement Test – Verbal (SAT-V). Suppose we know that the national mean for the SAT-V is 500. We want to know whether our sample of students, once having participated in the program, will now show evidence that the program has helped improve their SAT performance. The null and alternative hypotheses can be stated as

$$H_0 : \mu = 500$$

$$H_1 : \mu \neq 500$$

Although the alternative hypothesis is stated specifically, $\mu \neq 500$, it is a rather broad statement. We may reject H_0 and accept H_1 when μ is 550, and, under certain circumstances, we may reject H_0 and accept H_1 when μ is 501. Therefore, with H_0 stated as $\mu = 500$, the null hypothesis can be correct, a little bit incorrect, or very incorrect. It may not surprise us to learn that the likelihood a statistical test will find evidence of a difference increases as the difference between the null and alternative means increases. (Think about it this way: given equal error terms for denominators, the larger the difference between means in the numerator, the larger the resulting *t* value. Large *t* values are more likely to fall into the rejection region.) In other words, the power of the test will be influenced by the size of the difference between the means. A researcher, wanting to increase power, then, should create conditions that are believed to maximize the difference between means. In our example, assuming more time in the study program leads to greater SAT-V improvement, the power to show the

difference in an inferential statistical test should increase as participants spend time in the study program.

The second data-driven factor is the amount of variance in the distributions of raw data, which is reflected in the size of the standard deviation. If the two populations being sampled have similar means, high power can still be achieved if the variability of the scores in the populations is low. This concept is hard to represent in the SAT-V example because the standard deviations are known to be around 100. However, if we imagine the distribution of SAT-V scores to be much more narrow (e.g. $\sigma = 10$), then an improvement of only 5 points or so would be much more easily noticed. We will not spend too much time with this factor since it is usually very hard, if not impossible, for the researcher to influence the standard deviations of populations or the samples that are drawn from them. Nonetheless, as the researcher is able to shrink the standard deviation of the scores, power to detect differences between population means increases.

The third data-driven factor that influences power is the sample size. As the sample size increases, the estimate of the standard error decreases. If we are having trouble understanding this, imagine a situation where the sample size is so large that it is equal to $N - 1$ (one less than the total population). If a sampling distribution were created using a sample of that size, would not the standard error be virtually zero? Each and every sample mean found would be virtually identical. As n increases, the error term decreases. If the null hypothesis is false, t tests utilizing small error terms are much more likely to generate large t values. As the standard error decreases, power is increased.

The first two data-driven factors, the difference between the means of the populations and the standard deviation(s) of the distribution(s), are often combined and referred to as the "treatment effect." Recall that the null hypothesis can be false to varying degrees. The extent to which the H_0 is false is the size of the effect *in the population*. When determining the power of an anticipated statistical test, the size of the effect must be specifically stated ahead of time. The size of the effect is stated as the number of standard deviations the true population mean is from the null population mean. Since we do not know the true population mean (is it the null? is it something else?), questions of power take the form of "what if" questions. "*What if* the true population mean is 0.25 standard deviations from the null population mean, then what is the probability of detecting that difference with this inferential test in this research situation?" or "*What if* the true population mean is 0.89 standard deviations from the null population mean, then what is the probability of detecting that difference?" Obviously, we want to maximize our chances of discovering an effect in the population if one exists.

The second two data-driven factors, the standard deviation(s) of the distribution(s) and the sample size, can also be thought of as being combined; these two constitute the standard error. From this perspective, the researcher can think about a proposed study in terms of the difference between the means

(1) and the standard error (2 and 3). There is not *one* proper way to conceptualize the data-driven factors that influence power, but it is important for researchers to develop some strategy for thinking about measuring and adjusting the power of a study prior to running it.

11.4 Decision Criteria that Influence Power

As noted earlier, in addition to the data-driven factors, power is also influenced by policies adopted for decision-making; namely, the selection of an alpha level and Type of t test (one- or two-tailed). The more we avoid the risk of a Type I error, the more likely it is that we will make a Type II error if the null hypothesis is false. Smaller alpha values correspond to more extreme t_{crit} values, which, in turn, require larger t values to reject the null hypothesis. In these situations, a false null is harder to detect. This is why the decision regarding what is an acceptable Type I error rate influences the chance of making a Type II error. Some researchers may try to have their cake and eat it too by setting a low alpha value but then attempt to offset the implications of this decision by placing their entire rejection region in one tail of the distribution, thereby making the t_{crit} value a bit closer to zero and reducing the Type II error rate. (Recall the many concerns related to the use of one-tailed t tests discussed in Section 9.3.)

A hypothetical study based on our working example is presented to further clarify the concept of power and illustrate its calculation. The manner in which we go about computing power will vary, depending on the experimental design and type of statistical analysis used. The illustrative worked example below involves a one-tailed, single-sample z test.

Worked Example

A private company would like to offer a college entrance preparation course that will help students increase their SAT-V (verbal) scores. Before marketing the program, the company decides to evaluate the program's effectiveness. The president of the company comes to us with the following problem: "Our company is planning to evaluate the effectiveness of a course that we have developed. This course is designed to increase the SAT-V test scores of students, and, as you may know, the national average of the SAT-V is 500, with a standard deviation of 100. We plan to sample 36 graduates of our program and determine their mean SAT-V. We are not particularly interested in discovering if our program increases scores by an average of only 5 or 10 points. However, we would like to be fairly certain that our study will detect a 25-point difference. Given the information I have provided, what is the probability that we will detect an effect size of 25 points or more?"

The president has given us all the information we need to answer the question. As we work toward the solution, examine the problem from the perspective of the sampling distributions implied by the problem. Figure 11.1 depicts two sampling distributions. Figure 11.1a represents the case when H_0 is true; the distribution in Figure 11.1b represents the case when H_0 is false by 25 *raw points*. Given the findings of the study seem to be only meaningful if the training program increases performance, and given our desire to minimize the Type II error rate, we have decided to use a one-tailed z test. (Temporarily set aside any misgivings we may have about one-tailed tests for the purpose of this worked example.) Since we are conducting a one-tailed test, the critical value when $\alpha = .05$ is +1.65 (instead of ±1.96). Figure 11.1a is the null distribution because it depicts the sampling distribution of means when H_0 is true. Figure 11.1b is the alternative distribution because it illustrates the sampling distribution of means when the null hypothesis is incorrect by 25 raw points.

The first step in computing power is to identify the sample mean that corresponds to the critical value, +1.65. This sample mean sits 1.65 standard errors

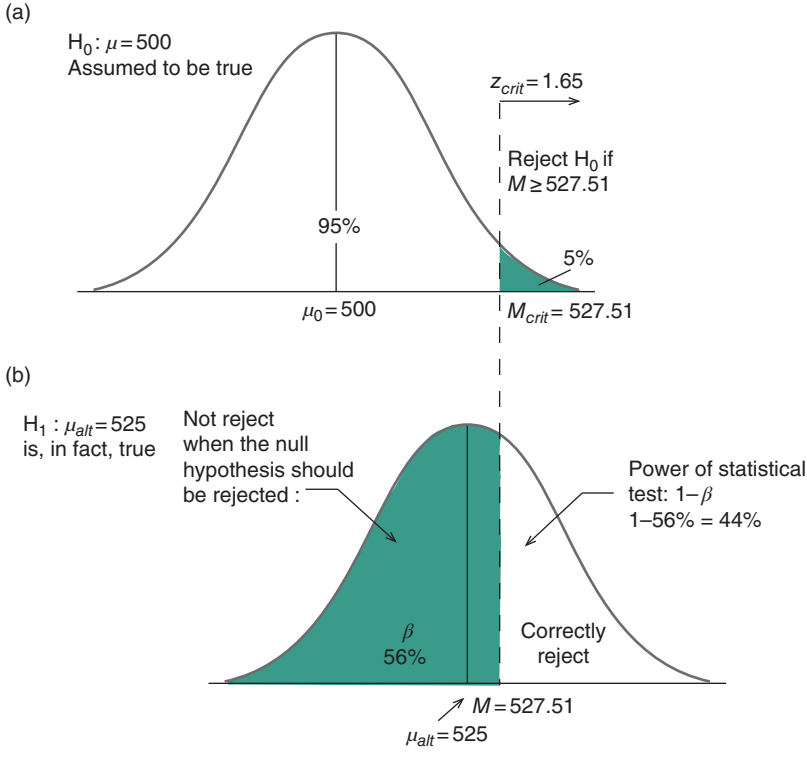


Figure 11.1 Sampling distributions of means. (a) Under $H_0: \mu = 500$. (b) Under $H_1: \mu = 525$.

above the mean of the null sampling distribution. Since the cutoff is +1.65, the formula used to identify the sample mean that is 1.65 standard errors above the mean of the null hypothesis is

$$M = \mu + (z_{crit})\sigma_M$$
$$M = 500 + (1.65)\left(\frac{100}{\sqrt{36}}\right)$$
$$M = 500 + (1.65)(16.67)$$
$$M = 527.51$$

All sample means equal to or larger than 527.51 fall into the rejection region of the null sampling distribution. Note that we are using the null sampling distribution to define the rejection region since it is the null hypothesis that is directly tested. If the null hypothesis is true, 5% of the means of the sampling distribution in a one-tailed test will fall above the mean of 527.51.

Now look at the alternative distribution in Figure 11.1b. Even though the true population mean is 525, obtaining a sample mean of 525 would not place it in the rejection region of the sampling distribution under the null hypothesis. We would need a sample mean equal to or greater than 527.51 to reject the null hypothesis. Now, *if* the true population mean is 525 and we need a sample mean greater than 527.51, what is the probability, when we run our study, of randomly selecting a mean from the *alternative sampling distribution* that falls at or above 527.51? It is another *z score* problem. A new symbol is specified here, μ_{alt}, to indicate that we are using the mean of the alternative sampling distribution. First, using the population mean of the alternative sampling distribution, transform 527.51 into a z value.

$$z = \frac{M - \mu_{alt}}{\sigma_M}$$
$$z = \frac{527.51 - 525}{16.67}$$
$$z = 0.15$$

The number 0.15 is the number of standard error units 527.51 is above the mean of the alternative distribution, 525. Next, using the third column of the z table (Table A.1 in the Appendix), look up a z of 0.15. We find that the probability of randomly selecting a sample mean above *the critical mean of the alternative distribution* is .44. (See the unshaded area of the alternative distribution in Figure 11.1b.) Therefore, the power of the test is .44. Forty-four percent of the sample means of the *alternative distribution* fall above the critical sample mean of 527.51. In other words, the probability of rejecting the null hypothesis in this particular research scenario is .44. This means that if we were to run this

identical study multiple times, less than half of the time would we expect to get a *t* value that would allow us to reject the null: the other times we would be failing to reject the null and (unknowingly) making at Type II error. Of course, all of this analysis is *assuming* that the effect of the training program is an increase of 25 points.

With the power analysis completed, we are now in a position to advise the president of the company offering the SAT-V study program. Do we think the study should be conducted as proposed? There is only a 44% chance of detecting the effect of interest. In fact, there is a higher probability of making a Type II error (.56) than there is of correctly rejecting the null hypothesis (.44)! We should probably advise against it.

At this point two questions can be raised. First, what can the company do to increase the power of the test? To answer this question, think about the three data-driven factors that influence power. Can any of them be altered? Well, perhaps the study program can be improved, thus increasing the hypothesized difference between μ_{null} and μ_{alt}. More easily, however, the sample size of 36 can be increased.

Second, how much power is enough? There is no set or easy answer to this question, but most researchers would like power to be in the range of .70–.90 with .80 as the conventional probability value. When a statistical test has a .80 probability of correctly rejecting the null hypothesis, there is a .20 probability of making a Type II error. We may note that a probability of .20 is four times greater than the conventional .05 probability of making a Type I error. Researchers are *usually* comfortable with that kind of trade-off. It reflects a conservative approach to hypothesis testing; concluding that there is an effect when there is no effect is considered a more serious mistake than failing to identity an effect when, in fact, there is one. However, each separate test of a hypothesis requires an investigator to consider what the acceptable probabilities are for different types of errors. The nature of the consequences for each type of error will guide the decision-making process.

11.5 Using the Power Table

The preceding section discussed how to compute the power of a one-tailed single-sample *z* test. The problem required knowing the sample size, the population standard deviation, the mean hypothesized under the null hypothesis, and a mean specified under the alternative hypothesis. There is a shortcut method, which can be used to arrive at the power of an inferential test. A couple simple calculations and a power table is all that is needed. Let us use the data from the worked example to illustrate this method for determining the power of a statistical test.

To use the power table (Table A.3), two new terms must be introduced: δ (delta) and γ (gamma), the symbol for the size of the treatment effect.

Formula for delta

$$\delta = \gamma\sqrt{n} \quad \text{(Formula 11.1)}$$

where

γ = the size of the treatment effect

Note that δ combines the size of the hypothesized treatment effect (γ) and the sample size (n) into a single measure. Delta will be used by the power table (Table A.3) to determine the proportion of the alternative sampling distribution that falls into the rejection region of the null sampling distribution (recall Figure 11.1).

Formula for gamma

$$\gamma = \frac{\mu_{alt} - \mu_0}{\sigma} \quad \text{(Formula 11.2)}$$

where

μ_{alt} = the mean of the alternative (treated) population
μ_0 = the population mean when the null hypothesis is true
σ = the population standard deviation; if σ is unavailable, use s

Note in Formula 11.2 that γ, the size of the treatment effect, determines the number of standard deviations between the population mean hypothesized by the null and the population mean specified when the null hypothesis is false (the proposed *true* population mean). If, by chance, the smaller mean is subtracted from the larger mean in the numerator, gamma will be a negative value. However, the resulting power value is always understood to be positive. There is no concept in statistics called negative power.

Keep in mind, this process does not *discover* the true population mean; rather, it is hypothetically specified ahead of time. A power analysis does not identify the difference between population means, and it does not test to see if there is a difference between population means. The power analysis merely calculates the probability of detecting a specified difference between population means (assuming that specified difference exists). Here is how the shortcut method works.

The steps below use the data from the worked example to demonstrate how to use the shortcut formulas.

Step 1. Use Formula 11.2 to specify an effect size, γ (gamma).

$$\gamma = \frac{\mu_{alt} - \mu_0}{\sigma}$$

$$\gamma = \frac{525 - 500}{100}$$

$$\gamma = \frac{25}{100}$$

$$\gamma = 0.25$$

The number 0.25 is the number of standard deviation units the treated population mean is from the mean of the population stated in the null hypothesis.

Step 2. Compute delta using Formula 11.1.

$$\delta = \gamma\sqrt{n}$$

$$\delta = .25\sqrt{36}$$

$$\delta = 1.50$$

Step 3. Determine the power by using the power table (Table A.3). In the worked example, use a one-tailed test, with $\alpha = .05$. With δ computed as 1.50, find δ in the first column of the table, and move to the column for $\alpha = .05$, one-tailed test. The power of the test is .44, the value found in the original worked example.

As a review, let us use the power table to demonstrate the relationship between effect size and power as presented in Section 11.3. Suppose the effect that we want to detect is actually 50 raw score units from the population mean of 500. The power question now becomes, "What is the power of my test to detect a population difference of 50 raw score units?" Use the data and sample size of the worked example.

$$\gamma = \frac{\mu_{alt} - \mu_0}{\sigma}$$

$$\gamma = \frac{550 - 500}{100}$$

$$\gamma = 0.50$$

$$\delta = \gamma\sqrt{n}$$

$$\delta = 0.50\sqrt{36}$$

$$\delta = 3.00$$

Now go to the power table with $\delta = 3.00$, for a one-tailed test when $\alpha = .05$. The power of the test jumps to .91. If the population mean for the alternative hypothesis is 550, we now have a 91% chance of detecting the effect.

Let us try one more. Recalculate δ when $\mu_{alt} = 505$. The power of the test drops to .12. There is only a paltry 12% chance of detecting a false null hypothesis if the effect of the training course creates a new population mean is 505. These exercises demonstrate how effect size (factors 1 and 2) influences the power of a statistical test. As the effect size increases, so does statistical power.

These hypothesized effect sizes do not need to be drawn out of thin air. Many times researchers establish hypothesized effect sizes based on existing theory and/or previous research findings (see Section 11.6 for a broader discussion). Please note that power analyses have nothing to do with what we *hope* to be the effect size. A power analysis essentially says, "Given a particular effect size, alpha, test type, and sample size, here is the probability that the null hypothesis will be correctly rejected." This information can be very valuable to a researcher who is contemplating the start of a research project.

11.6 Determining Effect Size: The Achilles Heel of the Power Analysis

Researchers can directly control the alpha level and the type of test to use (one- or two-tailed), and, within reason, they can select the sample size. The problem arises when they have to specify the size of the effect. The difficulty inherent in postulating an effect size might be one reason why researchers do not routinely perform a power analysis before conducting a study. However, we are not rendered helpless when having to state an effect size. There are two good sources of information for making a reasonable statement about effect size.

First, most studies have a heritage. Although we may be the first to conduct this exact study, it is likely that others have been working in this area as well, studying the same phenomenon. By familiarizing ourselves with effect sizes other researchers have detected, we can make an estimate of the effect size that applies to our experiment. It is true that this effort is hampered by unwillingness in the scientific community to publish failed findings (see Boxes 8.1 and 11.1). Nonetheless, we can oftentimes gain a better understanding of what effect size we should expect by familiarizing ourselves with the existing published literature on our topic. Second, we may be able to conduct a pilot study. A pilot study is a trial run of the study we wish to conduct. It is conducted with a sample smaller than we will ultimately use and provides an opportunity to adjust experimental procedures. By examining the data of the pilot study, we can get a feel for the strength of our independent variable, as well as the amount of variability in the data.

For example, one of the authors of this textbook (Grimm), as a graduate student, was involved in a research project to examine self-control techniques that could be taught to people to help them tolerate pain. A pilot study was run where it was discovered there is a tremendous variability among people in their ability (willingness) to tolerate a painful stimulus. The sample data allowed us to take a sneak peek at the variability relating to pain tolerance. As we recall from Section 11.3, increased variability reduces power.

Box 11.1 Is the Scientific Method Broken? The Need to Take Our Own Advice

Using the 1960 volume of the *Journal of Abnormal and Social Psychology*, Cohen (1962) conducted an interesting study. Although the authors of the articles in that volume did not use power analyses, Cohen computed the power of the statistical tests used in each of the studies. According to Cohen's (1962) early guidelines, a small effect size is about .20; a medium effect size is around .50; and a large effect size is around .80. Assuming that the researchers would want to detect a medium treatment effect size, the average power of the tests in that volume was .46. This means that, on the average, there was only a 46% chance of detecting a medium effect size! This realization prompts us to contemplate just how many other studies might have been potentially included in the literature but were abandoned, perhaps prematurely, because not enough power was marshaled to detect a treatment effect. Cohen's admonition to use power analyses became widely known among researchers (Cohen, 1977; Sedlmeier & Gigerenzer, 1989). Yet, 24 years later, a study similar to Cohen's (1962), using the 1984 volume of the *Journal of Abnormal Psychology*, found that the average statistical power for detecting a medium effect size had actually gone down to only .37 (Sedlmeier & Gigerenzer, 1989)! There is little reason to believe the situation is much different today. Despite the urging of statisticians, power analyses have not become a standard practice among researchers.

If a performed and reported power analysis were to become a standard step in the research process, it would both give studies that are investigating a genuine treatment effect a better chance of finding supporting evidence and give studies that end up failing to reject the null hypothesis more validity. For instance, if a study was set up to detect a small difference and the appropriate statistical power was generated for the test, then a finding of a failure to reject the null might be seen as theoretically and practically important to other researchers. For one thing, it might keep others from spending time and energy asking the same question, and, secondly, it might stimulate the development of different ideas about how that part of the world works. Unfortunately, power analyses tend not to be performed, and null findings tend not to be published (see Box 8.1). Scientists can be a stubborn bunch, and the standards of the scientific process can be hard to change. However, with each new generation of scientists comes a new opportunity to do things differently. Will a new generation of researchers choose to use the tools of power?

Formula 11.2, the formula for gamma, mathematically shows how increased variability decreases the effect size.

$$\gamma = \frac{\mu_{alt} - \mu_0}{\sigma}$$

Because the variability in the sample data was so large, it was clear that the treatment effect was going to be rather small. Now consider the formula for δ (Formula 11.1):

$$\delta = \gamma\sqrt{n}$$

We know that a larger delta is associated with greater power. If gamma is small, how can delta be increased? There is only one option: the sample size needs to be increased. Therefore, because of the pilot study, we were able to determine that our treatment effect was likely to be small, necessitating the need for a large sample size to detect the treatment effect we believed to be present.

11.7 Determining Sample Size for a Single-Sample Test

In this section, we will discuss how to explicitly arrive at a sample size that will allow a researcher to detect a given effect size. Sometimes we will not have any reliable information to help us specify a treatment effect size. In these situations, we can still perform a power analysis. All we need to do is ask ourselves, "What size of a treatment effect do we *want* to detect?" Once this is answered, the follow-up question becomes, "How many participants do we need to detect an effect of this size?"

To determine the sample size required to detect a specified effect size, we will need to state, beforehand, α, γ, the desired power, and our preference for a one- or two-tailed statistical test. Alpha is usually set at .10, .05, or .01, with .05 being the most common value for alpha. Desired power can be set at any value from just above 0 to just below 1. However, setting power low, for instance, .20, is saying that we will accept a 20% chance of detecting a treatment effect of that given size. (Stated in other terms, we are accepting an 80% chance of making a Type II error if the null hypothesis is false.) With the odds so low of correctly rejecting the null hypothesis, the study will most likely be a waste of time. Why not set the desired power at .99? The problem here is similar to the trade-off between Type I and Type II errors that was discussed in Chapter 8. Recall that in trying to minimize a Type I error, we could set alpha at .0001, but the probability of a Type II error would then become unacceptably large. Likewise, if we insist that the power of our statistical test be .99, we will pay a big price in the vast number of participants required. As stated earlier in the chapter, the conventional compromise is to set power at .80.

Let us return once again to the worked problem in which we are a consultant to a company that wants to market a course for improving SAT-V scores. Recall that the population mean is 500 with a standard deviation of 100; the company wants to detect an average increase in SAT-V scores of 25 points (525). The company proposed to use 36 participants. We found that if only 36 participants

are used, there is only a 44% chance of detecting an effect. Since 44% is unacceptably low, we now must advise the company how many participants should be used to detect a 25-point difference.

Formula for determining sample size for a single-sample *t* test

$$n = \left(\frac{\delta}{\gamma}\right)^2 \quad \text{(Formula 11.3)}$$

Remember that γ is not the number of points between 500 and 525; rather, γ is the number of standard deviations 525 is from 500. In our example, γ was calculated as 0.25: (525 – 500)/100 = 0.25. Now we need delta, δ, to find n in Formula 11.3. We can use Table A.3, locate the desired power in the proper column, and work backward to determine δ, or we can use Table A.4, input the desired power, and let the table determine δ. Let us use Table A.4 since it generates a δ that is more precise. Since our desired power is .80, look down the left-hand column and find .80. The column immediately to the right is for a one-tailed test, when $\alpha = .05$. Here we find $\delta = 2.49$. We now have the values needed to use Formula 11.3 to determine the needed sample size.

$$n = \left(\frac{\delta}{\gamma}\right)^2$$

$$n = \left(\frac{2.49}{.25}\right)^2$$

$$n = 99$$

To detect a 25-point difference in SAT-V scores, the study must include 99 participants.

■ **Question** *Suppose we wanted to be almost certain of detecting a treatment effect size of .25, so power is set at .99. How many participants would we now need?*

Solution Using Table A.4 to find δ for a power of .99, δ is found to equal 3.97.

$$n = \left(\frac{3.97}{0.25}\right)^2$$

$$n = 252$$

Increasing the probability of detecting a 25-point difference between the populations from .80 to .99 requires using 153 more participants. If participants are easy to come by and the cost of gathering data is low, the investigator may want to use a level of power greater than the conventional 80%. ■

11.8 Failing to Reject the Null Hypothesis: Can a Power Analysis Help?

Hypothesis testing is a probabilistic endeavor. Whether we are referring to a Type I error, a Type II error, or power, there is always some probability associated with every aspect of hypothesis testing. Moreover, there is no way to *prove* the null hypothesis. If we do not identify an effect, we say that we have "failed to reject" the null hypothesis. We do not claim that we have proved the null hypothesis to be true; we cannot even say that we have evidence that the null hypothesis is true. All we can say is that the null hypothesis was not rejected.

Researchers are usually in the position of wanting to reject the null hypothesis. Rejecting the null hypothesis typically means that an effect has been identified, which is something to report to other researchers. However, what do we say if we fail to find an effect? Studies that fail to reject the null hypothesis are often assigned to the circular file (the wastepaper basket). (See Boxes 8.1 and 11.1 for brief discussions as to why this practice is unfortunate.) However, a researcher may attempt to learn from a failed study, by either redesigning it or adjusting the hypothesis. A power analysis of a failed study may show that there was low power in the original design to detect the hypothesized effect. If so, the study can be run again with more participants. However, what if the researcher conducts a power analysis *before* the study, conducts the study so that the power of identifying a *small* effect is, say, .80, and then fails to reject the null hypothesis? Is this a potentially important finding? Perhaps. Although the researcher could not conclude that the null hypothesis is true, if the study is shown to have high power, it could be asserted that if there is a difference in population means, that difference is likely to be very small and perhaps not worthy of additional research. Even *null* findings like this can be important.

Consider a different example. There are times when failing to reject the null hypothesis can have important theoretical and/or practical significance. Suppose a researcher wants to see if there is a decrease in intellectual performance when antipsychotic medication is administered to schizophrenics. One sample of schizophrenic patients receives the drug, and another sample receives a placebo. The patients are later tested, and the two groups were found *not* to differ in intellectual performance. This is a useful finding but *only if the analysis has sufficient power to detect a small difference*. Box 11.2 more closely examines this issue by presenting a study where the researchers argue that a failure to reject the null hypothesis has theoretical value.

This chapter provides us with only a brief introduction to the topic of power. Power analyses can be conducted in many more test situations than are covered in this textbook. The interested reader is referred to textbooks pertaining to advanced statistical analysis.

Box 11.2 Psychopathy and Frontal Lobe Damage

Psychopathy (more generally referred to as antisocial personality disorder) is a diagnostic label applied to people who exhibit a reckless disregard for the rights of others, an inability to maintain relationships, irresponsibility, lying, lack of remorse for transgressions, interpersonal manipulations, and an inability to sustain employment. Some clinicians have noted a behavioral similarity between psychopaths and individuals who have frontal lobe damage (Elliott, 1978; Schalling, 1978), although the connection may not be as firm as once believed (e.g. Brower & Price, 2001).

A study by Gorenstein (1982) seems to support the notion that psychopathic behavior is, in part, due to cognitive deficits associated with the frontal lobe of the cortex. (One function of the frontal lobe is to inhibit impulsivity.) Gorenstein administered a number of problem-solving tasks to a group of psychopaths and two control groups. These tasks have previously shown sensitivity in identifying frontal lobe damage. Gorenstein found a significant difference between the means of the psychopathy and control groups, with the psychopaths showing poorer performance.

Robert Hare, a well-known researcher in the area of psychopathy, criticized the findings of Gorenstein. Hare raised two important issues (Hare, 1984). First, Hare questioned the adequacy of Gorenstein's diagnostic methods for classifying individuals as psychopaths. Second, Hare pointed to a confounding variable: there were a disproportionate number of individuals with substance abuse problems in the psychopathy group (85%). Using a different research design, Hare administered the same problem-solving tasks to three groups of participants classified as high, medium, and low on measures of psychopathy. Hare failed to reject the null hypothesis for any of the problem-solving measures. In other words, there was no statistical evidence of a difference found between the groups. Hare concluded that "…there is little support for the position that psychopaths have specific cognitive deficits in the processes associated with frontal lobe functioning" (p. 139).

Note that Hare is making an interpretation of the null hypothesis, which is a risky proposition since the null hypothesis cannot be proven true. However, let us see if a power analysis can justify Hare's "assertion of the null hypothesis."

Power Analysis

As previously noted, the power of a statistical test is influenced by several factors, including the treatment effect size. One difficulty in conducting a power analysis is the specification of the treatment effect size. In 1988, Cohen offered revised guidelines for interpreting effect sizes: .10 (small), .25 (medium), and .40 (large). The larger the effect size in the population, the more powerful the statistical test.

Suppose we postulate a medium effect size (using current standards) and ask the following question: "What is the probability that Hare's analysis would detect an effect size of .25?" Using the alpha level and sample size stated in the article, power was determined to be .299.[1] In other words, Hare had only a 30% chance of detecting a medium effect size. Accordingly, there was a 70% chance of making a Type II error; that is, failing to reject a false null hypothesis. When it is considered that power should ideally be approximately 80%, Hare's significance test seems inadequate. About 150 participants would have had to be included in the study to have had an 80% chance of detecting a medium effect size. Hare used only 46 participants.

Interpreting a failure to reject the null hypothesis as saying something important about the phenomenon under study, when there is only a 30% chance of rejecting the null hypothesis, is difficult to justify. However, in conducting our power analysis, we selected an effect size of .25. Statisticians recommend a useful strategy for specifying effect size: find other studies in the same area, and estimate the effect size from their sample data. Since Hare's study was based on the findings of Gorenstein, there is information available to estimate the effect size in the population.

Based on the group means, standard deviations, and the number of participants per group reported in Gorenstein's article, the estimated effect size was found to be .56, much larger than .25. Now, what implications does this have for the power of Hare's statistical analysis? Even with only 46 participants, the probability of detecting an effect size of .56 is .93! With a 93% chance of correctly rejecting the null hypothesis, and only a 7% chance of failing to reject a false null hypothesis (Type II error), there is a much stronger justification for interpreting the importance of failing to reject the null hypothesis.

Returning to the original intent of the research, what does all of this mean about psychopathy and frontal lobe damage? Can a researcher conclude that psychopaths do not have frontal lobe damage? In other words, is the null hypothesis true? There is no way of knowing. What *can* be asserted is that it is highly unlikely that there is a *large* difference between psychopaths and normal people in frontal lobe damage, at least as measured by the tests used in this

1 This power analysis was performed using Borenstein and Cohen's software package (1988), *Statistical Power Analysis: A Computer Program.* Obtained means and standard deviations for the Wisconsin Card Sorting Test were taken from the Gorenstein (1982) and Hare (1984) articles. A power analysis using the other "significant" dependent variables in the Gorenstein study would not alter the present findings.

research. On the other hand, since Hare's statistical power was inadequate for detecting a medium effect size, a researcher cannot conclude that there is not a moderate difference between the groups in cognitive functioning.

The plausibility of "asserting the null hypothesis" lies on a continuum. It is impossible to prove that there is no effect in the population. However, as the power of a statistical test to detect smaller and smaller effect sizes increases, an investigator can persuasively argue, "If there is an effect, it is probably quite small, and most likely trivial."

Readers should also know that power analysis software exists online. One very handy program is called G*Power. It is a free download and is easy to learn how to use. It allows users to select the type of test to be run (e.g. *t* tests) and the specific output variables needed (e.g. sample size, power, effect size). Once selected, the program asks for the necessary input variables and then generates the requested values.

Summary

The probability of rejecting a false null hypothesis is the power of a statistical test. Power is greatest when the magnitude of the treatment effect is large. (The treatment effect is the number of standard deviations between μ_{alt} and μ_0.) Increasing the sample size in a study will also increase the power of a statistical test. Furthermore, two features of the decision criteria for hypothesis testing influence power: the desired alpha value (Type I error rate) and the choice of a one- versus two-tailed statistical test.

The conventional figure for the desired power of a statistical test is .80, meaning there is an 80% chance of detecting a specified effect (if one exists). There is a complementary relationship between power and a Type II error (β) in that power equals $1 - \beta$. If an investigator wants to advance a substantive interpretation about the importance of failing to reject a null hypothesis, it is essential that there has been sufficient power to detect a meaningful effect size.

When calculating power, there are formulas and tables designed to aid in quick analysis. There are also helpful and free resources available online such as G*Power.

Key Formulas

Formula for delta

$\delta = \gamma\sqrt{n}$ (Formula 11.1)

Formula for gamma

$\gamma = \frac{\mu_{alt} - \mu_0}{\sigma}$ (Formula 11.2)

Formula for determining sample size for a single-sample *t* test

$n = \left(\frac{\delta}{\gamma}\right)^2$ (Formula 11.3)

Key Term

Power

Questions and Exercises

1 Power is the ability of a statistical test to:
- **a** Incorrectly fail to reject null hypotheses.
- **b** Incorrectly reject null hypotheses.
- **c** Correctly fail to reject null hypotheses.
- **d** Correctly reject null hypotheses.

2 How are the lenses of a microscope analogous to the concept of statistical power?

3 If β equals .10:
- **a** The Type I error rate = .10.
- **b** The Type II error rate = .10.
- **c** Alpha = .10.
- **d** Alpha = .90.

4 As the power of a statistical test increases, what happens to the Type I error rate?

5 As the power of a statistical test increases, what happens to the Type II error rate?

6 How is a hypothesized treatment effect calculated?

7 How does the size of the treatment effect influence power?

8 Compute treatment effect sizes for each of the following problems.
a $\mu_0 = 300$, $\mu_{alt} = 345$, $\sigma = 70$
b $\mu_0 = 300$, $\mu_{alt} = 345$, $\sigma = 20$
c $\mu_0 = 300$, $\mu_{alt} = 310$, $\sigma = 20$
d $\mu_0 = 300$, $\mu_{alt} = 310$, $\sigma = 50$

9 A researcher is interested in studying the effects of sleep deprivation on cognitive performance; however a power analysis performed on pilot data shows low power for detecting an effect for a loss of 3 hours of sleep (power = .24). If the researcher does not have enough money to increase their anticipated sample size, what other option do they have to increase power and demonstrate the relationship between loss of sleep and cognitive performance?

10 How does sample size influence the power of an inferential test?

11 Using the numbers found in Problem 8, how many participants would be needed to detect each effect in the population? (Assume $\alpha = .05$, two-tailed test, and desired power is .80.)

12 A researcher conducts several studies and performs single-sample t tests with each set of the following summary data. For each case, compute the power of the statistical test. Use s as an estimate of σ. (Assume $\alpha = .05$, two-tailed test, and round off delta to the nearest first decimal place.)
a $\mu_0 = 130$, $\mu_{alt} = 120$, $s = 15$, $n = 10$
b $\mu_0 = 130$, $\mu_{alt} = 120$, $s = 15$, $n = 40$
c $\mu_0 = 50$, $\mu_{alt} = 52$, $s = 10$, $n = 15$
d $\mu_0 = 50$, $\mu_{alt} = 52$, $s = 10$, $n = 100$
e $\mu_0 = 25$, $\mu_{alt} = 30$, $s = 7$, $n = 30$

13 A researcher is interested in running a power analysis before the research is started. Use the same means and standard deviations in Problem 12, parts a, c, and e, to determine what sample size is needed for each study to achieve power = .80, $\alpha = .05$, two-tailed test. (Round to the nearest whole integer.)

14 Which of the following three factors influencing power does the researcher typically have the least ability to adjust? Why?
a The difference between the null and hypothesized mean.
b The sampling error.
c The sample size.

15 How does increasing or decreasing alpha influence power?

16 How does a choice of a one- versus two-tailed statistical test influence power?

17 Formulate a scientific hypothesis in which support for the null hypothesis would have theoretical or practical significance. How would a researcher use power-related terminology to impress a journal editor if the null hypothesis is not rejected?

Part 4 Review

The *z* Test, *t* Tests, and Power Analysis

Review of Concepts Presented in Part 4

The purpose of this brief review section is to revisit both the similar concepts that hold Chapters 8–11 together and the concepts that distinguish them one from another. First let us look at the similarities. All four of the inferential tests presented in these chapters (single-sample *z* test, single-sample *t* test, independent-samples *t* test, and dependent-samples *t* test) are based on the same basic logic. That is, each one is designed to test a null hypothesis of no difference by taking a found difference between means and interpreting that difference in terms of the amount of sampling error we might expect to find if the null were true. In mathematical terms, a ratio is created where a difference between means is placed in the numerator and a measure of sampling error is placed in the denominator. The resulting value of this ratio is then compared with all of the values one might expect to find if we were working under conditions where the null is true, that is, where no difference between the population means actually exists. If it is determined that the difference between the means is not likely to be explained by sampling error, a "treatment effect" is said to occur. In these situations, a follow-up analysis measuring the size of the treatment effect can be run. Cohen's *d* statistic measures the size of a mean difference in standard deviation units, and various versions of this statistic exist that correspond to the various *z* and *t* tests.

Another common theme running through these tests is that each one is an inference, that is, an extrapolation from known sample data to unknown population data. As a result, there is always a likelihood that the sample data used for the test may misrepresent the population and render a misleading

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

conclusion. Nonetheless, a decision logic is employed that leads the researcher to cautiously conclude that either the null is wrong or the null may be true. This means that two types of decision errors can occur. A Type I error occurs when we conclude there is a difference or treatment effect when, in fact, there is none, and a Type II error occurs when we do not conclude there is a difference or treatment effect when, in fact, there is one. To help avoid making Type II errors, the concept of a power analysis is introduced in Chapter 11. This is an a priori technique using several factors, two key ones being the expected sample size and the expected treatment effect size. The analytical tool is designed to help researchers predict ahead of time how likely a sample mean will be generated that will allow the null to be rejected. If the likelihood is low, the researcher may opt to not conduct the study (thereby saving time and energy by moving on to a more promising area of investigation), or they may decide, prior to running the study, to make adjustments to the research design that would improve the chance of gathering data that would lead to a rejection of the null.

Each inferential test presented, however, is different and is used under different methodological situations. Chapter 8 tests compare a known population mean with a gathered sample mean. The null claims the sample mean came from the same population that produced the given population mean. For this test, the numerator is a sample mean subtracted from a known population mean. If the standard deviation of the population is known as well, then the standard error can be determined, and a single-sample z test can be run. The inferential decision is based on the values in the z table (Table A.1) with the typical critical values being ±1.96. If the population standard deviation is unknown, then the sample standard deviation is used to estimate the standard error. This technique is called the single-sample t test, and the resulting t value is compared against critical values found in the t table (Table A.2). For situations when a sample mean is compared with a known population mean, the key diagnostic question concerns whether or not σ is known.

The independent-samples t test found in Chapter 9 is employed when comparing two sample means coming from two independent samples. (Independent samples occur when two separate and unmatched groups of participants are used to create the two sample means.) The error term in the denominator uses the standard deviation from both samples to help estimate the standard error. If the two means to be compared are either matched or come from the same set of participants, the test to use is Chapter 10's dependent-samples t test. Exercises requiring this test usually convey the dependent relationship between the samples in the wording of the research scenario. However, the manner in which data are presented may also suggest each participant is generating two scores. This test, relative to the independent-samples t test, increases the statistical power (i.e. ability to reject correctly false null hypotheses) by reducing the amount of error in the estimate of the standard error, thus generating larger t values

(in terms of absolute value). Therefore, for situations when two samples are being compared, the key diagnostic question concerns whether the samples come from independent or dependent groups.

Since real-world research problems do not come with a label informing the researcher of which test to use for analysis, it is important for us to work on our diagnostic skills. Understandably, the exercises at the end of each particular chapter require the use of the test(s) found within that chapter for solution. These end-of-chapter work exercises are designed to get us familiar with solving a known type of problem; however, they are not designed to challenge our diagnostic skills (i.e. knowing which test to use for a given situation). The following review section, however, is designed to help us with this skill.

The exercises below will help us review the statistical differences between the various tests. The hypothesis testing exercises (numbers 3–9) will not identify which test is appropriate for the described scenario. We will need to use the available information to make that determination. (Note: Most of the exercises below can be solved either with or without the use of statistical software.)

Questions and Exercises

1 Which of the previously presented tests can be run prior to the gathering of any data?

2 Identify the critical values for the following situations (use online table if needed).

- **a** Dependent-samples t test, $\alpha = .05$, $df = 19$, two-tailed test.
- **b** Independent-samples t test, $\alpha = .10$, $df = 16$, two-tailed test.
- **c** Single-sample z test, $\alpha = .01$, $n = 40$, one-tailed test.
- **d** Dependent-samples t test, $\alpha = .10$, $df = 7$, two-tailed test.
- **e** Dependent-samples t test, $\alpha = .01$, $df = 100$, two-tailed test.
- **f** Single-sample t test, $\alpha = .01$, $df = 8$, two-tailed test.
- **g** Dependent-samples t test, $\alpha = .05$, $n_p = 5$, two-tailed test.
- **h** Independent-samples t test, $\alpha = .05$, $n_1 = 6$, $n_2 = 4$, one-tailed test.
- **i** Single-sample z test, $\alpha = .05$, $n = 22$, two-tailed test.
- **j** Single-sample t test, $\alpha = .10$, $n = 30$, two-tailed test.
- **k** Independent-samples t test, $\alpha = .01$, $n_1 = 14$, $n_2 = 15$, two-tailed test.
- **l** Dependent-samples t test, $\alpha = .01$, $n_p = 3$, two-tailed test.

3 A school psychologist in a rural area is concerned that the children in the local grade school do not have enough social interaction with peers. Suppose that previous research suggests that grade school children average 2.60 hours per day playing with friends. The psychologist samples 18 children at the school and records the following results.

Number of hours/day spent playing with friends

0.80	1.30	1.90	2.40	2.00	2.70
1.25	0.75	1.60	1.50	0.80	2.20
0.75	2.10	0.95	0.60	2.80	0.90

a State the null and alternative hypotheses.
b What is the appropriate inferential test? Why?
c What is the observed statistic?
d Identify the critical values for $\alpha = .05$, two-tailed test.
e Reject the null hypothesis?
f If so, what is the effect size?
g If the null is not rejected, what was the statistical power of this test if the size of the treatment effect is .5 ($\alpha = .05$; two-tailed test)?
h What type of decision error might have been made?
i Properly present the findings.

4 A sleep researcher believes that people will experience a different number of dreams depending on the temperature of the room in which they are sleeping. Adult volunteers are asked to sleep for 10 nights in an 80 °F room and for 10 nights in a 65 °F room. The temperature is alternated randomly to prevent habituation. The total number of dreams reported by each participant is given below.

Participant	80 °F room	65 °F room
P_1	5	8
P_2	7	7
P_3	15	20
P_4	12	17
P_5	10	11

a State the null and alternative hypotheses.
b What is the appropriate inferential test? Why?
c What is the observed statistic?
d Identify the critical values for $\alpha = .05$, two-tailed test.
e Reject the null hypothesis?
f If so, what is the effect size?
g If the null is not rejected, what was the statistical power of this test if the size of the treatment effect is .5 ($\alpha = .05$; two-tailed test)?
h What type of decision error might have been made?
i Properly present the findings.

5 A psychology professor believes the current class of statistics students is more intelligent than most of the previous classes. To test this hypothesis, the psychologist has the WAIS-R, an intelligence test, administered to a random sample of 12 students. Previous offerings of the WAIS-R to statistics students have generated a population mean of 110 and a standard deviation of 15. The psychologist obtains the following scores from the students.

WAIS-R scores

110	113
112	117
109	119
118	121
116	104
111	130

a State the null and alternative hypotheses.
b What is the appropriate inferential test? Why?
c What is the observed statistic?
d Identify the critical values for $\alpha = .05$, two-tailed test.
e Reject the null hypothesis?
f If so, what is the effect size?
g If the null is not rejected, what was the statistical power of this test if the size of the treatment effect is .5 ($\alpha = .05$; two-tailed test)?
h What type of decision error might have been made?
i Properly present the findings.

6 Oishi and Schimmack (2010) found that people who move frequently as children tend to have lower average levels of subjective well-being (happiness) as adults. To further explore this idea, suppose a psychologist samples 15 people who experienced 4 or more different homes before they were 12 years old. These participants were given a standard well-being questionnaire with a known $\mu = 50$. (Higher scores register greater subjective well-being.) The data from the 15 participants follows.

Subjective well-being scores

40	51	53
47	52	44
43	50	38
43	48	44
46	45	46

a State the null and alternative hypotheses.
b What is the appropriate inferential test? Why?
c What is the observed statistic?
d Identify the critical values for $\alpha = .05$, two-tailed test.
e Reject the null hypothesis?
f If so, what is the effect size?
g If the null is not rejected, what was the statistical power of this test if the size of the treatment effect is .5 ($\alpha = .05$; two-tailed test)?
h What type of decision error might have been made?
i Properly present the findings.

7 A physician is interested in comparing the relative effects of a synthetic anabolic steroid with a recently manufactured natural growth stimulant on weight gain. Sixteen patients in a nursing home are randomly assigned to two treatment conditions. One group (eight patients) receives the steroid for 30 days, and a second group (eight patients) receives the growth stimulant for 30 days. The dependent variable is the amount of weight gained at the end of the 30 days. The data follow.

Weight gained (lb)	
Steroid	**Growth stimulant**
6	2
5	5
7	0
2	1
6	2
5	3
4	4
8	7

a State the null and alternative hypotheses.
b What is the appropriate inferential test? Why?
c What is the observed statistic?
d Identify the critical values for $\alpha = .05$, two-tailed test.
e Reject the null hypothesis?
f If so, what is the effect size?
g If the null is not rejected, what was the statistical power of this test if the size of the treatment effect is .5? ($\alpha = .05$; two-tailed test)?
h What type of decision error might have been made?
i Properly present the findings.

8 Another researcher asks the same question posed in Exercise #7. However, in this research setting four participants are given the steroid for 30 days, followed by a 30-day period with the growth stimulant. A different set of four patients receive the two compounds in reverse order. Use the same data found in Exercise #7.

- **a** State the null and alternative hypotheses.
- **b** What is the appropriate inferential test? Why?
- **c** What is the observed statistic?
- **d** Identify the critical values for $\alpha = .05$, two-tailed test.
- **e** Reject the null hypothesis?
- **f** If so, what is the effect size?
- **g** If the null is not rejected, what was the statistical power of this test if the size of the treatment effect is .5 ($\alpha = .05$; two-tailed test)?
- **h** What type of decision error might have been made?
- **i** Properly present the findings.

9 A men's collegiate soccer coach wants to see if student/athletes in this program are more fit than collegiate student/athletes in general. Suppose a nationally normed collegiate-athlete fitness test exists which reports that the mean mile time for the population of biological male collegiate athletes is 5 minutes with a standard deviation of 30 seconds. The coach randomly samples a roster and gathers the following data. (Times have been adjusted to fractions of a minute to avoid the problem of converting 60 seconds into a minute.)

Mile times

5.3	4.7
4.4	4.9
4.8	4.6
5.1	4.6

- **a** State the null and alternative hypotheses.
- **b** What is the appropriate inferential test? Why?
- **c** What is the observed statistic?
- **d** Identify the critical values for $\alpha = .05$, two-tailed test.
- **e** Reject the null hypothesis?
- **f** If so, what is the effect size?
- **g** If the null is not rejected, what was the statistical power of this test if the size of the treatment effect is .5 ($\alpha = .05$; two-tailed test)?
- **h** What type of decision error might have been made?
- **i** Properly present the findings.

10 A researcher is interested in the effect of emotion on concentration. A two-sample study is designed in which anger is induced in one sample by having

a confederate provoke an argument in the lab waiting room. The control group does not undergo this mood induction. Both samples are then tested on a computer stunt driving game, and the number of times the participant runs the vehicle into an object (crashes) is counted. The data follows:

Angry group	Control group
6	6
9	5
13	8
11	6
5	9
10	7

a State the null and alternative hypotheses.
b What is the appropriate inferential test? Why?
c What is the observed statistic?
d Identify the critical values for $\alpha = .05$, two-tailed test.
e Reject the null hypothesis?
f If so, what is the effect size?
g If the null is not rejected, what was the statistical power of this test if the size of the treatment effect is .5? ($\alpha = .05$; two-tailed test)?
h What type of decision error might have been made?
i Properly present the findings.

11 A sports psychologist would like to compare the effects of different exercise programs on cardiovascular fitness. The measure of fitness is the resting heart rate of participants after they complete the program, with lower heart rates indicating greater physical fitness. Twelve college students are randomly assigned to three groups (four participants per group). Participants in the Aerobic condition walk a treadmill for 30 minutes, three times a week. Participants in the Circuit condition perform exercises on weight machines for 30 minutes, with a 10-second rest between exercises. In the Control condition, participants are asked to simply maintain their usual amount of exercise. The resting heart rates of all participants are taken after 10 weeks. Data from this hypothetical study are presented in the following table.

Aerobic	Circuit	Control
65	74	74
62	65	78
56	62	86
60	72	75

a What is the appropriate inferential test? Why?
b What is the observed statistic?
c Identify the critical values for $\alpha = .05$, two-tailed test.
d Reject the null hypothesis?
e If so, what is the effect size?
f If the null is not rejected, what was the statistical power of this test if the size of the treatment effect is .5? ($\alpha = .05$; two-tailed test)?
g What type of decision error might have been made?
h Properly present the findings.
i State the null and alternative hypotheses.

Part 5

Inferential Statistics

Analyses of Variance

12

One-Way Analysis of Variance

12.1 The Research Context

An independent-samples *t* test can be used to test a null hypothesis of no difference between two means. The *t* test is ideal for studies that require a comparison between two groups. However, it is frequently the case that a researcher will use more than two groups in a study. Comparing three education programs, contrasting the effectiveness of two psychotherapy groups and a control group, and evaluating the effects of three kinds of persuasive messages on attitude change are all examples that require an analysis of more than two group means. Indeed, any study that has more than two groups lends itself to the use of an **analysis of variance (ANOVA)**. It is true that we could avoid using an ANOVA and simply conduct *t* tests on all possible two-group (also called *pairwise*) comparisons. For example, in a study with three groups, we could calculate *t* values for group 1 versus group 2, group 1 versus group 3, and group 2 versus group 3. However, conducting multiple *t* tests raises a serious statistical problem.

Multiple *t* Tests and the Type I Error

Recall that Type I errors are committed when a true null hypothesis is mistakenly rejected. The probability of making a Type I error is determined directly when setting the alpha level. This means that when the alpha level is set at .05, and a *t* test is conducted, the probability of mistakenly rejecting a true null hypothesis is .05, *for that one t test*. Over a *series* of *t* tests, however, the probability of making a Type I error becomes *inflated*. For example, if we performed three *t* tests, the probability of making at least one Type I error is

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

closer to .14. If 10 t tests were conducted, the probability of making at least 1 Type I error would be an astonishing .40. Over a series of t tests, the Type I error rate inflates by $1 - (1 - \alpha)^c$, where c is the number of independent comparisons. (To help envision this problem, imagine a 20-sided die, each side representing the statistical outcome when a null hypothesis is being tested. Nineteen of the sides would be white, representing a proper decision to fail to reject the null hypothesis. However, one of the sides, representing 5%, is colored in red and marked "Type I error." With each t test run where the null is true, the die must be rolled. The cumulative probability of the die, at least once, landing on the "Type I error" side increases in an almost additive fashion as the number of rolls accumulate.)

This alpha inflation is eliminated by using an ANOVA since only one test is performed: the F test. The F test, named after its originator, Sir Ronald Fisher (see Spotlight 12.1), provides a comparison of all the population means in one test, just one roll of the 20-sided die. A sufficiently large F value means there is statistical evidence suggesting at least two of the sample means come from different populations. The problem is that a sufficiently large overall F test does not tell us which pairs of means are statistically different from one another. To make *pairwise* comparisons among all the means, special follow-up (or secondary) analyses are used that control the Type I error rate. A couple different versions of these analyses are presented near the end of this chapter.

A **one-way ANOVA** is used on designs having one independent variable (or factor) of three or more levels, each level having its own group of participants. (Even though ANOVAs are used to analyze data from nonexperimental designs, to simplify the language in the chapter, and in keeping with the terminology used in previous chapters, we will use experimental language when describing research designs.) Interestingly, an ANOVA can be used even if the design has only two levels. For these designs, ANOVAs are actually equivalent to t tests. Commonly, however, t tests are used for two-condition studies. If our design has two *factors*, a two-way ANOVA is used. (The two-way ANOVA is discussed in Chapter 13.) If our design uses the same participants for each level of the factor, a repeated-measures ANOVA is used. (The repeated-measures ANOVA is discussed in Chapter 14.)

The test's name, analysis of variance, may seem to suggest that it is a test of variances. In actuality, means are still being compared, but the comparison is based on the sources of variation within the data, including the variation between group means. Just as with the t tests, the ANOVA will help us decide if we can conclude that the sample group means are coming from different populations. How this is accomplished will become clear as the chapter progresses.

Spotlight 12.1 Sir Ronald Fisher

Ronald Fisher (1890–1962) was born in England. He is considered a child prodigy. His daughter and biographer offers the following story.

> At about age three when he had been set up in his high chair for breakfast, he asked: "What is a half of a half?" His nurse answered that it was a quarter. After a pause, he asked, "And what's a half of a quarter?" She told him that it was an eighth. There was a longer pause before he asked again, "What's a half of an eighth, Nurse?" When she had given her reply there was a long silence. Finally, Ronnie looked up, a plump pink and white baby face framed with waving red-gold hair, and said slowly, "Then, I suppose that a half of a sixteenth must be a thirty-toof." (Box, 1978, pp. 12–13)

Fisher's early mathematical ability flourished and led to the development of the most popular inferential test in experimentation: the analysis of variance (ANOVA).

Fisher received his training in mathematics at Cambridge and subsequently taught math in public schools. His daughter points out that he was a lousy teacher and did not like the profession of an educator. In 1917, he married Ruth Guinness, the cousin of the well-known Irish brewery operators. (We may recall that Gossett developed the *t* test while employed by Guinness.) When he quit teaching, Fisher was offered two jobs. One offer came from Karl Pearson, a person Fisher disliked because he, as the editor of *Biometrica*, kept rejecting Fisher's articles. The other offer, the one he accepted, was from the Rothamsted Experimental Station, the oldest agricultural research station in the world. It was during his tenure here that Fisher made many of his most brilliant statistical and research design contributions.

One of the problems in agricultural experimentation at that time was how to determine the effects of a multitude of variables on plant yields; factors for consideration were soil, fertilizer, weeds, seeds, and weather. Fisher devised several experimental designs suited to answer these questions. The nature of these experimental questions led Fisher to develop the factorial design and the ANOVA. To appreciate the context within which Fisher was working, it is interesting to note that the article that presented the ANOVA is titled "Studies in Crop Variation. II. The Manurial Response of Different Potato Varieties" (Fisher & MacKenzie, 1923).

Fisher's contributions were broad. He is responsible for coining many of the common terms in statistics including *variance*, *randomization*, and even the term *statistic* itself. Fisher also worked out many of the sampling distributions required for hypothesis testing and was an influential figure in establishing the 5% alpha level for rejecting null hypotheses.

Unfortunately, like many of the other statisticians of that time who were carving out the mathematical tools for the new social and behavioral sciences (e.g. Galton, Pearson, Spearman), Fisher was a strident eugenicist, holding, among other distasteful beliefs, the strong opinion that there were intellectual differences between various races of humans. This conviction even resulted in him registering a dissenting opinion to the United Nations Educational, Scientific and Cultural Organization (UNESCO) project on racism (UNESCO, 1952).

Although the personal beliefs of so many of the pioneers of inferential statistics are offensive, the power of the statistical tools they created should not be diminished by these associations. These tools, though originally created from motives to show differences between races and classes of people, are thankfully not limited to use for those sorts of questions.

12.2 The Conceptual Basis of ANOVA: Sources of Variation

Between-Group Variation

If we conduct a study with two or more groups, the resulting group means will almost never be identical. This would be true even if the null hypothesis were true. Our task is to decide if the independent variable has influenced the group means. Our decision will be based on the outcome of an inferential test. When the group means are similar, there is little variation between the means. When the means are very different, there is a greater degree of variation between the means. The difference between the group means is called **between-group variation.** When considering what can account for this between-group variation, three possible explanations emerge:

1) **Treatment variance** (or **primary variance**). This is the effect due to treatment, that is, the degree to which the mean differences are caused by the influence of an independent variable. This source of variation is what researchers attempt to maximize and then detect with statistical analysis. (In a nonexperimental situation, the more general term, "primary variance," may be more readily used.)
2) **Individual differences**. Each research participant comes to the experiment with a unique constellation of personality traits, called *participant variables* (see Chapter 1). Since each participant is unique, each one will inevitably respond differently to the task used to assess the influence of the independent variable. The random assignment of participants usually does a good job of balancing these individual differences across the conditions. However,

even when participants are randomly assigned to treatment conditions, it is possible for group means to differ due to differences between the individuals that comprise the groups. For example, consider a study comparing three teaching techniques. If, *by chance*, a few more of the sharper students are randomly assigned to one treatment condition than the other two, that particular group mean may reflect this lack of homogeneity. The larger the group of participants in the study, the less likely variation due to individual differences will become a problem. (See Chapter 6 to review the positive probabilistic benefits that come with large sample sizes.) In nonexperimental designs, individual differences are assumed to be largely balanced across groups by the random sampling of the respective populations.

3) **Experimental error**. This variance reflects the difference between a measurement and the true value. There are three potential sources of experimental error: the unreliability of measuring instrumentation (e.g. intelligence tests do not generate the same value each time they are administered to a given person), inconsistent interactions between the experimenter and the participants (e.g. the experimenter may state the instructions differently or with different enunciation or body language to different participants), and random forms of environmental disturbances during the experiment (e.g. the turning off and on of an air-conditioning system). In short, any uncontrolled aspect of the experiment could account for the variation among group means. These uncontrolled aspects of the experiment are assumed to be somewhat equally spread across the breadth of the research design and do not intrude *systematically*. (Do not confuse experimental error with confounding error. Confounds *systematically* or *somewhat systematically* vary among the groups and provide unintended, yet plausible explanations for differences among group means. Because experimental error influences data in an *unsystematic* manner, confounding variance is not introduced.)

Individual differences and experimental error are considered **random factors.** They are called random because they are not intentionally or systematically manipulated by the experimenter, yet they may intrude and produce differences between means. In summary, any one, or combination, of the foregoing reasons can influence the variation among group means. A second way to examine the variance in an experiment is to look at it from the within-group perspective.

Within-Group Variation

Between-group variation considers the variation between means. **Within-group variation** refers to the variation among scores *within* a group. Irrespective of whether a group of scores is a treatment or control group, would we

expect every person within a given group to produce the exact same score? Of course not, for the following two reasons:

1) **Individual differences.** Some variation among means could be due to individual differences. However, when considering the variation of *scores within a group*, individual differences always produce variability. Once again, let us use the teaching technique example. With respect to *between-group variation*, an individual difference variable (e.g. achievement motivation) could be unequally distributed *across* groups and account for differences among means. With respect to *within-group variation*, the focus shifts to each particular group. Students *within* a given group will vary in achievement motivation. Some measure of the effectiveness of the teaching technique will be administered. Whether or not a treatment effect exists, there will be *within-group variation* among the scores of the dependent variable because of the differences among the students in terms of their achievement motivation. Each group in the study will have within-group variability. For the study as a whole, the amount of within-group variability is based on a combination of the within-group variation from each group.
2) **Experimental error.** The variation among scores within a group could also be due to experimental error. These errors are the same as those that occur in between-group variation. Since individual differences and experimental error are random factors, the only sources of within-group variation are due to random factors.

It is important to realize that the treatment administered to each group *does not* contribute to within-group variability since each participant in a group is exposed to the same treatment. The ANOVA is a statistical technique that analyzes the sources of variability within the experiment. The sources of variation are depicted in the diagram of Figure 12.1.

Sources of Variation When the Null Hypothesis Can Be and Cannot Be Rejected

The amount of variation due to treatment is called treatment variance or primary variance. The variation due to random factors (individual differences and experimental error) is called **error variance** (or **secondary variance**). (These terms will be used somewhat interchangeably.)

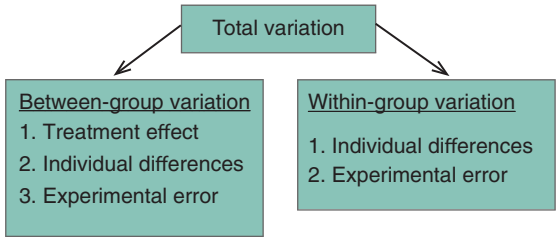


Figure 12.1 The ANOVA partitions the sources of variation in an experiment into between-group variation and within-group variation.

Suppose the H_0 is true (i.e. there is *no* primary variance, only secondary variance). In this instance, any difference between the means is completely due to the random variation of scores within the groups. Since group means will never be identical (even if the H_0 is correct) the statistical question becomes, "How different do the means have to be to conclude there is evidence of treatment variance?" To reject the null hypothesis, the between-group variation (treatment variance + error variance) has to be sufficiently greater than the overall within-group variation (error variance). The *F* test is the ratio between these two measures of variation:

$$F = \frac{treatment\ variance + error\ variance}{error\ variance}$$

Stated differently,

$$F = \frac{between\text{-}group\ variance}{within\text{-}group\ variance}$$

In the absence of a treatment effect, the between-group variance will be nothing but error variance. This means that when the H_0 is correct, the *F* ratio will be close to 1 (one measure of error variance divided by another measure of error variance). As the influence of treatment variance becomes stronger, the *F* ratio becomes larger. The larger the value of *F*, the more likely the null hypothesis should be rejected.

When the null hypothesis is correct, any differences between groups are due entirely to error variance. When the null hypothesis is false, at least some of the difference between means is due to the treatment effect. As the amount of between-group variation due to treatment increases, the numerator of the *F* ratio increases. If the effect of treatment is sufficiently great, then the *F* ratio will prompt the rejection of the null hypothesis. Figure 12.2 summarizes the sources of variance when the null hypothesis is correct and incorrect.

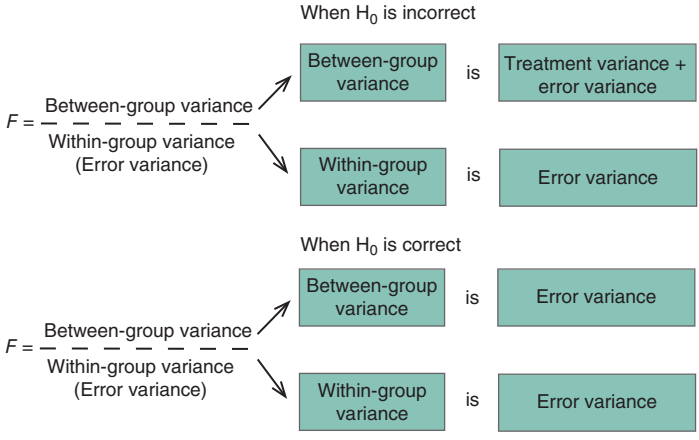


Figure 12.2 Sources of variance when the null hypothesis is correct and incorrect.

12.3 The Assumptions of the One-Way ANOVA

Following is a list of assumptions underlying the ANOVA. Note that the assumptions are the same as those for the independent-samples *t* test:

1) **Representativeness**. It is assumed that each sample is representative of the population from which it has been drawn. Random sampling is the best data gathering method to meet this assumption; however, other sampling methods might be sufficient. Meeting this assumption allows us to generalize from samples to populations.
2) **Independent observations**. Independent observations mean that the scores *within each sample* are independent of one another. If the behavior of one participant in the study is influenced by the behavior of another participant, then the scores from these two participants are *not* independent.
3) **Interval or ratio scale of measurement**. The one-way ANOVA utilizes means and standard deviations. These concepts only have meaning for data measured on a scale where the quantitative distance between integers is held constant, namely, an interval or ratio scale (see Chapter 2).
4) **The populations from which the samples are taken are normally distributed**. This assumption states that each sample is drawn from a population that is normally distributed.
5) **Homogeneity of variances**. The variances of each population distribution are the same; thus $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_k^2$, where k is the last group.

Since the *F* test is *robust*, it is not essential that the last two assumptions be met, particularly if each sample contains a sufficiently large and equal number of observations. However, gross violations of these assumptions will adversely affect the validity of the *F* test. Strong violations would require the use of analysis tools that do not make assumptions about the shape of the population distributions (see Chapter 18).

12.4 Hypotheses and Error Terms for the One-Way ANOVA

The Null and Alternative Hypotheses

The ANOVA provides a direct test of the null hypothesis, which is always

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_k.$$

A substantial difference between any two means will produce a large *F* ratio. The alternative hypothesis is always

H_1: at least two of the means are different.

Note that H_1 is not $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_k$. This erroneous statement of the alternative hypothesis would indicate that all the means must be substantially different from one another. The null hypothesis can be rejected when just two of the population means are different.

Mean Square Within: One Estimate of σ^2

Recall the two statistical assumptions underlying the ANOVA: The populations from which the samples are drawn are normal and the populations have the same variances. Since we have already learned the logic of hypothesis testing in previous chapters, we know that when testing a null hypothesis we set up the test *as if* the null hypothesis is true. Although the mathematics take place at the level of samples, the critical question concerns whether or not the samples come from the same population. (An equivalent manner of expression asks whether the samples come from identical populations. These are, in effect, the same question.) If we find evidence of at least one difference among the sample means, then we conclude that the populations from which at least two of the samples were drawn are different. If there is no evidence of differences between means, then we do not reject the idea that the means are from identical populations. (Notice this does not mean we conclude that the population means are equal; to do that would be to accept the null hypothesis.)

No matter what *t* test we ran in the previous chapters, we *always* generated an estimate of the population variance. To decide if a difference between sample means is evidence that they come from different populations, we need an estimate of how variable the null distribution is from which we are sampling. The ANOVA is no different in this regard; we need an estimate of σ^2.

In a study with three groups, there are three estimates of σ^2: s_1^2, s_2^2, and s_3^2. Which one should be used? Well, rather than rely on any one of them, we will use them all. The best estimate of σ^2 is made by pooling the variances of the samples. The term that is used for the pooled variance is the **mean square within (MS_W)**. We have encountered this concept before. We may recall from our discussion of the independent-samples *t* test that the denominator of the *t* test is the pooled variance, which is a weighted average of the variances of two samples. The MS_W term merely expands the pooled variance concept to incorporate more than two groups.

Pooled variance formula for MS_W

$$MS_W = \frac{s_1^2(n_1-1) + s_2^2(n_2-1) + \cdots + s_k^2(n_k-1)}{(n_1-1) + (n_2-1) + \cdots + (n_k-1)} \quad \text{(Formula 12.1)}$$

Since MS_W is a combination of sample variances, the degrees of freedom associated with MS_W is also the addition of each of their respective degrees of

freedom. More simply, this can be represented as the sample size for the study (N) minus the number of groups in the study (k):

$$df_W = (n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1) = N - k.$$

The MS_W is a good estimate of σ^2 (the error variance). Most importantly, it is a good estimate of σ^2 whether or not the H_0 is correct. The MS_W is based on the pooled average of the within-group variances. In the absence of a treatment effect, the sample variances are all taken from the same population. In the presence of a treatment effect, we are sampling from different populations. In either case, MS_W is a good estimate of σ^2 because the ANOVA assumes $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_k^2$ (see Section 12.3).

Mean Square Between: Another Estimate of σ^2

The numerator of the F ratio is the average variation among the group means, the **mean square between** (or mean square between groups), $\boldsymbol{MS_{BG}}$. When the H_0 is true, the only sources of variation among the group means are random factors (error variance). This means that a measure of variance between the group means is another way to measure error variance (σ^2).

We have previously learned of a statistic that serves as a measure of the amount of variation among means. A sampling distribution of means has a standard deviation, the standard error of the mean (σ_M). In fact, the relationship between the population standard deviation and the standard error of the mean is $\sigma_M = \sigma/\sqrt{n}$. Since variances, not standard deviations, are used in an ANOVA, both sides of the equation are squared to obtain

$$\sigma_M^2 = \frac{\sigma^2}{n}.$$

Multiplying each side of the equation by n gives

$$n\sigma_M^2 = \sigma^2.$$

When using sample means to estimate σ^2, simply treat the sample means as raw scores, apply our preferred formula for calculating the *variance* (not the standard deviation), and multiply the result by n (n = the number of participants in any group, assuming equal numbers of participants). Formula 12.2 gives the between-groups *estimate* of σ^2.

Between-group variance (MS_{BG}) as an estimate of σ^2

$ns_M^2 = \sigma^2$ (Formula 12.2)

This is the measure of between-group variance. Table 12.1 presents an example of how to calculate the between-group variance using this method. The means are borrowed from Table 12.2. Note that the three means are treated

Table 12.1 Between-group variation.

$M_1 = 4.75$
$M_2 = 8.75$
$M_3 = 2.25$
$n = 3$ (*number of groups*)
$\Sigma M^2 = 104.19$
$\Sigma M = 15.75$

$$s_M^2 = \frac{\Sigma M^2 - \left[(\Sigma M)^2/n\right]}{n-1}$$
$$s_M^2 = \frac{104.19 - \left[(15.75)^2/3\right]}{3-1}$$
$$s_M^2 = \frac{104.19 - [248.06/3]}{2}$$
$$s_M^2 = \frac{104.19 - 82.69}{2}$$
$$s_M^2 = \frac{21.50}{2}$$
$$s_M^2 = 10.75$$

Between-group variance = ns_M^2
$ns_M^2 = (4)(10.75)$
$ns_M^2 = \mathbf{43}$

The variance of the sample means multiplied by sample size.
The n in the formula for s_M^2 is the number of group means (3); the n in the formula ns_M^2 is the number of participants in one group (4), assuming equal numbers of participants per group.

not only as raw scores but also as a *sample* of scores that *estimate* σ^2. This means that the sample formula for s^2 can be used (using means, M's, in place of raw scores, X's). As we examine the computational flow of the problem, keep in mind that the n in the s_M^2 formula refers to the number of *groups*, which we are treating as if they are individual scores. (Since we are computing the variance of *means*, the symbol s_M^2 is used instead of s^2.) The n in the formula for between-group variance (Formula 12.2) refers to the number of participants in *each* of the samples ($n_1 = n_2 = n_3 = 4$), not the total number of participants across all groups.

Putting It All Together

Recall that MS_W is a good estimate of σ^2 whether H_0 is correct or incorrect. Since the ANOVA assumes $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_k^2$, whether there is one or several populations is not important; the same value is being estimated by the various sample variances.

The situation is different with respect to the variance of the group means (between-group variance, MS_{BG}). This variance, which conceptually parallels the

standard error of the mean, is a good estimate of σ^2 only when it is based on a sampling distribution established by taking repeated samples from *one* population, that is, when H_0 is correct. Here is the beauty of the ANOVA. When the H_0 is correct, ns_M^2 (between-group variance, MS_{BG}) is also a good estimate of σ^2. If ns_M^2 is a good estimate of σ^2, then it should be very close to the value of MS_W. If these two estimates of σ^2 are similar, the resulting F ratio will be close to 1. However, if the sample means come from different populations, ns_M^2 will be a poor estimate of σ^2, although MS_W will remain a good estimate. When ns_M^2 is a poor estimate of σ^2, it will only and always *overestimate* σ^2. The greater the difference among the sample means due to treatment, the larger the overestimate of σ^2; treatment variance *can only increase* the value of ns_M^2. Since ns_M^2 is the measure of between-group variance, it is placed in the numerator of the F ratio. The result of treatment variance is an F ratio that becomes larger as the value of ns_M^2 increases.

A point that has been stressed throughout this text, which cannot be emphasized enough, is that hypothesis testing uses samples to draw conclusions about populations. We fly blind, in a sense, because we *never* know the true nature of the populations. Inferential tests involve logic and mathematics as aids to allow the researcher to infer the characteristics of populations; yet all inferences involve a degree of uncertainty. Figures 12.3 and 12.4 illustrate the inferential situation in which a researcher is involved when using samples to make statements about populations. In Figure 12.3, a study with four groups is illustrated under the condition that the null hypothesis is true. When the null hypothesis is true, the four populations that are sampled are identical; that is, they have the same means and variances. Sampling from four identical populations is like sampling from one population. The four arrows at the bottom of Figure 12.3 reflect the sample means. Note that they are close together; they show little variability. This is just what we would expect when taking random samples from four identical populations (or four samples from the same population). In Figure 12.4, the sample means depicted by the arrows at the bottom of the diagram are more variable. Why? Because the population from which the fourth sample was drawn has a mean that is different from that of the other three populations.

12.5 Computing the *F* Ratio in a One-Way ANOVA

To recap, both MS_{BG} and MS_W are measures of variation. They are appropriately termed *mean* square between and *mean* square within because they reflect the average (mean) amount of variation. MS_{BG} is the average variation of the group means around the grand mean (the **grand mean** being the mean of all the scores in the experiment, irrespective of individual groups). MS_W is the average amount of variation of the individual scores with respect to the group from which the scores are taken.

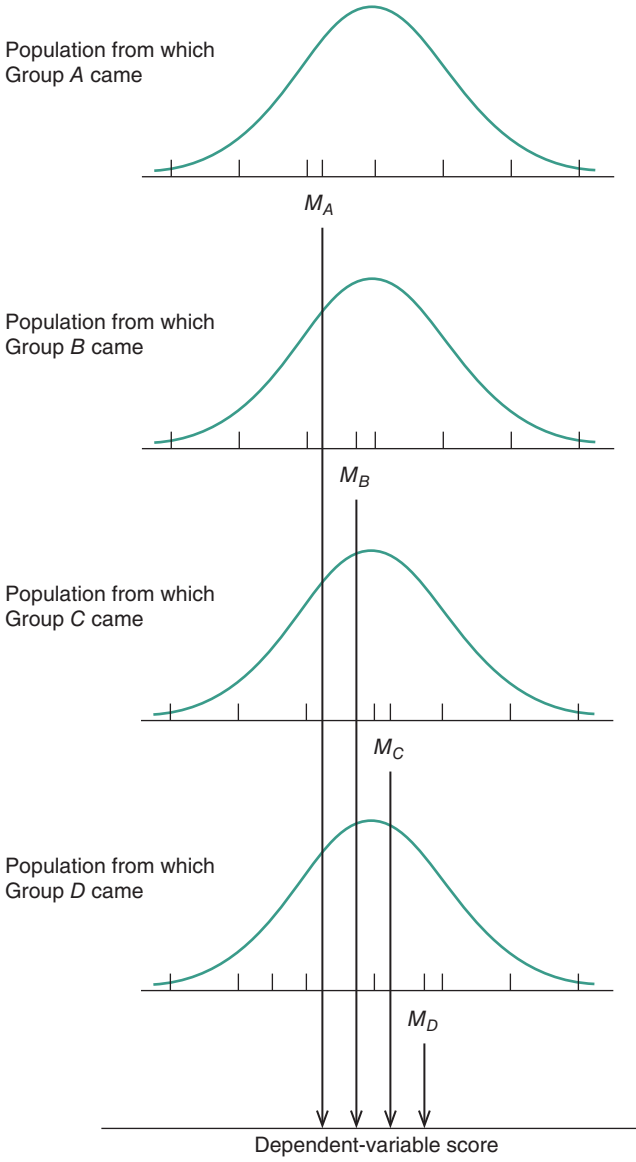


Figure 12.3 When the H_0 is true, the sample means are drawn from identical populations, which, in effect, is the same as saying "drawn from one population." Here we would expect the sample means to show little variation.

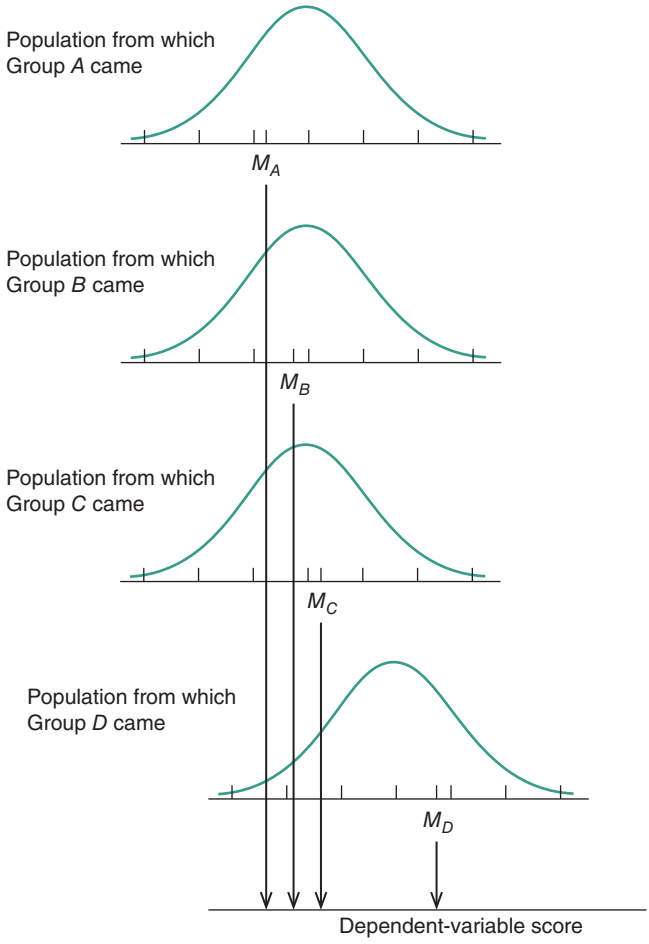


Figure 12.4 When the H_0 is false, at least one sample is drawn from a population that is different from the rest.

MS_{BG} and the Sum of Squares Between Groups (SS_{BG})

The definitional and computational formulas used to compute MS_{BG} and MS_W are presented in this section. Immediately following the presentation of formulas, a hypothetical data set will be used to illustrate the steps for computing MS_{BG}, MS_W, and the F ratio.

The MS_{BG} is a ratio of the sum of squares between the groups (SS_{BG}) divided by the degrees of freedom for SS_{BG}. (Degrees of freedom will be discussed later.) Formula 12.3 is the definitional formula for SS_{BG}.

Definitional formula for SS_{BG}

$SS_{BG} = \Sigma n_k(M_k - M_G)^2$ (Formula 12.3)

where

n_k = the number of participants in group k
M_k = the mean of group k
M_G = grand mean

This formula shows that SS_{BG} is the variation of group means about the grand mean. To use this formula in computing SS_{BG}, we would subtract the grand mean from the mean of group 1, square the value, and multiply by the number of participants in that group. Perform the same set of operations for each of the group means in the experiment. Finally, sum all the values.

Formula 12.4 is the computational formula for calculating SS_{BG}. Use this formula when calculating SS_{BG} by hand.

Computational formula for SS_{BG}

$$SS_{BG} = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right] \quad \text{(Formula 12.4)}$$

where

ΣX_1 = the sum of scores in group 1
ΣX_2 = the sum of scores in the second group and so on
ΣX_k = the sum of scores in the last group; if the experiment has four groups, then ΣX_k is the sum of scores in group 4
ΣX = sum of all the scores in the study
n_1 = the number of participants in group 1
n_k = the number of participants in the last group
N = total number of participants in the study

MS_W and the Sum of Squares Within Groups (SS_W)

The MS_W is a ratio of the sum of squares within groups (SS_W) divided by the degrees of freedom associated with SS_W. The definitional formula for SS_W (Formula 12.5) reminds us that within-group error variance is the amount of deviation among individual scores around the mean of the group from which the scores are taken.

Definitional formula for SS_W

$SS_W = \Sigma(X_1 - M_1)^2 + \Sigma(X_2 - M_2)^2 + \cdots + \Sigma(X_k - M_k)^2$ (Formula 12.5)

where

M_1 = the mean of group 1, M_2 is the mean of group 2, and so on
M_k = the mean of the last group
X_1 = each score in group 1, X_2 = each score in group 2, and so on
X_k = each score in the last group

If using the definitional formula for computing SS_W, subtract the mean of group 1 from the first score in that group and square the value. Repeat the operation for each raw score in group 1. Next, sum all these squared deviation scores. Do the same for each group in the study, remembering to use the relevant group mean. The sums of the squared deviations for each group are then summed.

The computational formula is given in Formula 12.6. Use this formula when calculating SS_W by hand.

Computational formula for SS_W

$$SS_W = \Sigma X^2 - \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k}\right] \quad \text{(Formula 12.6)}$$

where

ΣX^2 = sum of *all* squared scores
$(\Sigma X_1)^2$ = the sum of the scores in group 1, quantity squared
$(\Sigma X_2)^2$ = the sum of the scores in group 2, quantity squared
$(\Sigma X_k)^2$ = the sum of the scores in the last group, quantity squared
n_1 = number of participants in group 1
n_k = number of participants in the last group

The Total Sum of Squares (SS_T)

The total sum of squares (SS_T) is not used when computing the F ratio. Nonetheless, we need to calculate it for use in secondary analyses. The total sum of squares equals the sum of squares between groups plus the sum of the squares within groups:

$$SS_T = SS_{BG} + SS_W.$$

Independently computing SS_T also allows us to make sure that our SS_{BG} and SS_W calculations are accurate. The definitional formula for SS_T reveals the fact that the total variation in scores is the difference between each score in the study and the grand mean, squared and summed.

Definitional formula for SS_T

$$SS_T = \Sigma(X - M_G)^2 \quad \text{(Formula 12.7)}$$

where

X = each score in the study
M_G = grand mean

When using the definitional formula, simply subtract the grand mean from each score, squaring each difference as we go. Finally, sum all the squared values.

Formula 12.8 is the computational formula for SS_T. Use this formula when computing SS_T by hand.

Computational formula for SS_T

$$SS_T = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \quad \text{(Formula 12.8)}$$

where

ΣX^2 = the sum of *all* squared scores
$(\Sigma X)^2$ = the sum of *all* scores, quantity squared
N = the total number of participants

Degrees of Freedom

To arrive at the F ratio, SS_{BG} and SS_W must be divided by their appropriate degrees of freedom. This step generates MS_{BG} and MS_W. Dividing by df turns a sum of squares into an average sum of squares, which is what "mean squares" means. The degrees of freedom for the between-groups term (df_{BG}) is

$$df_{BG} = k - 1$$

where

k = the number of groups.

The degrees of freedom used for the within-groups term, df_W, is

$$df_W = (n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1) = N - k.$$

The degrees of freedom for SS_T is $N - 1$. Although SS_T is not divided by its degrees of freedom (df_T), calculating df_T can serve as a computational check since

$$df_T = df_{BG} + df_W.$$

Computing MS_{BG} and MS_W is accomplished in the following manner.

Calculating MS_{BG} and MS_W

$$MS_{BG} = \frac{SS_{BG}}{df_{BG}}$$

$$MW_W = \frac{SS_W}{df_W}$$

and, of course,

$$F = \frac{MS_{BG}}{MS_W}$$

Illustrating the Computational Steps with Raw Data

Table 12.2 presents the raw scores and summary statistics for a hypothetical study with three groups. These data will be used to illustrate the computational steps required to arrive at the *F* ratio. The by-hand calculations can be tedious, but we have had extensive practice in performing the arithmetic operations needed to compute *F*.

Step 1. Compute SS_{BG} using Formula 12.4.

Computational formula for SS_{BG}

$$SS_{BG} = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right]$$

$$SS_{BG} = \frac{(19)^2}{4} + \frac{(35)^2}{4} + \frac{(9)^2}{4} - \left[\frac{(19+35+9)^2}{12}\right]$$

$$SS_{BG} = 416.75 - 330.75$$

$$SS_{BG} = 86.$$

Table 12.2 Hypothetical raw data and summary statistics for a study with three groups.

Group 1	Group 2	Group 3
S_1 4	S_5 8	S_9 3
S_2 5	S_6 8	S_{10} 2
S_3 4	S_7 9	S_{11} 1
S_4 6	S_8 10	S_{12} 3
$M_1 = 4.75$	$M_2 = 8.75$	$M_3 = 2.25$
$\Sigma X_1 = 19$	$\Sigma X_2 = 35$	$\Sigma X_3 = 9$
$\Sigma X_1^2 = 93$	$\Sigma X_2^2 = 309$	$\Sigma X_3^2 = 23$
$n_1 = 4$	$n_2 = 4$	$n_3 = 4$

Step 2. Compute df_{BG}: $df_{BG} = k - 1 = 3 - 1 = 2$.
Step 3. Calculate MS_{BG}.[1]

$$MS_{BG} = \frac{SS_{BG}}{df_{BG}}$$

$$MS_{BG} = \frac{86}{2}$$

$$MS_{BG} = 43.$$

Step 4. Compute SS_W using Formula 12.6.
Computational formula for SS_W

$$SS_W = \Sigma X^2 - \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k}\right]$$

$$SS_W = (4)^2 + (5)^2 + (4)^2 + \cdots + (2)^2 + (1)^2 + (3)^2 - \left[\frac{19^2}{4} + \frac{35^2}{4} + \frac{9^2}{4}\right]$$

$$SS_W = 425 - 416.75$$

$$SS_W = 8.25.$$

Step 5. Compute df_W: $df_W = N - k = 12 - 3 = 9$.
Step 6. Compute MS_W.

$$MS_W = \frac{SS_W}{df_W} = \frac{8.25}{9}$$

$$MS_W = 0.92$$

Step 7. (Optional) Compute SS_T using Formula 12.8.
Computational formula for SS_T

$$SS_T = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$$SS_T = 4^2 + 5^2 + 4^2 + \cdots + 2^2 + 1^2 + 3^2 - \left[\frac{63^2}{12}\right]$$

$$SS_T = 425 - 330.75$$

$$SS_T = 94.25.$$

If the calculations are correct, then SS_T should equal $SS_{BG} + SS_W$:

$$SS_T = SS_{BG} + SS_W$$

$$94.25 = 86 + 8.25.$$

1 Refer to Table 12.1. Note that the value computed for MS_{BG} is the same as the value arrived at by using ns_M^2.

Even though SS_T is not used to calculate F, we are strongly urged to use SS_T as a computational check. The vast number of calculations required almost guarantees that we will make at least one undetected math error. Additionally, we may need this value for secondary analysis purposes.

Step 8. Compute the F ratio.

$$F = \frac{MS_{BG}}{MS_W}$$

$$F = \frac{43}{0.92}$$

$$F = \mathbf{46.74}.$$

12.6 Testing Null Hypotheses

The *F* Distributions

The shape of each distribution in a family of sampling distributions is affected by the sample size used in the repeated sampling process to construct all sampling distributions. Recall that when using t_{obt} to test a null hypothesis, we found the critical values by using the appropriate degrees of freedom. In the same manner, when using an F statistic to test a null hypothesis, we will need to use the appropriate degrees of freedom to find the critical value for the relevant sampling distribution – the ***F* distribution**. Here, two numbers are needed: one for the numerator of the F ratio and another for the denominator.

As is the case with all sampling distributions, an F distribution is theoretical; we do not need to go through the arduous task of creating it. Since the F statistic is a ratio of variances, it should make sense to us that the sampling distributions of F are based on ratios of variances.

Recall that the F ratio is

$$F = \frac{MS_{BG}}{MS_W} = \frac{treatment\ variance + error\ variance}{error\ variance}$$

When the null hypothesis is true, there is no treatment effect; that is, there is no treatment variance in the numerator of the F ratio. Therefore, when H_0 is true,

$$F = \frac{MS_{BG}}{MS_W} = \frac{error\ variance}{error\ variance}$$

The sampling distributions of F ratios are established under the assumption that the null hypothesis is true, $\mu_1 = \mu_2 = \mu_3 = \mu_k$. Here is how we would go about constructing a sampling distribution of F ratios. Imagine that we conduct an experiment with four groups, six participants in each group ($df_{BG} = 3$, $df_W = 20$), and the null hypothesis is true. MS_{BG} is obtained by computing the variance

of the four group means. MS_W is obtained by computing the pooled within-group variance. The F ratio is then calculated. Since the null hypothesis is true (i.e. no treatment variance is present), MS_{BG} and MS_W both reflect nothing but error variance. Since both values are estimates of the same thing, they should be rather similar, and the F ratio will be close to 1. Constructing a sampling distribution proceeds by storing that value and then repeating the process. We would conduct the same experiment again by sampling the same number of participants, placing them into the same number of groups, and computing another F ratio. This process repeats itself a near-infinite number of times. The stored F values would then generate an F distribution (sampling distribution) of all the possible F ratios, given $df_{BG} = 3$ and $df_W = 20$, under conditions when the null hypothesis is true.

Given the foregoing method for generating a sampling distribution of F ratios, what must be true about an F distribution?

1) Since variability can never be represented by a negative number, all F values must be positive.
2) The smallest value that F can obtain is 0.
3) Since the H_0 is true, MS_{BG} and MS_W independently estimate the same value; therefore, most F ratios cluster around 1.
4) Even when the H_0 is true, sampling error is still present. Ratios are bound by 0 when the numerator estimate of error is smaller than the denominator estimate of error, but not bound by a number when the numerator estimate of error is larger than the denominator estimate of error.
5) As a result, F sampling distributions are positively skewed.

When using an F ratio to test a null hypothesis, a sampling distribution is available that matches the degrees of freedom appropriate to the number of groups (numerator) and the number of participants (denominator) in the study. The exact shape of an F distribution will depend on the number of groups and sample sizes used to calculate the F ratio. Figure 12.5 shows two F distributions that are based on different degrees of freedom. Note that as the degrees of freedom change, and thus the shape of the F distribution changes, the critical values beyond which lie 5 and 1% of the F ratios shift accordingly.

As we examine Figure 12.5, bear in mind that the F distributions are distributions obtained when the H_0 is true. Note that the frequency of F ratios decreases as the size of the F ratio increases past 1. When deciding whether to reject the H_0, we have only one F ratio available. If the F ratio obtained from our study falls in the right tail of the distribution, say, beyond the point that marks the upper 5% of the distribution, we have a decision to make. Either the null hypothesis is true and sampling error has given rise to an unlikely large F ratio (a Type I error; less than 5% chance this is the case), or the null hypothesis is false and MS_{BG} is being influenced by a treatment effect. Here is where our decision rule

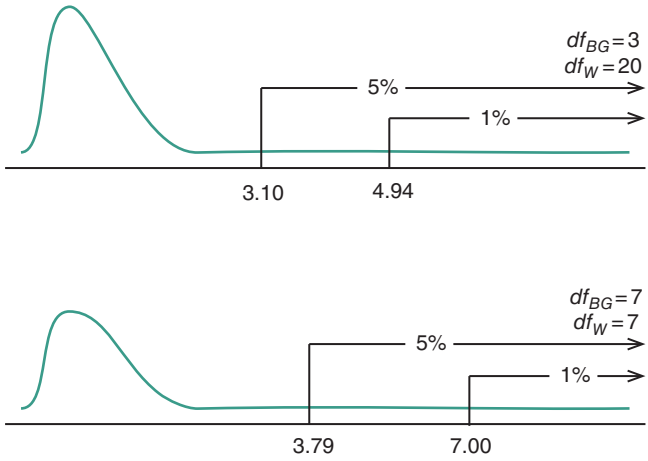


Figure 12.5 Two *F* distribution for $df_{BG} = 3$ and $df_W = 20$, and $df_{BG} = 7$ and $df_W = 7$.

associated with inferential testing gives us direction. In this situation, we are to reject the null hypothesis. We have found statistical evidence that the H_0 is false.

Using the *F* Table

The *F* table (Table A.5) is used to determine if an *F* ratio falls in the far right tail of the sampling distribution. The following is a portion of the table. The *F* table provides critical values for alphas of .05 and .01. The boldface values in the body of the table are used when α = 1%. The lightface values are used when α = 5%.

.05 (roman) and .01 (boldface) α levels for the F distribution

	Degrees of freedom (numerator)			
Degrees of freedom (denominator)	⋯	3	4	5 ⋯
⋮				
19		3.13	2.90	2.74
		5.01	**4.50**	**4.17**
20		3.10	2.87	2.71
		4.94	**4.43**	**4.10**
21		3.07	2.84	2.68
		4.87	**4.37**	**4.04**
⋮		⋮	⋮	⋮

Locating the relevant critical value is accomplished by entering the table with the degrees of freedom from the numerator of the F ratio, df_{BG}, and the degrees of freedom from the denominator of the F ratio, df_W. The degrees of freedom for MS_{BG} is found in the row at the top of the table, and the degrees of freedom for MS_W is found in the left column of the table. When testing the null hypothesis with an F ratio where $df_{BG} = 4$ and $df_W = 20$, the critical values for an alpha of 5 and 1% are 2.87 and 4.43, respectively.

Let us test a null hypothesis with the F ratio found in the worked example using an alpha of .05. The obtained F was 46.74. The $df_{BG} = 2$ $(k - 1)$ and $df_W = 9$ $(N - k)$. Referring to the F table (Table A.5), the critical value is stated as 4.26. Since the obtained F value is larger than the critical value of 4.26, we reject the null hypothesis that $\mu_1 = \mu_2 = \mu_3$.

12.7 The One-Way ANOVA Summary Table

Table 12.3 is one customary way to summarize the results of an ANOVA. The values in the table are taken from the worked example. Some source tables substitute the word "Treatment" for "Between groups," "Error" for "Within groups," and "Sig." for "*p*." The ANOVA summary table supplies the most relevant information used in calculating the F ratio. It will also be a resource for values needed for secondary analyses.

Table 12.3 An ANOVA summary table.

Source of variation	*SS*	*df*	*MS*	*F*	*p*
Between groups	86	2	43	46.74	<.05
Within groups (error)	8.25	9	.92		
Total	94.25	11			

See text for computations.

12.8 An Example of an ANOVA with Unequal Numbers of Participants

When conducting an experiment, it is always desirable to use the same number of participants in each experimental condition. One reason is that it makes any violation of the population assumptions underlying the test less serious. A second reason is that the researcher maximizes power (relative to the number of participants being used) by equally distributing them across conditions. However, sometimes participants drop out of the study, and replacing them

is difficult or impossible. The same formulas that were used in the worked example, in which the number of participants in each group were equal, can be used when there are an unequal number of participants in the groups. This is possible since the formulas used sample sizes as weights when multiplying variances.

Consider the following hypothetical study. A child psychologist is interested in evaluating the effectiveness of two treatments for children who are afraid of the dark. One treatment, "emotive imagery," teaches the children to imagine that they are brave superheroes, like Wonder Woman or Superman, on a mission to save a friend. A second treatment, "relaxation," involves training the children to breathe slowly and deeply when in the dark. A third condition serves as a control condition, and these children are simply asked to remain in the dark as long as they can. The dependent variable is the number of seconds elapsed before the child turns on the light. Table 12.4 presents raw data, summary statistics, the calculation of the F ratio, and the ANOVA summary table.

12.9 Measuring Effect Size for a One-Way ANOVA

The F test provides information about evidence of a difference between at least two means of a study. A sufficiently large F ratio indicates that the observed differences between means are unlikely to occur by chance. Stated differently, all of the sample means are unlikely to have come from the same population. However, the F ratio value concerns the *certainty* of an effect, not necessarily the *size or strength* of the effect. Studies with large amounts of power, for instance, may generate very large F's even when the treatment effect may be quite modest.

As with t tests in previous chapters, we need to use a different statistic to measure the size of the effect. Over the years, statisticians have developed several measures of the *strength of relationship* between the independent variable and the variation of scores on the dependent variable. We will look at two of them. The first statistic offered here is a measure of the estimated magnitude of the treatment effect in the population, **omega-squared (ω^2)**. Omega-squared is easy to calculate, requiring only values from the ANOVA summary table.

Formula for omega-squared, ω^2

$$\varpi^2 = \frac{SS_{BG} - df_{BG}(MS_W)}{SS_T + MS_W} \quad \text{(Formula 12.9)}$$

Table 12.4 An ANOVA with unequal numbers of participants.

Emotive imagery	Relaxation training	Control
40	33	23
45	39	32
45	51	33
53	40	29
49	42	40
	40	25
		28
$M_1 = 46.40$	$M_2 = 40.83$	$M_3 = 30.00$
$\Sigma X_1 = 232$ $\Sigma X_2 = 245$ $\Sigma X_3 = 210$ $\Sigma X_G = 687$ $\Sigma X^2 = 27\,527$	$n_1 = 5$ $n_2 = 6$ $n_3 = 7$ $N = 18$	$df_{BG} = k - 1 = 2$ $df_W = N - k = 15$ $df_T = N - 1 = 17$

$$SS_{BG} = \frac{232^2}{5} + \frac{245^2}{6} + \frac{210^2}{7} - \left[\frac{687^2}{18}\right] = 848.47$$

$$MS_{BG} = \frac{SS_{BG}}{df_{BG}} = \frac{848.47}{2} = 424.24$$

$$SS_W = 27\,527 - \left[\frac{232^2}{5} + \frac{245^2}{6} + \frac{210^2}{7}\right] = 458.03$$

$$MS_W = \frac{SS_W}{df_W} = \frac{458.03}{15} = 30.54$$

$$SS_T = 27\,527 - \frac{678^2}{18} = 1306.50$$

$$F = \frac{MS_{BG}}{MS_W} = \frac{424.24}{30.54} = \mathbf{13.89}$$

$$F_{df} = 2, 15$$

$$\alpha = .05$$

$$F_{crit} = 3.68$$

$$H_0: \mu_1 = \mu_2 = \mu_3$$

Since **13.89** > 3.68, reject the null hypothesis.

Source of variation	*SS*	*df*	*MS*	*F*	*p*
Between groups	848.47	2	424.24	13.89	<.05
Within groups (error)	458.03	15	30.54		
Total	1306.50	17			

The dependent variable is the number of seconds until the child turns on the light. The variation among means is sufficiently large to reject the null hypothesis. Therefore, we conclude statistical evidence exists suggesting the treatment approaches differentially affect the altering of children's fear of the dark.

Using the values from Table 12.2, what is ω^2?

$$\varpi^2 = \frac{86 - 2(0.92)}{94.25 + 0.92}$$

$$\varpi^2 = \frac{84.16}{95.17}$$

$$\varpi^2 = \mathbf{.88}$$

The interpretation of ω^2 = .88 is that 88% of the variation among the scores is accounted for by the levels of the independent variable. We can think of it as the ratio of primary variance (or treatment variance) to the total amount of variation in the study (primary variance + secondary variance). The difference between ω^2 and 100% is the amount of variation due to random factors. Therefore, if ω^2 = 88%, then 12% of the variation in scores is due to random factors. Logically, ω^2 can range from 0 to 100 percent. Is 88% a large treatment effect? Yes, very large! In fact, it would be rather unlikely that such an effect size would be found in the real world of social and behavioral science research. Some researchers (e.g. Cohen, 1977) have created guidelines for interpreting the size of ω^2; however, these are rather arbitrary designations. We simply need to understand that as ω^2 increases, so does the effect size.

The most frequently used measure of effect size is **eta-squared (η^2)**. Even though ω^2 generates a more accurate measure, η^2 is currently preferred, presumably due to its relative ease of calculation. The formula is presented below. As the worked example demonstrates, η^2 tends to overestimate the amount of primary variance.

Formula for eta-squared, η^2

$$\eta^2 = \frac{SS_{BG}}{SS_T} \quad \text{(Formula 12.10)}$$

Using the values from Table 12.2, what is η^2?

$$\eta^2 = \frac{SS_{BG}}{SS_T}$$

$$\eta^2 = \frac{86}{94.25}$$

$$\eta^2 = \mathbf{.91}$$

Historically, published studies did not necessarily report effect sizes. For instance, a published study by Hupka and Eshett in 1988 claimed an effect was found with an F value of 1.23. Looking at the F table (Table A.5), we can see that an F of 1.23 should not lead to a rejection of the null hypothesis. However, the degrees of freedom for this test were 192 and 13 440! Upon calculating ω^2, it was found that the effect size was .003%![2] This clearly shows us that the

2 The authors would like to thank Professor Hupka for making the summary statistics available that allowed for the computation of ω^2.

independent variable had only the slightest of effects on the dependent variable. Thankfully, it is becoming common practice to report effect sizes when publishing in behavioral and social science journals.

12.10 Locating the Source(s) of Significance

After rejecting the overall null hypothesis, the researcher has statistical evidence suggesting that at least two of the means have come from different populations. That, however, is all that is known. It is hard to imagine a situation in which a researcher would not want to discover *which* groups appear to be different from one another. This topic, however, is very complicated. For example, an investigator may wish to make all possible comparisons between the group means; this is the most typical course of action. In a study with three groups, three comparisons would be required; with four groups, six comparisons would be made; and so on. The number of all possible comparisons is $k(k-1)/2$. However, while some comparisons between group means may have important theoretical implications, other comparisons may be of no interest to the investigator. Tests comparing group means that the researcher decides to run *after* observing the sample data are called **post hoc** or **posteriori tests**. When a researcher decides to test a set of specific null hypotheses *before* collecting the data, the subsequent analyses are called **a priori tests** or **planned comparisons**. A priori tests often involve only a subset of all possible comparisons among sample means. Whether a priori or post hoc tests are performed is determined by the nature of the hypotheses in the study. In addition, controlling the probability of a Type I error is still an issue, even when making comparisons after obtaining a significant F value. The topic of **multiple comparisons** is complex, and the investigator has many statistical issues and options to consider. For example, some post hoc tests are so conservative that they can be used without even performing an F test, the first post hoc test we will look at, **Tukey's HSD test**, being one example. (Spotlight 12.2 will tell us more about John Tukey, the creator of this statistical tool.) Most post hoc tests, however, are only allowed if the obtained F allows the researcher to reject the H_0. The second post hoc test we will look at, **Fisher's *LSD* test** (also called the **protected *t* test**), is a case in point. When reading professional journal articles, be alert to the names of some of the common procedures for multiple comparisons – Scheffé, Newman–Keuls, Duncan, and Bonferroni corrected t tests being a few of the common ones. Unfortunately, even a cursory coverage of this area is beyond the scope of this book; the interested reader is referred to any number of advanced statistics books or websites, Cohen (2013) being one example.

One of the most commonly used post hoc statistics is Tukey's *HSD*. This test allows us to compute a single value, called the *honestly significant difference*, or *HSD*, to use as the minimal difference between any two group means, provided

Spotlight 12.2 John Wilder Tukey

John Tukey (1915–2000) was born in New Bedford, Massachusetts. He was homeschooled by his educator parents who responded to his numerous questions not with direct answers but with clues and follow-up questions designed to help him solve his own problems (McCullagh, 2003). This philosophy produced a remarkable student, culminating in two degrees from Brown University in chemistry and a PhD in mathematics from Princeton in 1939, where he was asked to stay on as a professor upon graduation. He stayed at Princeton for his entire career.

During World War II he decided to serve the war effort in the Fire Control Research Office (think "artillery fire") where he tackled many mathematical issues related to ballistics, gun and artillery control, and range firing (e.g. Sande, 2001). After the war, he continued to stay involved in several government projects including the enrichment of uranium and the development of the U-2 spy plane, even representing the US government at a conference in Geneva addressing the discontinuance of nuclear weapons testing.

Concurrent with his academic career, he also worked for AT&T Bell Laboratories where, among other contributions, he created several neologisms including "bit" for binary digit and perhaps most notably "software" as a contrast to hardware (e.g. Leonhardt, 2000). Over the course of his career, Tukey was awarded, among other recognitions, entrance into the prestigious National Academy of Sciences, the National Medal of Science from President Nixon, and the Medal of Honor from the Institute of Electrical and Electronic Engineers.

As an academic, Tukey's contributions were numerous and significant (pun intended!), publishing and producing both individually and collaboratively in the fields of chemistry, mathematics, environmental research, probability and statistics (forming and chairing Princeton's Department of Statistics in 1966), and philosophy (McCullagh, 2003). He also served as a scientific advisor to several presidents and was tapped to be a critical reviewer of influential notable publications such as the *Kinsey Report* (1953) and *Silent Spring* (Carson, 1962). His contributions to the field of statistics are numerous. Some of his most notable statistical creations include the box-and-whisker plot, the stem-and-leaf diagram, and the Tukey range test (we know it as Tukey's *HSD* test). These tools will likely be long-lasting contributions to the field of behavioral and social statistics.

each group in the design has an equal n. If any difference between two group means exceeds Tukey's *HSD*, we can conclude that statistical evidence exists for a difference. Here is the formula.

Formula for Tukey's *HSD*

$$HSD = q\sqrt{\frac{MS_W}{n}} \quad \text{(Formula 12.11)}$$

where

q = the studentized range statistic (Table A.6)
n = the number of scores in each group (must be the same)

To determine the proper q value, we must know the number of groups in the study (k), the degrees of freedom for MS_W (df_W), and select an alpha level (generally the same α selected for the ANOVA).

Since the data in Table 12.4 comes from groups with unequal n's, we will use the data from Table 12.2 to determine the *HSD*:

$$HSD = q\sqrt{\frac{MS_W}{n}}.$$

Given that $k = 3$, $df_W = 9$, and $\alpha = .05$, $q = 3.95$ (see Table A.6),

$$HSD = 3.95\sqrt{\frac{0.92}{4}}$$

$$HSD = \mathbf{1.89}.$$

The group means from Table 12.2 are $M_1 = 4.75$, $M_2 = 8.75$, and $M_3 = 2.25$. Applying Tukey's *HSD* as our post hoc analysis tool directs us to conclude that statistical evidence for population differences exist between all three groups: groups 1 and 2 (4.75 – 8.75 = –4), groups 1 and 3 (4.74 – 2.25 = 2.5), and groups 2 and 3 (8.75 – 2.25 = 6.5). (Mean differences should be understood in absolute value terms since negative values merely indicate that the larger mean was subtracted from the smaller.)

The second posttest we will explore, Fisher's *LSD* (least significant difference) test, has been shown, when compared with other tests of multiple comparisons, to perform well under many circumstances (Cramer & Swanson, 1973). However, Hochberg and Tamhane (1987) advise against using this test when making all possible comparisons with more than three sample means. Under these conditions alpha begins to inflate beyond the 5% level.

To use the protected t test, it is essential that the overall null hypothesis has first been rejected. With this requirement, the probability of a Type I error is much less than if the ANOVA were bypassed. Formula 12.12 is the formula for the protected t test.

Formula for Fisher's *LSD* test

$$t = \frac{M_i - M_j}{\sqrt{MS_W\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \quad \text{(Formula 12.12)}$$

where

M_i, M_j = the means for the two groups being compared
n_i, n_j = the number of scores in each of the two groups being compared

The formula for the protected *t* test is different from what we have encountered when conducting independent- or dependent-samples *t* tests. Even though only two means are contrasted in Formula 12.12, the pooled variance used is MS_W is based on the variance estimates from *all* the groups in the study. The result is a more stable estimate of the population variance. The protected *t* test, consequently, has more power than an independent-samples *t* test when multiple comparisons are performed.

If our study has three groups and we wish to make all possible comparisons, we will have to perform three protected *t* tests (group 1 versus group 2, group 1 versus group 3, group 2 versus group 3). The means in the numerator will change with each test, n_i and n_j may also change, but MS_W is taken from the ANOVA summary table and remains *the same* for each analysis. The t_{crit} value is based on $N - k$ degrees of freedom and is found in the *t* table, *not* the *F* table. Using the data from Table 12.4, Table 12.5 provides an example of how protected *t* tests are conducted. Each comparison uses an alpha level of .05.

The results of the protected *t* tests show that statistical evidence has been found suggesting both emotive imagery and relaxation are more effective than the control condition in reducing children's fear of the dark, $t(15) = 5.09$, $p < .05$; $t(15) = 3.46$, $p < .05$, respectively. No evidence of a difference between the two treatment conditions was found, $t(15) = 1.66$, *n.s.* The three *t* test comparisons have allowed us to locate the source(s) of significance detected by the *F* test.

Table 12.5 Protected *t* tests following a significant ANOVA.

Emotive imagery	Relaxation training	Control
$M_1 = 46.40$	$M_2 = 40.83$	$M_3 = 30.00$
$n_1 = 5$	$n_2 = 6$	$n_3 = 7$

$MS_w = 30.54$ (taken from the ANOVA)

$$M_1 \textit{ versus } M_2 = \frac{46.40 - 40.83}{\sqrt{30.54(1/5 + 1/6)}} = \frac{5.57}{3.36} = \mathbf{1.66}\ (n.s.)$$

$$M_1 \textit{ versus } M_3 = \frac{46.40 - 30.00}{\sqrt{30.54(1/5 + 1/7)}} = \frac{16.40}{3.22} = \mathbf{5.09}\ (p < .05)$$

$$M_2 \textit{ versus } M_3 = \frac{46.83 - 30.00}{\sqrt{30.54(1/6 + 1/7)}} = \frac{10.83}{3.13} = \mathbf{3.46}\ (p < .05)$$

df for t's $= df_W = N - k = 18 - 3 = 15$
$\alpha = .05$
$t_{crit} = \pm 2.13$

The *t* tests are based on the ANOVA in Table 12.4.

Box 12.1 presents a study on the topic of people's loyalty to a group. The experiment has three groups, yet the researchers did not use an ANOVA. This raises an important statistical issue.

Box 12.1 Initiation Rites and Club Loyalty

Many clubs require an initiation. The rites of passage may range from simply learning a password and a secret handshake to initiations that are life threatening. The more dramatic displays of fraternity hazing have been banned by universities amid reports of fatalities. The Marine Corps is infamous for its treatment of recruits during basic training. Yet it is a common observation that clubs that require severe initiations have a great deal of group cohesion and many members who remain loyal for a lifetime. One explanation for this observation is that clubs that require severe initiations attract a certain type of person who is prone to develop strong loyalties. However, might there be something about the kind of initiation experienced by club members and their subsequent feelings about the group? This is the question asked by researchers Aronson and Mills in their classic 1959 study, a study that challenged behaviorism, the then ruling paradigm of psychology. Due to a self-selection factor that invariably operates in club membership, only a randomized controlled experiment could discover if there is a causal relationship between severity of initiation and attitudes of the initiate toward the club.

One theory the researchers felt that could better account for the initiation-induced loyalty was *cognitive dissonance* (Festinger, 1957). According to this theory, we are strongly inclined to maintain consistency between our attitudes and behavior. When an attitude that we hold is inconsistent with our behavior, a state of dissonance arises, creating an unpleasant psychological state. A reduction in dissonance can be accomplished by altering our behavior or, more commonly, by changing our attitude so that consonance is achieved. How does this relate to initiations and club loyalty? No matter how attractive a group is to someone, there are always some negative aspects of the group. After going through an unpleasant initiation, a state of dissonance arises. It is as if people ask themselves, "How could I have gone through all this when there are things about this group I do not like?" To resolve the dissonance, people can either view the initiation as not that bad or disregard the negative aspects of the group and magnify the positive characteristics of the group. The more unpleasant the initiation, the more difficult it is to view the initiation as not that bad. The only avenue left to achieve consonance is to see the group as more positive. The more severe the initiation, the more positive people will tend to see the group. In other words, the greater the dissonance, the larger the shift in how one perceives the group. Aronson and Mills tested this theory in the following manner.

Sixty-three female undergraduates volunteered to join a discussion group, ostensibly so that the researchers could study the dynamics of group interaction. The participants were told that they would have to be screened before being allowed to join the group, and since the discussion topic was to be sexual behavior, it was important that they be able to discuss the topic freely. The "screening" constituted the experimental manipulation. In the "Severe" condition, participants were required to read aloud a number of obscene words related to sex and body parts. The participants were told that the experimenter was rating them on how embarrassed they appeared while reading the words. Furthermore, they were told that the ratings would be used to determine if they would be admitted to group membership. Participants in the "Mild" condition read words that were related to sex but were not obscene. Those participants assigned to the "Control" condition were not required to read any words; their admission to the group was based only on their willingness to discuss the topic of sex. Of course, irrespective of performance, all participants were admitted to the "club."

The participants were told that their participation would begin at the next meeting. However, in order for the participants to "become familiar with the group discussion," they were allowed to listen, via intercom, to a discussion among other initiates. Participants actually heard a tape recording of three women discussing the sexual practices of lower animals. The discussion was dull, trite, and filled with contradictions. Depending on the condition to which the participant was assigned, dissonance had been created. After going through the initiation, they would become a member of a club that would include rather unimpressive people who discussed a potentially interesting topic in a boring manner.

After listening to the discussion, participants completed a questionnaire, which asked them to rate the discussion and the participants along a number of dimensions (e.g. dull–interesting, intelligent–unintelligent). The sum of the ratings served as the dependent variable. One dependent variable was the ratings of the *discussion*; a second dependent variable was the ratings of the *participants*. The authors predicted that the more severe the initiation, the more positive would be participants' ratings of the discussion *and* the three group members heard on the tape. To test their hypothesis, they performed multiple *t* tests, making all possible comparisons among group means, for each dependent variable.

With respect to the ratings of the discussion, participants in the "Severe" condition offered more positive ratings of the discussion than those in the "Mild" and "Control" conditions. No evidence of differences was found between the "Mild" and "Control" groups. Therefore, it would appear that people who experience a severe initiation reduce dissonance by increasing their positive evaluations of the discussion among the members of the group.

The second dependent variable was especially important because it reflected the participants' favorable regard toward the other members of the group. Did

the experimental manipulation influence the participants' liking for the other participants in the group discussion? According to the *t* tests, the answer was a qualified "yes." The difference between the "Severe" and "Control" conditions was statistically significant. However, there were no differences between the "Severe" and "Mild" conditions, nor the "Mild" and "Control" conditions. Based on all of these analyses, the authors concluded, "The results clearly substantiate the hypothesis: persons who undergo a severe initiation to attain membership in a group increase their liking for the group" (p. 181).

We have learned that multiple *t* tests should not be conducted unless the overall null hypothesis has been rejected. Aronson and Mills neglected to perform the requisite *F* tests on the dependent variables. Instead, they proceeded straight to the multiple comparisons. What would have happened if the *F* tests were conducted? Since the authors reported means and standard deviations, one of the textbook authors (Grimm) used Formulas 12.1 and 12.2 to compute an *F* ratio for each dependent variable. The *F* test for the discussion ratings was significant, $F(2, 60) = 6.54$, $p < .05$. For this variable, the authors were warranted in conducting the *t* tests. However, the *F* test performed on the ratings of the participants was not significant, $F(2, 60) = 2.81$, *n.s.* Consequently, for this dependent variable, the authors should *not* have performed *t* tests when making all possible comparisons among the means of the groups.

In conclusion, the findings of this study are not as "clean" as they originally appeared. The reanalysis suggests a more limited conclusion than the one made by the authors. A proper interpretation of the findings would conclude that a severe initiation process increases liking for the *opinions* of the group but not necessarily the *members* of the group.

12.11 How to Present Formally the Conclusions for a One-Way ANOVA

Every journal has a standard format for reporting the results of a statistical test in the text of an article. Most social and behavioral science journals rely on the format offered by the American Psychological Association (2009). When formally reporting the rejection of a null hypothesis, we must include the df_{BG}, df_W, the *F* value, and the alpha level used to make our decision. For instance, "Statistical evidence suggests the time-of-day the drug is administered influenced its effectiveness, $F(3, 69) = 9.59$, $p < .05$. Further analysis found evidence suggesting the effects of the drug were most pronounced if administered in the morning compared to the afternoon, $t(24) = 3.55$, $p < .05$; and morning compared to the evening, $t(22) = 4.59$, $p < .05$. No difference was found between afternoon and evening, $t(23) = 1.22$, *n.s.*" A failure to reject might read, "There was no statistical evidence to suggest the drug was more pronounced at one time of day compared to others, $F(3, 69) = 1.59$, *n.s.*" Measures of effect size can be added at the end of the sentence when a null hypothesis has been rejected.

Many other principles common to the proper reporting of all types of statistical findings were first laid out in Section 8.8. For instance, notice once again that the critical value for the test is usually not presented.

Summary

A one-way ANOVA can be used when a study has two or more levels of one independent variable and participants are placed into independent groups (not repeatedly measured). When conducting an ANOVA it is assumed that the samples are representative of the populations from which they come, these populations are normally distributed and have roughly equivalent variances. Minor violations of the statistical assumptions (normality and homogeneity of variance) are not serious since the F test is considered robust to these assumptions. However, it is essential that the sample data be representative of the population, that each participant's score on the dependent variable be independent of every other participant's score, and that the data be measured on an interval or ratio scale.

The variation among scores within each group is due to random factors: individual differences and experimental error. The influence of random factors creates error variance (also called secondary variance). Mathematically it is referred to as *within-group variance*. *Between-group variance* is caused by error variance as well as the effect of the treatment variable (if there is any). Variance due to a treatment effect can also be called primary variance. To determine if the null hypothesis can be rejected, the between-group variance (MS_{BG}, primary and secondary variance) is compared in ratio form with the within-group variance ($MS_{W,}$ secondary variance only). This ratio is called an F statistic.

If there is no primary variance, the resulting ratio is close to 1. However, as primary variance exists, it only inflates the numerator (MS_{BG}), making the resulting ratio increasingly larger than 1. If the resulting ratio falls into the outermost 5% of the null F sampling distribution, evidence of a treatment effect has been found.

The F sampling distribution that is used to test the significance of the F ratio is determined by the df_{BG} $(k - 1)$ and df_W $(N - k)$. Although the shape of the F distribution will vary depending on sample sizes and the number of conditions, the F distribution is always positively skewed.

The size of the F ratio is not an indication of the magnitude of the treatment effect. To assess the strength of association between the independent variable and the variation of scores on the dependent variable, ω^2 or η^2 can be used. To locate the source(s) of a statistically significant F test, Tukey's *HSD*, Fisher's *LSD*, or any number of other tests can be used to make pairwise comparisons between group means. The selection of the proper test best suited for comparisons depends on an understanding of many factors – a topic that is beyond the scope of this text.

Using Microsoft® Excel and SPSS® to Run a One-Way ANOVA

Excel

General instructions for data entry into Excel can be found in Appendix C.

Data Entry

Enter all of the scores from the samples into adjacent columns (the number of columns equaling the number of conditions in the research design), one sample in each column. Label the columns appropriately. (See Figure 12.6 for an example.)

Cond1	Cond2	Control
4	7	2
4	8	2
5	7	3
5	6	2
4	7	2
5	9	3

Anova: Single factor

Summary

Groups	*Count*	*Sum*	*Average*	*Variance*
Cond1	6	27	4.5	0.3
Cond2	6	44	7.333 333	1.066 667
Control	6	14	2.333 333	0.266 667

ANOVA

Source of variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between groups	75.444 444	2	37.72 222	69.28 571	2.7E-08	3.68 232
Within groups	8.1 666 667	15	0.544 444			
Total	83.611111	17				

Figure 12.6 A worked example using Microsoft Excel to calculate a one-way ANOVA.

Data Analysis

1) Excel has built-in programs for many inferential tests, including the one-way ANOVA test. To access it, click on the Data tab on the top menu and then click **Data Analysis**. (Some versions of Excel have a "Tools" tab. The Data Analysis function may be under this tab.) If this option is not found, the Data Analysis ToolPak needs to be installed. See Excel instruction materials for how to install this feature.
2) With the Data Analysis box open, select **Anova: Single Factor**.
3) Input the data range by dragging over the entire data set and placing those coordinates into the **Input Range** box. (If we included the labels in the data range, make sure to click the **Labels** box to exclude those cells.)
4) Decide on an Output option. The default is to place it on a separate worksheet.

5) Click **OK**.
6) The first output box labeled "Summary" will present the count, sum of all values, means (average), and variance for all conditions. The second output box will be an ANOVA summary table (labeled "ANOVA") that will be very similar to the ANOVA summary table described earlier in the chapter with an additional column identifying the F_{crit} value. (See Figure 12.6 for a worked example.)

SPSS

General instructions for inputting data into SPSS can be found in Appendix C.

Data Entry

In SPSS, each row of the data file represents a participant. Since all samples in a one-way ANOVA test have different participants, all of the dependent variable data from all samples will need to be placed in one column. Within **Variable View**, label this variable appropriately. However, also create a second variable that will allow the user to identify which data goes with which group. A typical label for this variable might be "condition." Then, go to **Data View**. Input the sample data to the appropriate column, and use a nominal variable in the "condition" column to distinguish between the samples (for example, "1," "2," "3," etc.). See Figure 12.7 for an example.

	hoursslept	Condition
1	4	1
2	4	1
3	5	1
4	5	1
5	4	1
6	5	1
7	7	2
8	8	2
9	7	2
10	6	2
11	7	2
12	9	2
13	2	3
14	2	3
15	3	3
16	2	3
17	2	3
18	3	3

Figure 12.7 An example of entered data for a one-way ANOVA in SPSS.

Data Analysis

1) Click **Analyze** on the tool bar, select **Compare Means**, and then click **One-Way ANOVA**.
2) Highlight the dependent variable column label in the left box and click the arrow to move it into the **Dependent List** box. Move the "condition" variable to the **Factor** box.
3) If we want to make specific group comparisons at the time of the ANOVA, click on **Post Hoc** and make the appropriate selections. If not, simply skip this step.
4) If we want to get basic descriptive statistics, click on **Options** and then **Descriptive**. If not, simply skip this step.
5) Click **OK**.
6) The output will generate an ANOVA summary table very similar to the one described earlier in the text. However, "Sum of Squares" is spelled out in this table, and instead of a column labeled *p* for probability, SPSS generates a column labeled Sig. for significance. The meaning, however, is the same. We are looking to see if the *F* obtained falls in the most extreme 5% of the null *F* distribution. If the value found under Sig. is .05 or less, we have evidence to reject the null hypothesis. See Figure 12.8 for a worked example.

One way

ANOVA

hoursslept

	Sum of squares	*df*	Mean square	*F*	Sig.
Between groups	75.444	2	37.722	69.286	.000
Within groups	8.167	15	.544		
Total	83.611	17			

Figure 12.8 An output table from worked example using SPSS to calculate a one-way ANOVA.

Key Formulas

Pooled variance formula for MS_W

$$MS_W = \frac{s_1^2(n_1-1) + s_2^2(n_2-1) + \ldots + s_k^2(n_k-1)}{(n_1-1) + (n_2-1) + \ldots + (n_k-1)} \quad \text{(Formula 12.1)}$$

Between-group variance (MS_{BG}) as an estimate of σ^2

$$ns_M^2 = \sigma^2 \quad \text{(Formula 12.2)}$$

Definitional formula for SS_{BG}

$$SS_{BG} = \Sigma n_k(M_k - M_G)^2 \quad \text{(Formula 12.3)}$$

Computational formula for SS_{BG}

$$SS_{BG} = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right] \quad \text{(Formula 12.4)}$$

Definitional formula for SS_W

$$SS_W = \Sigma(X_1 - M_1)^2 + \Sigma(X_2 - M_2)^2 + \cdots + \Sigma(X_k - M_k)^2 \quad \text{(Formula 12.5)}$$

Computational formula for SS_W

$$SS_W = \Sigma X^2 - \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \ldots + \frac{(\Sigma X_k)^2}{n_k}\right] \quad \text{(Formula 12.6)}$$

Definitional formula for SS_T

$$SS_T = \Sigma(X - M_G)^2 \quad \text{(Formula 12.7)}$$

Computational formula for SS_T

$$SS_T = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \quad \text{(Formula 12.8)}$$

Formula for omega-squared, ω^2

$$\varpi^2 = \frac{SS_{BG} - df_{BG}(MS_W)}{SS_T + MS_W} \quad \text{(Formula 12.9)}$$

Formula for eta-squared, η^2

$$\eta^2 = \frac{SS_{BG}}{SS_T} \quad \text{(Formula 12.10)}$$

Formula for Tukey's *HSD*

$$HSD = q\sqrt{\frac{MS_W}{n}} \quad \text{(Formula 12.11)}$$

Formula for Fisher's *LSD* test

$$t = \frac{M_i - M_j}{\sqrt{MS_W\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \quad \text{(Formula 12.12)}$$

Key Terms

Analysis of variance (ANOVA)
One-way ANOVA
Between-group variation
Treatment (or primary) variance

Individual differences
Experimental error
Random factors
Within-group variation
Error (or secondary) variance
Mean square within, (MS_W)
Mean square between, (MS_{BG})
Grand mean
***F* distribution**
Omega-squared, (ω^2)
Eta-squared, (η^2)
Post hoc (or posteriori) tests
A priori tests (or planned comparisons)
Multiple comparisons
Tukey's *HSD*
Fisher's *LSD* (or protected *t*) test

Questions and Exercises

1 What is the acronym for "analysis of variance?"

2 A one-way ANOVA is used to analyze data from research designs having _____ or more groups that are _____ each other.

3 For designs with more than two conditions, why is an ANOVA preferred over several *t* tests?

4 What terms can be used to describe variance associated with participants being in different conditions?

5 What terms can be used to describe variance found within groups or conditions?

6 The symbol used to represent the variation between group means is _____.

7 The symbol used to represent the variation between scores with groups is _____.

8 The symbol used to represent the number of groups in a research design is _____.

9 How is the MS_{BG}/MS_W ratio used to detect the presence of primary variance?

10 Describe, in brief terms, the shape of *F* distribution.

11 The assumptions for the one-way ANOVA are the same as they are for the _____.

12 Which of the five assumptions can we consider the one-way ANOVA to be robust against slight violations?

13 The table used to organize the ANOVA analysis is called the _____ _____ _____.

14 Complete the following ANOVA summary table. Test to see if the null hypothesis can be rejected.

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	280.3	3	_____	_____	_____
Within groups	_____	_____	7.28		
Total	527.98	_____			

15 Complete the following ANOVA summary table. Test to see if the null hypothesis can be rejected.

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	_____	4	1.47	_____	_____
Within groups	6.30	_____	_____		
Total	_____	15			

16 A sports psychologist would like to compare the effects of different exercise programs on cardiovascular fitness. The measure of fitness is the resting heart rate of participants after they complete the program, with lower heart rates indicating greater physical fitness. Twelve college students are randomly assigned to three groups (four participants per group). Participants in the aerobic condition walk a treadmill for 30 minutes, three times a week. Participants in the circuit condition perform exercises on weight machines for 30 minutes, with a 10-second rest between exercises. In the control condition, participants are asked simply to maintain their usual amount of exercise. The resting heart rates of all participants are taken after 10 weeks. Data from this hypothetical study are presented in the following table.

Aerobic	Circuit	Control
65	74	74
62	65	78
56	62	86
60	72	75

a State the null and alternative hypotheses.
Calculate the following values.
b SS_{BG}
c SS_W
d df_{BG}
e df_W
f MS_{BG}
g MS_W
h SS_T
i df_T
j F ratio
k What is F_{crit} when $\alpha = .05$?
l Should we reject or fail to reject the H_0?
m Make an ANOVA summary table.
n Calculate omega-squared.
o Conduct Fisher's *LSD t* tests if appropriate and properly report the findings.
p Does this F test allow for a causal interpretation?

17 A political scientist hypothesizes that persons from the US Midwest are more conservative in their political views than individuals from either coast. Five participants between the ages of 21 and 60 are randomly selected from the Western, Midwestern, and Eastern parts of the nation ($N = 15$). A questionnaire measuring conservatism is administered, with higher scores reflecting greater conservatism. The data are presented in the following table.

West	Midwest	East
3	9	4
6	15	9
2	9	1
4	4	2
3	6	3

a State the null and alternative hypotheses.
Calculate the following values.
b SS_{BG}
c SS_W
d df_{BG}
e df_W
f MS_{BG}
g MS_W
h SS_T
i df_T
j F ratio
k What is F_{crit} when $\alpha = .05$?
l Should we reject or fail to reject the H_0?
m Make an ANOVA summary table.
n Calculate eta-squared.
o Conduct Tukey's *HSD* and interpret.
p Does this F test allow for a causal interpretation?

18 A clinical psychologist is interested in evaluating treatments for panic attacks. The number of reported panic attacks during the 6-month program of treatment is used as the dependent variable. Fifteen clients suffering from panic disorder are randomly assigned to three conditions (five participants per group). In the breathing condition, clients are taught how to breathe slowly and deeply at the first sign of an attack. Clients in the medication condition are administered a sedative, three times a day. Clients in the control condition are not provided with any treatment. The data are presented in the following table.

Breathing	Medication	Control
16	12	9
22	15	12
15	13	16
9	18	18
13	12	10

a State the null and alternative hypotheses.
Calculate the following values.
b SS_{BG}
c SS_W
d df_{BG}
e df_W

f MS_{BG}
g MS_W
h SS_T
i df_T
j F ratio
k What is F_{crit} when $\alpha = .05$?
l Should we reject or fail to reject the H_0?
m Make an ANOVA summary table.
n Provide an interpretation of the results.
o Calculate omega-squared and interpret.
p Conduct protected t tests if appropriate and interpret.
q Does this F test allow for a causal interpretation?

19 A researcher is interested in the effect of emotion on concentration. A two-sample study is designed in which anger is induced in one sample by having a confederate provoke an argument in the lab waiting room. The control group does not undergo this mood induction. Both samples are then tested on a computer stunt driving game and the number of times the participant runs the vehicle into an object (crashes) is counted. The data follows.

Angry group	Control group
6	6
9	5
13	8
11	6
5	9
10	7

Conduct a one-way ANOVA on these data (even though there are only two levels of the independent variable). Compare the conclusions to Part 4, Problem 10, in which this same study should have been analyzed using an independent-samples t test.

a State the null and alternative hypotheses.
Calculate the following values.
b SS_{BG}
c SS_W
d df_{BG}
e df_W
f MS_{BG}
g MS_W

h SS_T
i df_T
j F ratio
k What is F_{crit} when $\alpha = .05$?
l Should we reject or fail to reject the H_0?
m Make an ANOVA summary table.
n Provide a proper reporting of the findings.
o Calculate omega-squared.
p Does this F test allow for a causal interpretation?

20 We observe that people seem to be happier when they are wearing a new article of clothing. We would also like to test whether level of happiness depends on the particular type of new clothing worn. To test this, we provide a random sample of five of our classmates with new T-shirts and five with new shoes, and instruct them to wear the articles of clothing all day. At the end of the day, we ask these participants to rate, on a 10-point scale, how happy they are. A control group of five classmates is also asked for this self-rating, but without the experimental manipulation. Ratings for each participant are reported below. Higher scores indicate greater happiness. Conduct a one-way ANOVA on these data.

New T-shirt	New shoes	Control
6	8	4
6	7	6
7	9	5
5	7	3
8	10	5

a State the null and alternative hypotheses.
Calculate the following values.
b SS_{BG}
c SS_W
d df_{BG}
e df_W
f MS_{BG}
g MS_W
h SS_T
i df_T
j F ratio

k What is F_{crit} when $\alpha = .05$?
l Should we reject or fail to reject the H_0?
m Make an ANOVA summary table.
n Calculate eta-squared.
o Conduct Tukey's *HSD* and interpret.
p Provide a proper reporting of the findings.
q Does this *F* test allow for a causal interpretation?

21 State the sources of variance of the numerator of the *F* ratio when H_0 is correct and when H_0 is incorrect.

22 Why is an *F* distribution always positively skewed?

23 True or False. In a study that has two levels of one independent variable, it is better to conduct an *F* test rather than a *t* test because the *F* test is more powerful.

24 What are the three sources of variation that can account for mean differences?

25 What are the two sources of variation that can account for within-group variability?

26 When can a multiple comparison tool like Fisher's *LSD* be used?

27 *LSD* stands for _____ _____ _____.

28 *HSD* stands for _____ _____ _____.

29 Select the right answer: In Tukey's test, if the difference between two group means meets or exceeds the *HSD* value, there *is/is not* evidence of a significant difference between those two groups.

Computer Work

30 A clinical psychologist hypothesizes that tension produced by frustration can be relieved if the person is allowed to respond aggressively. However, it is unknown what form the aggression must take in order for tension reduction to occur. All participants in the experiment are asked to complete an intellectually demanding task. While working on the task, the experimenter keeps interrupting the participant, correcting mistakes, offering advice, and slowing the progress of the participant. After this phase of

the experiment, the independent variable is defined by the opportunity afforded the participant to express aggression. In the overt aggression condition, participants become a "teacher" and are required to administer a loud, noxious noise when a confederate learner makes a mistake on a memory task. In the written aggression condition, participants are asked to write an evaluation of the experimenter, which will be made available to the experimenter's supervisor. In the fantasy aggression condition, participants are administered the thematic apperception test. This test is composed of several pictures depicting, for the most part, interpersonal scenes. The participant is asked to make up a story for each card, consequently allowing for the expression of aggressive fantasies. The dependent variable is the change in systolic blood pressure from just after the frustration induction experience to just after the opportunity for participants to express aggression. Use $\alpha = .05$ to test the null hypothesis. Conduct all possible post hoc comparisons if the null hypothesis is rejected. Either use Fisher's *LSD* or another method of pairwise comparisons offered in our statistical package.

Overt	Written	Fantasy
−10	−2	0
−5	+2	−4
−8	0	0
−3	−1	+5
−11	−5	0
+3	+1	−2
−15	−9	0
+3	−1	−2
+4	0	−6
−12	−3	−2
−3	−5	−4
+6	−1	0

31 Researchers have noted that chronic severe muscle contraction headaches respond quite well to antidepressant medication, as well as biofeedback for relaxing the muscles of the forehead (Bourianoff & Stubis, 1988). A health psychologist is interested in making a direct comparison between these two modes of treatment. Forty-five headache sufferers are randomly

assigned to three conditions: medication, biofeedback, and no treatment control. Treatment lasts for five months, during which time the number of weekly headaches is recorded. Conduct an F test ($\alpha = .05$) and post hoc comparisons of our preference to determine the relative effects of these three treatment conditions. The raw scores are the average number of headaches per week, over the five-month period of treatment.

Medication	Biofeedback	Control
2	4	5
1	2	7
2	3	8
6	5	10
7	4	8
8	2	2
6	7	8
3	4	8
2	0	2
0	3	5
1	0	1
2	5	6
0	1	2
4	2	1
5	3	8

32 Sarah believes there is a relationship between a person's wardrobe and the type of major they pursue at the university. The theory states that majors described as a "science" attract people who are not particularly interested in self-presentation. Those students who are interested in the arts, however, have a greater interest in self-presentation. Humanities majors fall in between. Sarah hypothesizes that the number of shoes brought to school might be a way to start to investigate the theory. Below are data from four different majors: a "hard" science, "social" science, humanity, and art. To try to control for any gender effects, Sarah has only recruited biological males for the study. Use $\alpha = .05$ to test the null hypothesis. Conduct all possible post hoc comparisons if the null hypothesis is rejected. Either use Tukey's *HSD* or another method of pairwise comparisons offered in a statistical package.

Chemistry	Psychology	History	Theater
2	5	5	7
3	2	7	4
2	5	8	9
6	5	10	8
4	4	8	9
5	8	4	5
6	7	8	11
3	4	9	4
2	9	2	6
4	5	5	11
2	3	6	7
4	5	6	9
3	1	3	5
4	2	9	10
7	6	8	9
3	5	5	8
4	4	7	5

13

Two-Way Analysis of Variance

13.1 The Research Context

Factorial Designs

In previous chapters, research designs were presented with only one factor, or in experimental terms, one independent variable.[1] However, we do not live in a "one-variable world." Our behavior is constantly affected by the combined influence of two or more sets of conditions. For this reason, many research designs in the social and behavioral sciences allow the researcher to evaluate the **interaction** between two independent variables and measure their combined influence on behavior. **Factorial (or complex) designs** are a blend of two or more single-variable designs and serve at least two purposes. For one thing, an investigator can ascertain, in one study, information about the effect of more than one independent variable, thereby saving the time, expense, and effort required to conduct separate, single-variable experiments. Most importantly, by combining independent variables, information can be gained about the combined effect of the independent variables on the dependent variable. When factorial designs are used, there is typically a prediction made regarding the combined effect. The researcher believes that the effect of one independent variable will be altered, depending on the value of the second independent variable. An example serves to illustrate this point.

If we saw a person in trouble, would we come to their aid? As we think about this question, we may answer, "it depends." It may depend on how many people are present, how dangerous the situation is, or even what the person in trouble looks like (Latane & Darley, 1970). "It depends" qualifies our answer, which, in effect, says our action is determined by the joint presence of certain conditions.

1 As in previous chapters, concepts introduced here will be presented from an experimental perspective, even though two-way ANOVAs can be used to analyze data gathered in nonexperimental designs as well.

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

For instance, we may only help if it is not too dangerous, and we are the only person around to offer assistance. In the language of factorial designs, this is an interaction effect since our behavior depends on the joint occurrence of two or more variables. It might also be true that we would help if there are no other people around or if the person was dressed like us, but we would not help if both of these conditions were true, thinking it might be a trap. It is one thing to learn how one variable (e.g. the presence of other people) influences our behavior, and it is quite another thing to look at the effect of this variable among other variables. The ability to look at combined effects is the great attraction of factorial designs.

Examples of Factorial Designs and Cell Notations

A system of terminology and notation is employed to identify the features of a factorial design. Each design is broken up into **cells**, each cell corresponding to a unique combination of treatment conditions and housing its own group of participants. Each independent variable in a factorial design is referred to as a **factor** (thus the term "factorial design"); this term is more generally appropriate since many designs involve nonexperimental variables. When there are two independent variables, one variable is designated Factor *A* and the other Factor *B*. A cell designation of A_1B_3 refers to the group of participants who are in the first level of Factor *A* and the third level of Factor *B*. In Example 13.1, A_1B_3 refers to the participants who have an anxiety disorder with panic attacks (first level of Factor *A*) and receive Imipramine (third level of Factor *B*). By convention, when the design is depicted with two rows and three columns, it is referred to as a 2 × 3 (pronounced "2 by 3") factorial design (rows × columns). If one of the independent variables has more levels than the other independent variable, the variable with more levels is usually placed at the top of the figure, thereby creating more columns than rows. There is no special logic for designating the letters A and B to the two factors.

► **Example 13.1** A psychiatrist is interested in comparing the effectiveness of three different psychopharmacological treatments for anxiety with two types of patients, those with and without panic attacks. Factor *A* is *patient type*, Factor *B* is *treatment*, and the dependent variable is the participants' self-reports of anxiety. This is a 2 × 3 factorial design.

		Factor *B*		
		Valium	Alprazolam	Imipramine
Factor *A*	Anxiety with Panic	A_1B_1	A_1B_2	A_1B_3
	Anxiety without Panic	A_2B_1	A_2B_2	A_2B_3

◄

► **Example 13.2** A social psychologist hypothesizes that two advertising techniques will be differentially effective, depending on the product. Factor *A* is the *advertising technique*, Factor *B* is the *product*, and the dependent variable is the participants' attitudes toward the product. This design is a 2 × 2 factorial design.

		Factor *B*	
		Autos	Cell Phones
Factor *A*	Image Appeal	A_1B_1	A_1B_2
	Technical Information	A_2B_1	A_2B_2

◄

► **Example 13.3** A psychologist hypothesizes that aggression is more likely when a person is physiologically aroused *and* is exposed to aggressive cues. Factor *A* is *arousal*, Factor *B* is the presence or absence of *aggressive cues*, and the dependent variable is aggression.

		Factor *B*	
		Present	Absent
Factor *A*	Aroused	A_1B_1	A_1B_2
	Not Aroused	A_2B_1	A_2B_2

◄

Main Effects and Interactions

To illustrate how two single-independent variable experiments can be combined into one factorial design, consider two hypothetical studies on memory.

In Experiment 1, a psychologist hypothesizes that memory will be better if a person is in a good mood when attempting to *recall* a list of words. On the first day of the study, all participants are given a list of words to memorize. A test for recall is conducted on the second day, with half of the participants tested immediately after reading a mood-elevating passage. The other half of the participants is tested after reading a depressing passage. If the psychologists' hypothesis is correct, those participants who just finished the mood-elevating passage should remember more words than the participants who just finished the depressing passage (see Experiment 1, Table 13.1).

In Experiment 2, a researcher hypothesizes that memory will be best if an individual is in a good mood while *memorizing* a list of words. This means the mood manipulation (reading a happy or depressing passage) is accomplished on the first day, just prior to having the participants memorize the word list. On the second day, all participants are asked to recall as much of the list as they can. This experiment is also depicted in Table 13.1.

Now suppose a psychologist develops a state-dependent theory of memory in which it is predicted that memory will be facilitated when a person recalls a list of words while in the same emotional state that existed when the list was memorized. Neither one of the experiments in Table 13.1 can test this hypothesis since the prediction requires experimentally manipulating mood state during memorization *and* inducing a mood state during recall. However, by combining the two studies, we can obtain all the information gathered from each single-variable experiment, as well as the interactive effects of the two independent variables. An effect found among the conditions of one independent variable, independent of the influence of another independent variable, is called a **main effect**. As described earlier in the chapter, an effect produced by the combination of independent variables is an *interaction*. The **two-way ANOVA** is an analytical procedure that allows us to investigate both main effects and interactions.

The two independent variables are Mood During Recall and Mood During Memorization. Each independent variable, or factor, has two levels: Happy and Sad. The formal name of this design is a *Completely Randomized 2 × 2 Factorial Design*. It is completely randomized in that participants are randomly assigned to the four treatment conditions. It is a factorial design because there is more than one factor. It is a 2 × 2 design because there are two levels of each

Table 13.1 Two separate between-group designs, each with one independent variable.

Day 1	Day 2
Experiment 1: Mood State During Recall	
Both groups memorize list	Group 1: Recall While Happy
	Group 2: Recall While Sad
Experiment 2: Mood State During Memorization	
Group 1: Memorize While Happy	Both groups recall list
Group 2: Memorize While Sad	

In Experiment 1, the independent variable is "mood during recall." In Experiment 2, the independent variable is "mood during memorization." In both experiments, the dependent variable is the number of words recalled correctly.

Table 13.2 A 2 × 2 factorial design with a significant interaction and no significant main effects.

		Factor *B*: Mood During Recall		
		Happy	Sad	
Factor *A*: Mood During Memorization	Happy	(1) 20	(2) 10	$M_{A_1} = 15$
	Sad	(3) 10	(4) 20	$M_{A_2} = 15$
		$M_{B_1} = 15$	$M_{B_2} = 15$	

The numbers in parentheses are group designations, also called cell numbers.

factor (Happy/Sad During Memorization and Happy/Sad During Recall).[2] Table 13.2 illustrates this 2 × 2 factorial design.

Using Diagrams and Graphs to Examine Main Effects and Interactions

Viewing diagrams and graphs of data from factorial designs allow us to speculate about the presence of main effects and interactions. Of course, we must conduct statistical analyses to discover if there really is statistical evidence for main effects and interactions. However, it can be helpful to graph group means before performing a statistical analysis, and researchers often display graphs in their publications as a visual complement to the verbal summary of study results.

Throughout this section, we will stay with the 2 × 2 memory experiment. Group means will be altered in each successive example so that we can learn what the diagrams and graphs look like as the experimental results change. A word of caution: If we read the text without referring to the relevant diagrams and graphs, we can easily get lost. We need to bounce back and forth between the illustrations and the text to understand the discussion.

In Table 13.2, the numbers within each cell are group means representing the average number of words recalled for the participants in that group. The means in the margins (called marginal means) represent the average number of words recalled for participants across that condition (identified by row or column, as the case may be). Let us first consider each independent variable separately. When addressing only Factor *B*, the design can be simplified by collapsing the conditions of Factor *A* leaving us with just two cells (Factor *B* Happy; Factor

2 Some descriptions may use the term "full factorial." This term means that every possible combination of the levels of the factors is realized. Since this is almost always the case, the simpler term "factorial" is typically used.

B Sad) in the design, and the question is, "What is the effect of mood state during recall on the number of words remembered?" The average of collapsed cells 1 and 3 versus the average of collapsed cells 2 and 4 are found in the marginal means and reveal no difference (M_{B_1} = 15 and M_{B_2} = 15). Comparing the levels within one independent variable allows us to determine if there is a main effect for that factor. Since, in this example, there is clearly no evidence of a main effect for Factor *B,* we might be tempted to conclude that the mood state present *during recall* has no effect on the number of words remembered.

The same logic can be applied when considering Factor *A*. Now the design is simplified by collapsing the conditions of Factor *B* leaving us with just two cells (Factor *A* Happy; Factor *A* Sad) in the design, and the question is, "What is the effect of mood state during memorization on later recall?" The average of collapsed cells 1 and 2 versus 3 and 4 are found in the marginal means and reveal no difference (M_{A_1} = 15 and M_{A_2} = 15). There is no evidence for a main effect for Factor *A*. Once again, we might be tempted to conclude that the mood state present *during memorization* has no effect on the number of words remembered. Bear in mind that these data are hypothetical. The marginal means in Table 13.2 have purposely been presented as identical to simplify the example. With real data, the marginal means are rarely identical, even when there is no main effect. Nevertheless, if each independent variable were examined in separate, single-variable studies, each study would be a washout. However, by using a factorial design, a third question is possible: "Does the mood state during memorization interact with the mood state during recall?" Although there are no main effects in this study, it would appear that there is an interaction. By examining the pattern of means within the cells, we may conclude that recall is facilitated when there is congruence between mood during memorization and mood during recall.[3] It does not seem to matter what the moods are as long as they are similar. These results lend support to the proposed state-dependent theory of memory.

When speculating about an interaction, a graph of cell means can be helpful. Figure 13.1 is a graph of the means presented in Table 13.2.

When graphing means from a factorial design, the levels of one independent variable are indicated on the horizontal axis. (Which variable to select is largely arbitrary. Sometimes visual analysis is a bit clearer depending on which variable is chosen.) In Figure 13.1, Mood During Recall has been placed on the horizontal axis. The dependent variable is represented as means on the vertical axis. The second independent variable is placed within the graph. In the example, the

3 Since this example uses a 2 × 2 design, the interaction can be examined by taking the average of the diagonals. The average of cells 1 and 4 is 20 and the average of cells 2 and 3 is 10. The difference between the marginal means of 10 and 20 (not depicted) likely indicates an interaction effect. This method *only* applies to a 2 × 2 factorial design since any other factorial design (e.g. a 2 × 3) will not have diagonals.

Figure 13.1 A graph of the group means in Table 13.2. There is an interaction but no main effects.

second independent variable is Mood During Memorization. Since there are two levels of this variable, there will be two lines on the graph: one line for each level of the factor that is embedded into the graph. Do not make the mistake of thinking that each line depicts each independent variable.[4]

In Figure 13.1, the line drawn from the upper left to the lower right of the graph represents the Happy Mood During Memorization level of Factor *A*. The line connects the means of two groups: Happy During Memorization/Happy During Recall (Group 1) and Happy During Memorization/Sad During Recall (Group 2). Carefully examine the means in Table 13.2 and the graph in Figure 13.1. Match up the cell means from the table with the points in the graph, and see how the lines on the graph follow the cells from the perspective of the Mood During Memorization perspective.

Table 13.2 illustrates data in which there are no main effects but there is an interaction. On a graph, *an interaction is revealed when the lines are not parallel*. Figure 13.1 reveals the interaction; the lines are not only nonparallel, they also cross. The lines, however, do not have to cross when there is an interaction; they just have to be nonparallel. How nonparallel the lines need to be for there to be an interaction depends on the statistical power in each particular study. In

4 Another common mistake among students is thinking that the number of cells equals the number of independent variables. Keep in mind that independent variables (factors) *always* have levels.

Table 13.3 A 2 × 2 factorial design with one main effect (Factor *B*) and no interaction.

		Factor *B*: Mood During Recall		
		Happy	Sad	
Factor *A*: Mood During Memorization	Happy	(1) 30	(2) 12	$M_{A_1} = 21$
	Sad	(3) 28	(4) 10	$M_{A_2} = 19$
		$M_{B_1} = 29$	$M_{B_2} = 11$	

general, however, the more they depart from being parallel, the more likely it is that evidence for an interaction will be found when the statistical analysis is performed.

Up to this point, data has been used in which there is an interaction but no main effects. We will now use the same experiment, adjust the cell means, and illustrate various combinations of main effects and interactions.

The data in Table 13.3 show a main effect for Factor *B* but no main effect for Factor *A* and no interaction. Although the marginal means for Factor *A* are not identical, they are rather close; most likely, there is not a main effect. Factor *B* is another matter. The marginal means of 29 and 11 are strikingly different, indicating that a main effect would likely be found when the data is analyzed.

The interpretation of the presumed experimental results depicted in Table 13.3 is as follows. The number of words remembered is greatest when participants are happy during the recall task (main effect for Factor *B*). It makes no difference what mood participants are in when they memorize the word list (no main effect for Factor *A*). Further, whether the mood states during memorization and recall are congruent or incongruent is of no consequence (no interaction effect).

Figure 13.2 graphically displays the means in Table 13.3. Notice that the lines are parallel, reflecting the absence of an interaction. When examining a graph for a main effect, look to see if one of the lines is higher than the other. If so, then there may be a main effect. In identifying a main effect, to what extent does one line have to be higher than the other? It depends on the statistical power in that study, so only a statistical analysis can identify evidence of a main effect. In general, however, the greater distance between the lines, the more likely it is that evidence for a main effect will be found when the statistical analysis is performed.

In Figure 13.2, Mood During Memorization is placed on the horizontal axis; therefore, two lines are drawn, which indicate the two levels of Mood During Recall. Why the switch from Figure 13.1 to 13.2? By changing the variable on the horizontal axis, the main effect is shown more clearly. Figure 13.3 also

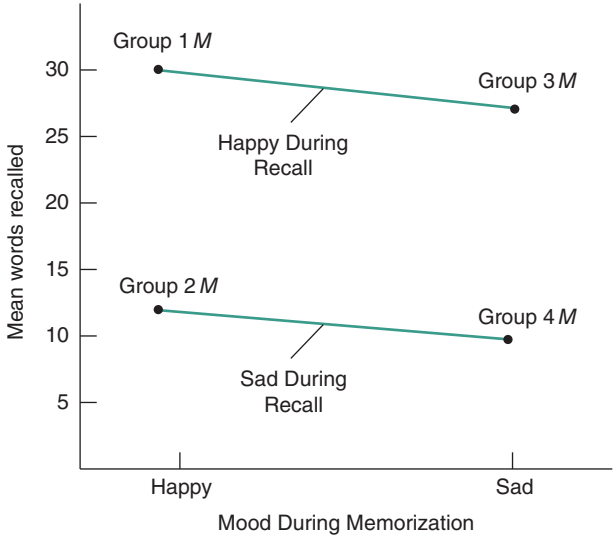


Figure 13.2 A graph of the means in Table 13.3. There is one main effect but no interaction.

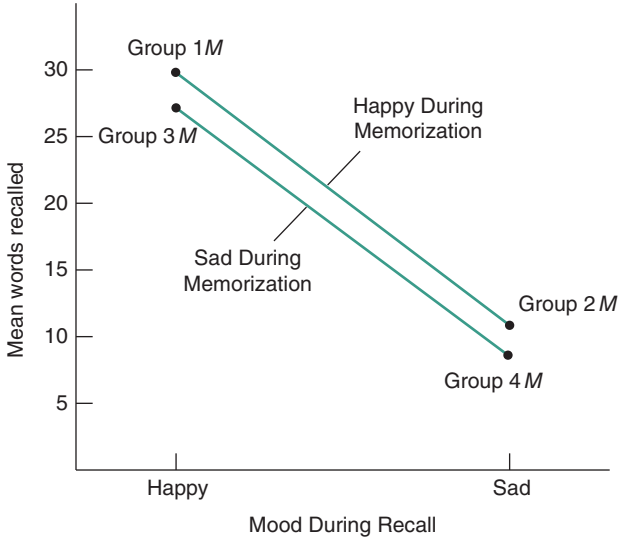


Figure 13.3 Group means from Table 13.3 are graphed. There is a main effect for Mood During Recall. However, by placing Mood During Recall on the *X* axis, the main effect is more difficult to identify. Figure 13.2 is a more useful display of the main effect presented in Table 13.3.

illustrates a graph using the means in Table 13.3. This time, however, the *X* axis is Mood During Recall. The lines are still parallel, indicating the absence of an interaction. The Happy During Memorization line is only slightly higher than the line drawn for Sad During Memorization. Displaying the means in this manner might obscure the main effect for Factor *B* (Mood During Recall); this version of the graph would not be as helpful.

After the data are analyzed, we will know which factors (if any) show a main effect, and the graph can be drawn accordingly. When computing a two-way ANOVA by hand, it is a good idea to draw the graph both ways beforehand – with Factor *A* on the *X* axis and then with Factor *B* on the *X* axis. This will enable us to know ahead of time what the results of the ANOVA are likely to reveal. The guiding principle in graphing is to draw our figure so that it can be easily interpreted at a glance.

At this point, we should be able to place means in the cells of a diagram that illustrate a main effect for Factor *A*. Table 13.4 is one way to illustrate a main effect for Mood During Memorization.

The pattern of means in Table 13.4 shows that recall is facilitated when participants are happy while memorizing the list of words (main effect for Factor *A*). Irrespective of the level of Factor *A*, recall is not affected by the mood state present during recall (no main effect for Factor *B*). Finally, the manner in which mood states combine during memorization and during recall does not have an influence on the number of words recalled (no interaction).

In Figure 13.4, the lines are nearly parallel: no interaction. The main effect for Mood During Memorization is revealed by the different heights of the lines. The Happy During Memorization line is highest, which reflects the greater recall of information between the two Happy During Memorization groups in comparison to the two Sad During Memorization groups.

Table 13.4 A 2 × 2 factorial design with one main effect (Factor *A*) and no interaction.

		Factor I: Mood During Recall		
		Happy	Sad	
Factor *A*: Mood During Memorization	Happy	(1) 30	(2) 28	$M_{A_1} = 29$
	Sad	(3) 20	(4) 20	$M_{A_2} = 20$
		$M_{B_1} = 25$	$M_{B_2} = 24$	

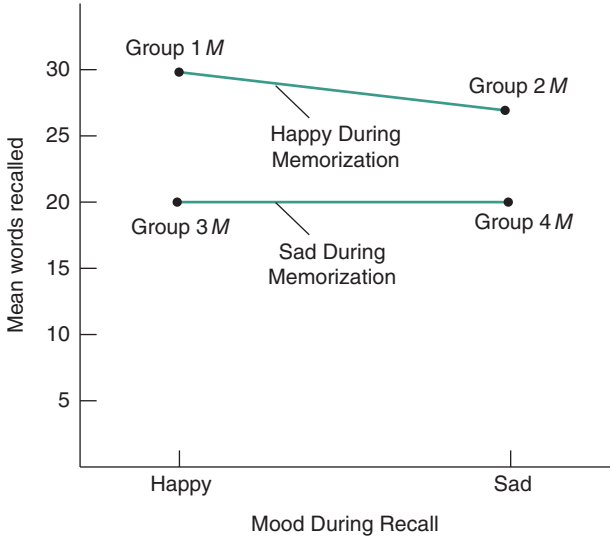


Figure 13.4 The graph of the means in Table 13.4. There is a main effect for Factor *A* but no interaction.

Up to this point, we have looked at several different analytical scenarios. Table 13.5 presents yet another. In this scenario, we find a main effect for Factor *A*, as well as an interaction, but no main effect for Factor *B*.

Table 13.5 shows that mood state during recall does not influence the number of words recalled (no main effect for Factor *B*). However, the main effect revealed for Factor *A* indicates that recall is enhanced when words are memorized while in a happy mood. Now examine Figure 13.5. One line is higher than the other, revealing the main effect for Mood During Memorization. Because the lines are not parallel, an interaction is also indicated. When more than

Table 13.5 A 2 × 2 factorial design with one main effect (Factor *A*) as well as an interaction.

		Factor *B*: Mood During Recall		
		Happy	Sad	
Factor *A*: Mood During Memorization	Happy	(1) 40	(2) 28	$M_{A_1} = 34$
	Sad	(3) 20	(4) 26	$M_{A_2} = 23$
		$M_{B_1} = 30$	$M_{B_2} = 27$	

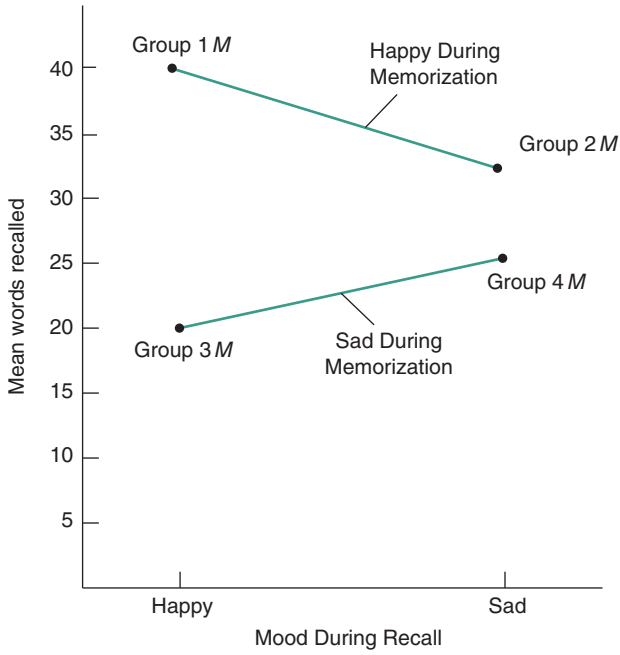


Figure 13.5 The graph of the means from Table 13.5. There is a main effect for Factor *A* and an interaction.

one *type* of effect is found, keep this rule in mind, *always interpret the* ***higher-order effect*** *first, and then interpret lower effects cautiously in light of the initial analysis*. The term *higher-order effect* means the effect involving the most number of factors. In this situation, it means we need to first interpret the interaction before we interpret the main effect. The interaction would be interpreted as follows: Recall is, on average, greatest when the mood states between memorization and recall are the same.

Refer to Table 13.5. Notice the marginal means for Mood During Memorization (34 vs. 23). That is the main effect. However, the mean in cell 1 (40) is pulling up the marginal mean of 34. The interaction may be *carrying* the main effect. The Happy/Happy condition shows the greatest influence on recall. The congruence between the two Sad conditions (cell 4) does not show a similar influence on recall ($M_4 = 26$). Therefore, not only does the Happy/Happy condition contribute to the interaction, it also seems to account for the main effect for Factor *A*. In situations like this, we must be very cautious about drawing conclusions regarding the main effect for Factor *A*. This main effect might be an *artifact* of the higher-order interaction and may not be *real*. Main effects that can be explained by an interaction are referred to as *illusory* main effects.

Some methodologists take the extreme position that main effects should not be interpreted when an interaction is found. A more moderate position is to interpret the interaction first and then cautiously interpret any lower-order effects (main effects) within the context of the interaction analysis and the theory used to justify the study.

Our illustrations of some of the outcomes of a 2×2 factorial design have not exhausted all possible outcomes. The entire set of possible outcomes is:

1) Factor *A* main effect, no Factor *B* main effect, and no interaction.
2) Factor *B* main effect, no Factor *A* main effect, and no interaction.
3) Main effects for Factors *A* and *B*, no interaction.
4) Main effect for Factor *A*, no main effect for Factor *B*, but an interaction.
5) Main effect for Factor *B*, no main effect for Factor *A*, but an interaction.
6) Main effects for both factors and an interaction.
7) No main effects for either factor but an interaction.
8) No main effects and no interaction.

To recap, a main effect addresses the differences among levels of an independent variable. The number of independent variables is the same as the number of potential main effects. An interaction is the combined influence of two or more independent variables. A significant interaction means that the influence of an independent variable changes based on the level of a second independent variable. Interactions may generate illusory main effects; interpret cautiously.

Factorial Designs with More Than Two Independent Variables

When three independent variables are combined in a factorial design, the number of possible outcomes is increased. Not only are there potential main effects for each independent variable, but more than one interaction can occur. When there are three independent variables, the investigator will test for four interactions: $A \times B$, $A \times C$, $B \times C$, and $A \times B \times C$. The latter interaction is called a three-way interaction; it is a higher-order effect than the two-way interactions. It is possible to design experiments with four-and five-way interactions. However, a significant four- or five-way interaction is often very difficult to interpret; for this reason, researchers tend to avoid overly complex factorial designs. This text will only examine designs with no more than two factors.

13.2 The Logic of the Two-Way ANOVA

The Null and Alternative Hypotheses

There are three separate null hypotheses when conducting a two-way ANOVA, each requiring a separate *F* ratio to test them.

Main Effect for Factor *A*

The independent variable designated as Factor A has two or more levels. The null hypothesis states no differences between the population means of the *levels* of Factor A. Symbolically this can be represented as

$$H_0: \mu_{A_1} = \mu_{A_2} = \mu_{A_k}$$

The subscript k refers to the last level of Factor A. The alternative hypothesis is that at least one of the levels of the Factor A population means is different from one of the other levels of Factor A. This can be expressed as

$$H_1: \text{At least two Factor} A\ \mu\text{'s are not equal}$$

Main Effect for Factor *B*

The independent variable designated as Factor B has two or more levels. The null hypothesis states no differences between the population means of the *levels* of Factor B. Symbolically this can be represented as

$$H_0: \mu_{B_1} = \mu_{B_2} = \mu_{B_k}$$

Now the subscript k refers to the last level of Factor B. The alternative hypothesis is that at least one of the levels of the Factor B population means is different from one of the other levels of Factor B. This can be expressed as

$$H_1: \text{At least two Factor} B\ \mu\text{'s are not equal}$$

The $A \times B$ Interaction

The null hypothesis for the interaction is that there is no interaction. That is, the effect of Factor A is independent of the effect of Factor B. This can be represented as

$$H_0: \text{There is no interaction}$$

The alternative hypothesis is that at least one unique effect, not reducible to a main effect, will be found. This can be expressed as

$$H_1: \text{There is an interaction}$$

Partitioning Variability

The structure of the two-way ANOVA depicted in Figure 13.6 is an extension of the one-way ANOVA. In fact, the first stage of the structure for a two-way ANOVA is identical to the one-way ANOVA: Total variance is due to between-group variance plus within-group variance. The second stage of the model is a further partitioning of the between-group variance into the variance

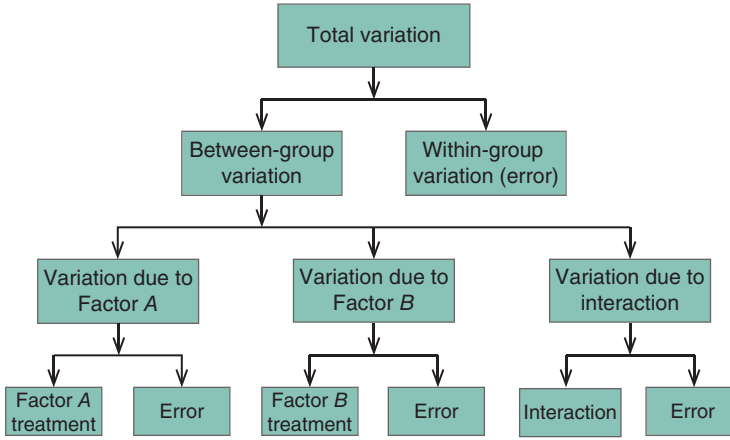


Figure 13.6 Partitioning the total variation in the two-way ANOVA.

due to Factor *A*, the variance due to Factor *B*, and the variance due to the interaction. The second stage of the model yields the *F* ratios that are used to test null hypotheses. An explanation of each source of variability is provided in the following sections.

Between-Group Variability

Between-group variability refers to the variability among *all* the means in the study.[5] In the one-way ANOVA, all of the treatment variance goes into the numerator of the *F* ratio. In the two-way ANOVA, the treatment variance is associated with either Factor *A*, Factor *B*, or the combination of Factor *A* *and* Factor *B*.

Factor *A* Variability

Factor *A* variability refers to the difference among the means of the levels of Factor *A*. These mean differences are due to Factor *A* treatment variance plus error variance.

Factor *B* Variability

Factor *B* variability refers to the difference among the means of the levels of Factor *B*. These mean differences are due to Factor *B* treatment variance plus error variance.

5 Refer to Chapter 12 to review the concepts of between-group and within-group variability, treatment variance, error variance, individual differences, experimental error, and random factors.

Interaction Variability

The variability of the interaction is due to the combined influence of Factors *A and B*, plus error variance.

Within-Group Variability

Within-group variance is the average of the variances *within* each group. Within-group variance is also called *error variance* or simply *error*. The variability within each group is due to two sets of factors: individual differences and experimental error. Recall that individual differences refer to the influence of participant variables on the dependent variable. Experimental error is the variability among scores due to such things as the unreliability of measuring instrumentation, inconsistent interactions between the experimenter and the participants, and random forms of environmental disturbances. Individual differences and experimental error are unsystematic random factors that do not introduce confounding variance into the study.

The Conceptual Form of the Three *F* Ratios

The two-way ANOVA yields three different *F* ratios: one for Factor *A*, Factor *B*, and an interaction. The conceptual basis of these *F* ratios should have a familiar look.

$$F_A = \frac{Treatment\ A\ effect + error\ variance}{error\ variance}$$

$$F_B = \frac{Treatment\ B\ effect + error\ variance}{error\ variance}$$

$$F_{A \times B} = \frac{Treatment\ A \times B\ effect + error\ variance}{error\ variance}$$

As with the one-way ANOVA, the effect due to treatment is placed in the numerator. If there is no treatment effect, or, in other words, if H_0 is true, the *F* ratio is reduced to a measure of error variance divided by another measure of error variance. Therefore, when the null hypothesis is correct, the *F* ratio should be close to 1. As the effect due to treatment increases, the numerator grows and the *F* ratio will become increasingly greater than 1. At some point, the *F* ratio will exceed the critical *F* value associated with the test, and our decision rule will direct us to reject the null hypothesis.

We now turn to the computational steps and formulas used to calculate the three *F* ratios of the two-way ANOVA.

13.3 Definitional and Computational Formulas for the Two-Way ANOVA

Summary of the Computational Steps

The two-way ANOVA involves the computation of the following values.

1. SS_T	8. MS_A
2. SS_{BG}	9. MS_B
3. SS_W	10. $MS_{A \times B}$
4. SS_A	11. F_A
5. SS_B	12. F_B
6. $SS_{A \times B}$	13. $F_{A \times B}$
7. *df* for each *SS*	

Formulas for the Sums of Squares

The formulas, both definitional and computational, will be presented in this section. Analyzing two-way ANOVA's by hand calculation is a long and arduous task. The formulas are presented here to provide a mathematical understanding of the procedure. Use of statistical processing software is strongly suggested for computation.

The Total Sum of Squares, SS_T

The definitional formula for SS_T is the sum of the squared deviations of all the scores from the grand mean, M_G. The grand mean is the mean of *all* the scores in the study.

Definitional formula for SS_T

$$SS_T = \Sigma(X - M_G)^2 \quad \text{(Formula 13.1)}$$

The definitional and computational formulas for SS_T are identical to those used in conducting a one-way ANOVA. The SS_T computes the sum of squares for the *entire* set of scores, N.

Computational formula for SS_T

$$SS_T = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \quad \text{(Formula 13.2)}$$

The Sum of Squares Between Groups, SS_{BG}

The definitional formula for SS_{BG} reminds us that the overall between-group variability is the amount of variation obtained by the sum of the squared differences between each group's mean and the grand mean.

Definitional formula for SS_{BG}

$$SS_{BG} = \Sigma n_k (M_k - M_G)^2 \quad \text{(Formula 13.3)}$$

This definitional formula is identical to the formula for SS_{BG} used in the one-way ANOVA. However, with the two-way ANOVA, remember that the SS_{BG} is not used in an F ratio. The SS_{BG} is only used as a computational check. The computational formula for SS_{BG} for the two-way ANOVA is conceptually identical to the one-way ANOVA. The increased complexity is entirely due to the additional cells in a factorial design.

Computational formula for SS_{BG}

$$SS_{BG} = \frac{(\Sigma X_{A_1B_1})^2}{n_{A_1B_1}} + \frac{(\Sigma X_{A_1B_2})^2}{n_{A_1B_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left(\frac{(\Sigma X)^2}{N}\right) \quad \text{(Formula 13.4)}$$

where

$(\Sigma X_{A_1B_1})^2$, $(\Sigma X_{A_1B_2})^2$ = the sum of the scores in Group 1, quantity squared; the sum of the scores in Group 2, quantity squared, etc.
$(\Sigma X_k)^2$ = the sum of the scores in the last Group, quantity squared
$(\Sigma X)^2$ = the sum of all the scores in the study, quantity squared
$n_{A_1B_1}$, $n_{A_1B_2}$, n_k = the number of participants in Groups 1, 2, and the last group, respectively
N = the total number of participants

The Sum of Squares Within Groups, SS_W

The SS_W is found by calculating the sum of squares within each cell (group) and adding them together. The definitional formula reflects the fact that within-group variability is derived from the deviation of single scores about the mean of the group from which the scores are taken. These formulas have the same form as the SS_W formulas used in the one-way ANOVA. The subscripts of X identify specific cells.

Definitional formula for SS_W

$$SS_W = \Sigma(X_{A_1B_1} - M_{A_1B_1})^2 + \Sigma(X_{A_1B_2} - M_{A_1B_2})^2 + \Sigma(X_{A_2B_1} - M_{A_2B_1})^2 + \cdots + \Sigma(X_k - M_k)^2 \quad \text{(Formula 13.5)}$$

Computational formula for SS_W

$$SS_W = \Sigma X^2 - \left[\frac{(\Sigma X_{A_1B_1})^2}{n_{A_1B_1}} + \frac{(\Sigma X_{A_1B_2})^2}{n_{A_1B_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k}\right] \quad \text{(Formula 13.6)}$$

Computational Check

Formulas have been presented for SS_{BG}, SS_W, and SS_T. In the first stage of the two-way ANOVA, total variability is partitioned into between-group and within-group variability (see Figure 13.6). Therefore,

$$SS_T = SS_{BG} + SS_W.$$

The second stage of the two-way ANOVA partitions the sum of squares between groups into the sum of squares for Factor A, SS_A, the sum of squares for Factor B, SS_B, and the sum of squares for the interaction, $SS_{A\times B}$.

The Sum of Squares for Factor *A*, SS_A

We are used to working with deviations of raw scores around group means (SS_W) and the deviations of group means around the grand mean (SS_{BG}). When calculating the sum of squares for a factor, the means of the *levels* of the factor are used (the marginal means), not the individual cell means. The definitional formula for SS_A reveals that SS_A is a measure of the deviations of the means of each level of Factor A around the grand mean.

Definitional formula for SS_A

$$SS_A = n_{A_1}(M_{A_1} - M_G)^2 + n_{A_2}(M_{A_2} - M_G)^2 + \cdots + n_k(M_k - M_G)^2 \quad \text{(Formula 13.7)}$$

where

n_{A_1}, n_{A_2}, n_k = the number of participants in Factor A levels 1, 2, to the last level, respectively

M_{A_1}, M_{A_2}, M_k = the Factor A means for levels 1, 2, to the last level, respectively; also called *marginal means of each level*

Computational formula for SS_A

$$SS_A = \frac{(\Sigma X_{A_1})^2}{n_{A_1}} + \frac{(\Sigma X_{A_2})^2}{n_{A_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right] \quad \text{(Formula 13.8)}$$

The computational formula requires us to sum scores *across* the cells of each level of Factor A. Therefore, ΣX_{A_1} is the sum of the scores in cell A_1B_1 plus the sum of the scores in cell A_1B_2, plus the sum of scores in A_1B_k. The ΣX_{A_2} is the sum of scores in cell A_2B_1 plus A_2B_2, plus A_2B_k.

The Sum of Squares for Factor *B*, SS_B

Everything said about Factor A applies to Factor B. Just keep in mind that the action now occurs with the *levels* of Factor B. Remember to attend to the subscripts.

Definitional formula for SS_B

$$SS_B = n_{B_1}(M_{B_1} - M_G)^2 + n_{B_2}(M_{B_2} - M_G)^2 + \cdots + n_k(M_k - M_G)^2 \quad \text{(Formula 13.9)}$$

Computational formula for SS_B

$$SS_B = \frac{(\Sigma X_{B_1})^2}{n_{B_1}} + \frac{(\Sigma X_{B_2})^2}{n_{B_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right] \quad \text{(Formula 13.10)}$$

The Interaction Sum of Squares, $SS_{A \times B}$

The sum of squares for the interaction, $SS_{A \times B}$, involves subtracting various marginal means from cell means, adding the grand mean, and multiplying everything by the number of participants in one group. Giving a detailed verbal description of the formula runs the risk of creating more confusion than clarity. So without further comment, Formula 13.11 is used to compute $SS_{A \times B}$ for a 2×2 design.

Computational formula for $SS_{A \times B}$

$$SS_{A \times B} = n_k[(M_{A_1B_1} - M_{A_1} - M_{B_1} + M_G)^2 + (M_{A_2B_1} - M_{A_2} - M_{B_1} + M_G)^2 + (M_{A_1B_2} - M_{A_1} - M_{B_2} + M_G)^2 + (M_{A_2B_2} - M_{A_2} - M_{B_2} + M_G)^2] \quad \text{(Formula 13.11)}$$

This formula can only be used when each group has the same number of participants. The value n_k is the number of participants in *one* group, not the total number of participants in the study. It is assumed that $n_1 = n_2 = n_3 = n_k$.

Computational Checks

In the second stage of the two-way ANOVA, the SS_{BG} is partitioned into SS_A, SS_B, and $SS_{A \times B}$ (see Figure 13.6). Therefore,

$$SS_{BG} = SS_A + SS_B + SS_{A \times B}$$

Even though SS_{BG} is not used in an F ratio, its calculation is justified because it serves as a computational check for SS_A, SS_B, and $SS_{A \times B}$.

The $SS_{A \times B}$ term can be computed by

$$SS_{A \times B} = SS_{BG} - SS_A - SS_B$$

However, it is recommended that $SS_{A \times B}$ be computed separately, using Formula 13.11, and then a computational check performed.

Unequal Numbers of Participants

Researchers strive to include the same number of participants in each experimental condition. Statisticians have noted that the *F* test is more robust under minor violations of the population assumptions (i.e. normality and equivalent variances) when there are the same numbers of participants in each group. Most importantly, unequal sample sizes present serious difficulties when they occur in the context of a factorial design. For a discussion of the conceptual issues and computational adjustments related to factorial designs with unequal sample sizes, refer to Keppel and Wickens (2004).

Partitioning Degrees of Freedom

Each *SS* in the analysis of variance has corresponding degrees of freedom. Partitioning the degrees of freedom follows the same logic as partitioning the variability. Figure 13.7 shows how the degrees of freedom are partitioned. Table 13.6 lists the various degrees of freedom and their computation. Note that

$$df_{BG} = df_A + df_B + df_{A \times B} \text{ and } df_T = df_{BG} + df_W$$

Therefore,

$$df_T = df_A + df_B + df_{A \times B} + df_W$$

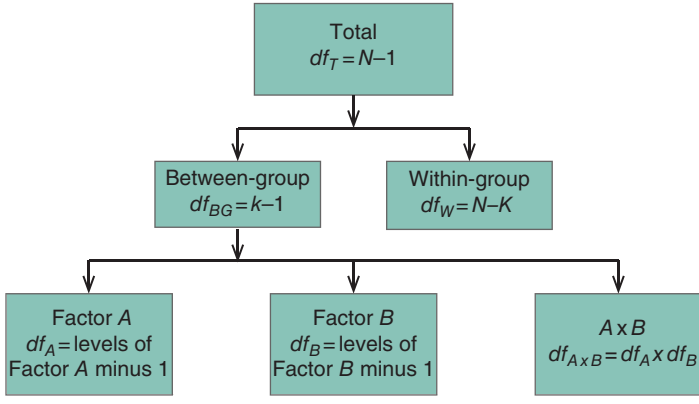


Figure 13.7 Partitioning the degrees of freedom in the two-way ANOVA.

Calculating Mean Squares and *F* Ratios

The last step in the analysis of variance is to calculate the mean squares and the *F* ratios for Factor *A*, Factor *B*, and the interaction. As in the one-way ANOVA, a mean square is a sample variance and has the general form

$$MS = \frac{SS}{df}$$

The *MS* for the factors, interaction, and error term are

$$MS_A = \frac{SS_A}{df_A}$$

$$MS_B = \frac{SS_B}{df_B}$$

$$MS_{A \times B} = \frac{SS_{A \times B}}{df_{A \times B}}$$

$$MS_W = \frac{SS_W}{df_W}$$

When calculating the *F* ratios, the denominator of each *F* is MS_W. Therefore,

$$F_A = \frac{MS_A}{MS_W}$$

$$F_B = \frac{MS_B}{MS_W}$$

$$F_{A \times B} = \frac{MS_{A \times B}}{MS_W}$$

Table 13.6 The degrees of freedom and their computation.

Source	Degrees of freedom	Symbol
Total	$N - 1$	df_T
Within groups	$N - k$	df_W
Between groups	$k - 1$	df_{BG}
Factor *A*	Levels of Factor *A* minus 1	df_A
Factor *B*	Levels of Factor *B* minus 1	df_B
Interaction	$df_A \times df_B$	$df_{A \times B}$

Worked Problem

To provide an experimental context for the calculations of the two-way ANOVA, the following hypothetical experiment will be used. Table 13.7 presents the design. A clinical psychologist is interested in the relative effectiveness of two popular forms of therapy: behavioral therapy and psychoanalysis. However, let us assume that there is some reason to believe that behavioral therapy may be more effective with anxiety problems and psychoanalysis may be a more effective treatment for depression. In other words, an interaction is predicted.

Factor A is the type of clinical problem and has two levels: anxiety and depression. Note that Factor A is not experimentally manipulated. This factor is based on a participant variable. If it turns out that there is a significant main effect for this factor, no cause–effect interpretation can be advanced. This is not the case for Factor B. Factor B also has two levels: behavioral therapy and psychoanalysis. Since this factor is *created* and manipulated by the experimenter, it is an independent variable, and a mean difference between the levels of this factor can be interpreted using causal language. Since both Factor A and Factor B have two levels, this is an example of a 2×2 factorial design. The dependent variable is improvement ratings offered by an independent observer, with higher numbers indicating greater improvement. Table 13.8 contains the raw data for each group in the study. Table 13.9 shows the summary statistics of the raw data. Figure 13.8 is a graph of the means. By examining the graph, we should be able to predict how the analysis will turn out.

Table 13.7 The 2×2 factorial design of the worked problem, including cell notations.

		Factor B		
		Behavioral Therapy	Psychoanalysis	
Factor A	Anxiety	A_1B_1	A_1B_2	A_1
	Depression	A_2B_1	A_2B_2	A_2
		B_1	B_2	

Table 13.8 The raw data for the worked problem.

		Factor B	
		Behavioral Therapy	Psychoanalysis
Factor A	Anxiety	8 6 6 9 8	3 4 1 6 2
	Depression	4 7 4 5 5	8 9 7 7 8

Table 13.9 Summary statistics for the raw data of Table 13.8.

		Factor B		
		Behavioral Therapy	Psychoanalysis	
Factor A	Anxiety	$M_{A_1B_1} = 7.40$ $\Sigma X^2_{A_1B_1} = 281$ $\Sigma X_{A_1B_1} = 37$ $n_{A_1B_1} = 5$	$M_{A_1B_2} = 3.20$ $\Sigma X^2_{A_1B_2} = 66$ $\Sigma X_{A_1B_2} = 16$ $n_{A_1B_2} = 5$	$M_{A_1} = 5.30$
	Depression	$M_{A_2B_1} = 5.00$ $\Sigma X^2_{A_2B_1} = 131$ $\Sigma X_{A_2B_1} = 25$ $n_{A_2B_1} = 5$	$M_{A_2B_2} = 7.80$ $\Sigma X^2_{A_2B_2} = 307$ $\Sigma X_{A_2B_2} = 39$ $n_{A_2B_2} = 5$	$M_{A_2} = 6.40$
		$M_{B_1} = 6.20$	$M_{B_2} = 5.50$	$M_G = 5.85$

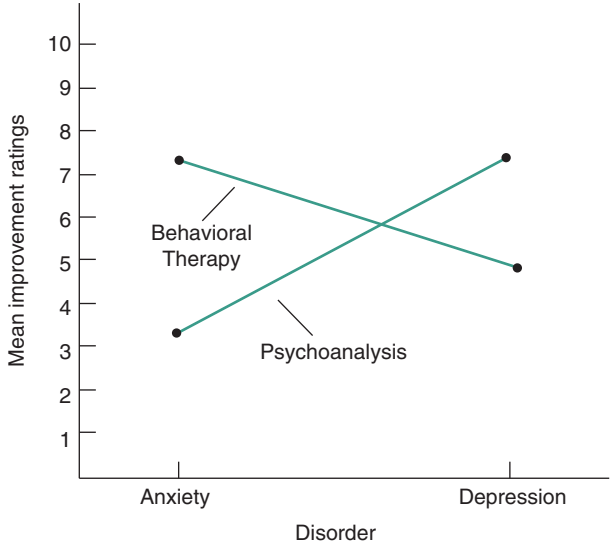


Figure 13.8 A graph of the cell means from Table 13.9.

Calculating the Sums of Squares

Step 1. Find the total sum of squares, SS_T.

$$SS_T = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \quad \text{(Formula 13.2)}$$

$$SS_T = 785 - \frac{(117)^2}{20}$$

$$SS_T = \mathbf{100.55}\ .$$

Step 2. Find the sum of squares between groups, SS_{BG}.

$$SS_{BG} = \frac{(\Sigma X_{A_1B_1})^2}{n_{A_1B_1}} + \frac{(\Sigma X_{A_1B_2})^2}{n_{A_1B_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left(\frac{(\Sigma X)^2}{N}\right) \text{(Formula 13.4)}$$

$$SS_{BG} = \frac{(37)^2}{5} + \frac{(16)^2}{5} + \frac{(25)^2}{5} + \frac{(39)^2}{5} - \left[\frac{(117)^2}{20}\right]$$

$$SS_{BG} = 754.20 - 684.45$$

$$SS_{BG} = \mathbf{69.45}$$

Step 3. Find sum of squares within groups, SS_W.

$$SS_W = \Sigma X^2 - \left[\frac{(\Sigma X_{A_1B_1})^2}{n_{A_1B_1}} + \frac{(\Sigma X_{A_1B_2})^2}{n_{A_1B_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k}\right] \quad \text{(Formula 13.6)}$$

$$SS_W = 785 - \left[\frac{(37)^2}{5} + \frac{(16)^2}{5} + \frac{(25)^2}{5} + \frac{(39)^2}{5}\right]$$

$$SS_W = 785 - 754.20$$

$$SS_W = \mathbf{30.80}$$

Step 4. Perform a computational check for SS_T.

In the first stage of the ANOVA, total variability is partitioned into between-group and within-group variability. Therefore,

$$SS_T = SS_{BG} + SS_W$$

Using the obtained sums of squares values,

$$100.55 = 69.75 + 30.80$$

The second stage of the two-way ANOVA partitions the sum of squares between groups into the sum of squares for Factor A, SS_A, the sum of squares for Factor B, SS_B, and the sum of squares for the interaction, $SS_{A\times B}$.

Step 5. Find the sum of squares for Factor A, SS_A.

$$SS_A = \frac{(\Sigma X_{A_1})^2}{n_{A_1}} + \frac{(\Sigma X_{A_2})^2}{n_{A_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right] \quad \text{(Formula 13.8)}$$

$$SS_A = \frac{(53)^2}{10} + \frac{(64)^2}{10} - \left[\frac{(117)^2}{20}\right]$$

$$SS_A = 690.50 - 684.45$$

$$SS_A = \mathbf{6.05}$$

Refer to Table 13.9 and note that $\Sigma X_{A_1} = \Sigma X_{A_1B_1} + \Sigma X_{A_1B_2} = 37 + 16 = 53$. Likewise, $\Sigma X_{A_2} = \Sigma X_{A_2B_1} + \Sigma X_{A_2B_2} = 25 + 39 = 64$.

Step 6. Find the sum of squares for Factor *B*, SS_B.

$$SS_B = \frac{(\Sigma X_{B_1})^2}{n_{B_1}} + \frac{(\Sigma X_{B_2})^2}{n_{B_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right] \quad \text{(Formula 13.10)}$$

$$SS_B = \frac{(62)^2}{10} + \frac{(55)^2}{10} - \left[\frac{(117)^2}{20}\right]$$

$$SS_B = 686.90 + 684.45$$

$$SS_B = \mathbf{2.45}$$

Step 7. Find the interaction sum of squares, $SS_{A \times B}$.

$$\begin{aligned} SS_{A \times B} = n_k[&(M_{A_1B_1} - M_{A_1} - M_{B_1} + M_G)^2 \\ &+ (M_{A_2B_1} - M_{A_2} - M_{B_1} + M_G)^2 \\ &+ (M_{A_1B_2} - M_{A_1} - M_{B_2} + M_G)^2 \\ &+ (M_{A_2B_2} - M_{A_2} - M_{B_2} + M_G)^2] \end{aligned} \quad \text{(Formula 13.11)}$$

$$\begin{aligned} SS_{A \times B} = 5[&(7.40 - 5.30 - 6.20 + 5.85)^2 \\ &+ (5.00 - 6.40 - 6.20 + 5.85)^2 \\ &+ (3.20 - 5.30 - 5.50 + 5.85)^2 \\ &+ (7.80 - 6.40 - 5.50 + 5.85)^2] \end{aligned}$$

$$= \mathbf{61.25}$$

Step 8. Perform a computational check for SS_{BG}.

In the second stage of the two-way ANOVA, the SS_{BG} is partitioned into SS_A, SS_B, and $SS_{A \times B}$ (see Figure 13.6). Therefore,

$$SS_{BG} = SS_A + SS_B + SS_{A \times B}$$

A check of the calculations shows

$$69.75 = 6.05 + 2.45 + 61.25$$

Step 9. Compute MS_W using SS_W and df_W. The df_W = the total number of participants minus the number of groups ($df_W = N - k = 20 - 4 = 16$).

$$MS_W = \frac{SS_W}{df_W} = \frac{30.80}{16} = \mathbf{1.93}$$

Step 10. Compute MS_A using SS_A and df_A. The df_A = the number of levels of Factor A minus 1 (2 − 1 = 1).

$$MS_A = \frac{SS_A}{df_A} = \frac{6.05}{1} = \mathbf{6.05}$$

Step 11. Compute F_A.

$$F_A = \frac{MS_A}{MS_W} = \frac{6.05}{1.93} = \mathbf{3.13}$$

Step 12. Compute MS_B using SS_B and df_B. The df_B = the number of levels of Factor B minus 1 (2 − 1 = 1).

$$MS_B = \frac{SS_B}{df_B} = \frac{2.45}{1} = \mathbf{2.45}$$

Step 13. Compute F_B.

$$F_B = \frac{MS_B}{MS_W} = \frac{2.45}{1.93} = \mathbf{1.27}$$

Step 14. Compute $MS_{A \times B}$ using $SS_{A \times B}$ and $df_{A \times B}$. The $df_{A \times B} = df_A \times df_B$ (1 × 1 = 1).

$$MS_{A \times B} = \frac{SS_{A \times B}}{df_{A \times B}} = \frac{61.25}{1} = \mathbf{61.25}$$

Step 15. Compute $F_{A \times B}$.

$$F_{A \times B} = \frac{MS_{A \times B}}{MS_W} = \frac{61.25}{1.93} = \mathbf{31.74}$$

The remaining steps involve using the F values to test the null hypotheses.

13.4 Using the *F* Ratios to Test Null Hypotheses

To determine if the F ratios are sufficiently large to reject null hypotheses, the obtained F's must be compared with the appropriate F_{crit} values found in Table A.5. This is the same table used to test the F ratio of the one-way ANOVA. Each of the three F ratios has degrees of freedom for the numerator and the denominator (refer to Table 13.6).

The degrees of freedom for F_A is df_A, df_W (1 and 16 in the worked example).
The degrees of freedom for F_B is df_B, df_W (1 and 16 in the worked example).
The degrees of freedom for $F_{A \times B}$ is $df_{A \times B}$, df_W (1 and 16 in the worked example).

For this design, the pair of *df* for each *F* ratio is the same. Different research designs will generate pairs of *df*'s, which are not all the same.

Since in our example the *df* is the same for each *F* ratio, there is only one F_{crit} value needed for comparison. According to the *F* table, F_{crit} is 4.49 when $\alpha = .05$ and 8.53 when $\alpha = .01$. We will use the critical value for $\alpha = .05$.

Since F_A (3.13) is smaller than F_{crit} when alpha is .05 (4.49), we will fail to reject the null hypothesis for Factor *A*. The same conclusion is reached for Factor *B* since $F_B = 1.27$ and $1.27 < 4.49$. However, $F_{A \times B} = 31.74$; this value is larger than F_{crit} (4.49). As a result, we will reject the null hypothesis that there is no interaction.

Interpreting the Findings

What do these findings mean about the therapy study? Our only rejected null hypothesis concerns the interaction. Interpreting this interaction is straightforward. Examining the group means indicates that psychoanalysis seems to be a superior treatment for people suffering from depression, whereas the treatment of choice for people troubled by anxiety seems to be behavioral therapy. Discovering a significant interaction always leads to a qualified interpretation regarding the role of an independent variable. Neither main effect null hypotheses were rejected. If one or both would have been rejected, we would need to interpret these lower-order findings carefully. They may be legitimate, or they may be illusory. A close inspection of the cell means (perhaps in graphical form) can be helpful.

The ANOVA Summary Table

The results of the foregoing analysis are summarized in Table 13.10. The structure of the two-way ANOVA summary table is similar to the one-way ANOVA.

Table 13.10 The ANOVA summary table based on the analysis of the data in Table 13.8.

Source	*SS*	*df*	*MS*	*F*	*p*
Factor *A* (Disorder)	6.05	1	6.05	3.13	*n.s.*
Factor *B* (Treatment)	2.45	1	2.45	1.27	*n.s.*
$A \times B$	61.25	1	61.25	31.74	< .05
Within groups	30.80	16	1.93		
Total	100.55	19			

However, the "between-groups" row is replaced by rows for Factor A, Factor B, and the $A \times B$ interaction. As we examine the values in the table, verify that SS_T equals the sum of all the sum of squares and that df_T equals the sum of all the df's.

Box 13.1 presents a study on aggression. The researchers used a factorial design to examine the independent and combined effects of arousal and the presence of aggressive cues on aggression.

Box 13.1 Do Firearms Create Aggression?

All of us have displayed aggressive behavior at times to our benefit, and at times, perhaps, to our embarrassment. Psychologists disagree on the roots of aggression. Freud (e.g. 1922) believed that everyone is born with an aggressive instinct. More recently, theorists believe that aggression is largely a learned response (e.g. Anderson & Bushman, 2002; Bandura 1973, 1983). Because aggression can have such a profound effect on the course of our lives, ranging from verbal abuse between spouses to wars among nations, the factors that influence aggression have received a great deal of attention from psychologists. Leonard Berkowitz conducted research on aggression for over 30 years. Common sense tells us that we are more likely to act aggressively when we are angry. Berkowitz took this observation one step further by hypothesizing that, once angry, people would behave even more aggressively if there were aggressive cues present. An aggressive cue is anything that we associate in our minds with aggression. A gun is one of the best examples of an aggressive cue. To test the hypothesis that aggressive cues augment aggression, Berkowitz conducted the following, now classic, psychological study (Berkowitz & LePage, 1967).

Overview of the Design

The design was a 2×3 factorial. One independent variable manipulated participants' anger; the two levels of this variable were low anger and high anger (Factor *A*). The second independent variable manipulated the presence of aggressive cues; this factor (Factor *B*) had three levels: presence of aggressive cues associated with someone in the study, presence of aggressive cues not associated with someone in the study, and the absence of aggressive cues. The design is illustrated in Table 13.11.

Experimental Procedure

Only biological male undergraduates served as participants in this study. They believed the purpose of the research was to examine physiological reactions to stress. Each participant would have to solve a problem, with the foreknowledge

that the partner would evaluate the adequacy of the given solution. The partner was actually a confederate; someone the student thought was another participant but who was actually working for the experimenter. The student was asked to list ideas that a publicity agent could use to increase the popularity of a professional singer. The confederate evaluated the participants' ideas by delivering electric shocks. The number of electric shocks administered to the participant served as the Anger manipulation. In the Low-Anger condition, the student received only one shock, which meant that the confederate deemed the ideas to be good. In the High-Anger condition, the participant received seven shocks, which meant that the ideas were poor. Then the situation was reversed so that the participant had an opportunity to evaluate the confederate's ideas by administering anywhere from one to ten shocks to the confederate. If we consider only this aspect of the experiment, we have a nonfactorial design. We could address the question of whether angered participants are more aggressive than those not angered. The means of the two groups could be analyzed with a *t* test. Now we will consider the second independent variable.

When it came time for the participant to evaluate the confederate's answers, the experimental instructions and the table upon which rested the shock apparatus were rearranged. In the Associated Weapons condition, a 12-gauge shotgun and a .38-caliber handgun were placed in full view of the participant, next to the shock apparatus. These participants were told that the guns were to be used by the "confederate" in another experiment.[6] In the Unassociated Weapons condition, the participants were told that the guns "belonged to someone else who must have been doing an experiment in here." The Associated/Unassociated manipulation was included in the design to test the hypothesis "that aggressive stimuli, which also were associated with the anger instigator, would evoke the strongest aggressive reaction from the participants." The third level of this independent variable, No Aggressive Cues, had the participants use the shock apparatus in the absence of the guns. Again, the dependent variable was the number of shocks the participants delivered to the confederate. Factor *A* allowed Berkowitz and LePage to examine the effect of anger on aggression. Factor *B* allowed them to see if there are differences in aggression due to aggressive cues. Furthermore, because the design is factorial, the researchers could look for an interaction. The interaction of interest was, "Will the most aggression be observed among participants who are angered *and* exposed to aggressive cues?" Table 13.11 shows the mean number of shocks delivered by the participants to the confederate. Table 13.12 presents the ANOVA summary table.

6 Understandably, the data from 20% of the participants in this condition had to be discarded because they did not believe the experimenter.

Table 13.11 Mean number of shocks delivered to the confederate.

	Factor *B*: Aggressive Cues (Weapons)		
Factor *A*: Anger	Associated	Unassociated	None
One shock (low)	2.60	2.20	3.07
Seven shocks (high)	6.07	5.67	4.67

Table 13.12 The ANOVA summary table.

Source	*SS*	*df*	*MS*	*F*	*p*
Anger (*A*)	182.04	1	182.04	104.62	<.01
Weapons Associated (*B*)	3.80	2	1.90	1.09	n.s.
$A \times B$	17.46	2	8.73	5.02	<.01
Within groups	146.16	84	1.74		
Total	349.46	89			

Examine Tables 13.11 and 13.12 and interpret the findings (we may want to compute all the marginal means to help us see the main effect). By examining the rows of Table 13.11 and noting the significant main effect for anger (shocks received), it is clear that the number of shocks delivered by the participants in these two conditions is affected by how many shocks they received from the confederate. More aggression was displayed by those participants who had just received the high number of shocks. What about the main effect for Factor *B*? The nonsignificant *F* ratio in Table 13.12 means that the manipulation of aggressive cues alone has no effect on aggression. In other words, simply viewing guns does not induce aggression.

The significant interaction tells us that aggression is also affected by the *combined* influence of anger and aggressive cues. Exactly how these variables combine to influence aggression cannot be determined by the two-way ANOVA. Follow-up tests are required to identify which groups differ among each other. After conducting these tests, the authors concluded that while anger increases aggression, the presence of aggressive cues during an episode of anger further increases a person's level of aggression.

13.5 Assumptions of the Two-Way ANOVA

The assumptions for the two-way ANOVA are the same as those for the one-way ANOVA. Stated succinctly,

1) The samples are representative of the populations from which they come.
2) Observations are independent of one another.
3) Gathered data comes from an interval or ratio scale.
4) The populations from which the data come are normally distributed.
5) The variances of each population distribution are the same.

Once again, the *F* test is robust and therefore can be performed when the last two assumptions are not *strictly* met. However, gross violations of these assumptions require the use of statistical tests (nonparametric tests), which do not require the populations to be normally distributed with equal variances.

13.6 Measuring Effect Sizes for a Two-Way ANOVA

The *F* tests in the two-way ANOVA allow us to test the null hypothesis for each independent variable as well as the interaction. However, the *F* values do not provide information about the *size* of each effect. Chapter 12 used omega-squared (ω^2) and eta-squared (η^2) to determine the amount of variability in the scores due to the independent variable. With a two-way ANOVA, we can use either of these same two measures to compute a size for the effect of Factor *A*, Factor *B*, and the interaction. Of course, the issue of effect size only arises when a null hypothesis has been rejected and, if being a lower-order effect, it has been interpreted to be more than just an artifact of a higher-order effect. The formulas for calculating ω^2 are slightly different for each potential effect.

Formulas for omega-squared, ϖ_A^2, ϖ_B^2, $\varpi_{A\times B}^2$

$$\varpi_A^2 = \frac{SS_A - (df_A)MS_W}{SS_T + MS_W} \quad \text{(Formula 13.12)}$$

$$\varpi_B^2 = \frac{SS_B - (df_B)MS_W}{SS_T + MS_W} \quad \text{(Formula 13.13)}$$

$$\varpi_{A\times B}^2 = \frac{SS_{A\times B} - (df_{A\times B})MS_W}{SS_T + MS_W} \quad \text{(Formula 13.14)}$$

Since, in the worked problem, only the interaction was significant, we would only calculate $\omega_{A\times B}^2$. Using values from Table 13.10 we find

$$\varpi_{A\times B}^2 = \frac{61.25 - (1)1.93}{100.55 + 1.93} = \mathbf{0.58}$$

This value reflects the ratio between the amount of primary variance associated with the interaction effect and the total variance in the study. Larger values reflect larger effect sizes. An effect size of 58% is unusually large.

Recall from Chapter 12 that although ω^2 is a more refined measure of effect size, the most common current practice in the behavioral and social sciences is to report a simpler statistic, **eta-squared** (η^2). The formulas are presented below.

Formulas for eta-squared, η_A^2, η_B^2, $\eta_{A\times B}^2$

$$\eta_A^2 = \frac{SS_A}{SS_A + SS_W} \quad \text{(Formula 13.15)}$$

$$\eta_B^2 = \frac{SS_B}{SS_B + SS_W} \quad \text{(Formula 13.16)}$$

$$\eta_{A\times B}^2 = \frac{SS_{A\times B}}{SS_{A\times B} + SS_W} \quad \text{(Formula 13.17)}$$

Since, in the worked problem, only the interaction was significant, we would only calculate $\eta_{A\times B}^2$. Using the values from Table 13.10 we find

$$\eta_{A\times B}^2 = \frac{61.25}{61.25 + 30.80} = \mathbf{0.67}$$

When we compare the two measures, we see that η^2 generates a larger estimate of the effect size compared to ω^2.

13.7 Multiple Comparisons

Multiple comparisons are conducted to help interpret why a null hypothesis is being rejected. If we were to conduct a study with only two groups, perform a *t* test, and find evidence to reject the null hypothesis, would we need to follow with other comparisons? No, with only two groups, a significant *t* test is all that is needed to locate the difference. Likewise, in a 2 × 2 factorial design, a significant main effect tells us that the null hypothesis of no difference between the two levels of the factor can be rejected. However, a significant interaction means that we have evidence that at least two cell means are sufficiently different from each other. Follow-up tests allow us to make pairwise comparisons between pairs of group means that share one factor. The pattern of pairwise comparisons either rejecting a null hypothesis or not is interpreted by the researcher to gain a clearer understanding of the interaction.

To illustrate the use of multiple comparisons for interpreting rejected null hypotheses, another hypothetical study is provided. The multiple comparison procedures used are the same ones presented in Chapter 12: Tukey's *HSD* and Fisher's *LSD* (or protected *t* test). First, we will use Tukey's *HSD* procedure.

If the null hypothesis for the interaction effect is rejected, then an *HSD* for the interaction effect can be used to see which pair(s) of cell means is driving the interaction. If the null hypothesis for either Factor *A* or Factor *B* is rejected, and if there are more than two conditions for that factor, then an *HSD* can be used to determine which pairs of means are driving the main effect. Here are the formulas for the various Tukey's *HSD* tests associated with a two-way ANOVA.

Formulas for Tukey's *HSD* values, HSD_A, HSD_B, $HSD_{A \times B}$

$$HSD_A = q_A \sqrt{\frac{MS_W}{n_A}} \quad \text{(Formula 13.18)}$$

$$HSD_B = q_B \sqrt{\frac{MS_W}{n_B}} \quad \text{(Formula 13.19)}$$

$$HSD_{A \times B} = q_{A \times B} \sqrt{\frac{MS_W}{n_{A \times B}}} \quad \text{(Formula 13.20)}$$

where

q_A, q_B, $q_{A \times B}$ = the studentized range statistic (Table A.6); for $q_{A \times B}$ use adjusted k as indicated below.
n_A = the number of scores in each level of Factor A (must be the same)
n_B = the number of scores in each level of Factor B (must be the same)
$n_{A \times B}$ = the number of scores in each cell, i.e. combined level of Factor A and Factor B (must be the same)

Design of study	Number of cell means	Adjusted value of *k*
2 × 2	4	3
2 × 3	6	5
2 × 4	8	6
3 × 3	9	7
3 × 4	12	8
4 × 4	16	10

Our hypothetical study for demonstrating the use of multiple comparisons is an extension of the previous study, which involves the effectiveness of two therapy techniques for the treatment of anxiety and depression. By adding a control group, the design becomes a 2 × 3 factorial design. Table 13.13 illustrates the design.

We will forego the computational steps of this study and proceed directly to the results of the analysis of variance. Table 13.14 presents the cell and marginal

Table 13.13 A 2 × 3 factorial design.

		Factor *B*		
		Behavioral Therapy	Psychoanalysis	Control
Factor *A*	Anxiety	A_1B_1	A_1B_2	A_1B_3
	Depression	A_2B_1	A_2B_2	A_2B_3

Table 13.14 The sample sizes, cell, and marginal means for a hypothetical therapy study.

		Factor *B*			
		Behavioral Therapy	Psychoanalysis	Control	
Factor *A*	Anxiety	$M_{A_1B_1} = 5.80$ $n_{A_1B_1} = 10$	$M_{A_1B_2} = 7.70$ $n_{A_1B_2} = 10$	$M_{A_1B_3} = 4.60$ $n_{A_1B_3} = 10$	$M_{A_1} = 6.03$
	Depression	$M_{A_2B_1} = 5.50$ $n_{A_2B_1} = 10$	$M_{A_2B_2} = 7.80$ $n_{A_2B_2} = 10$	$M_{A_2B_3} = 5.80$ $n_{A_2B_3} = 10$	$M_{A_2} = 5.37$
		$M_{B_1} = 5.65$	$M_{B_2} = 7.75$	$M_{B_3} = 3.70$	

means, as well as the sample sizes. Table 13.15 presents the ANOVA summary table.

The results of the ANOVA show a significant main effect for Factor *B*, mode of therapy. The Factor *A* main effect, clinical disorder, and the interaction are not statistically significant. Therefore, the issue of locating the source(s) of significance only arises for Factor *B*.

Tukey's *HSD* for Factor *B*

$$HSD_B = q_B\sqrt{\frac{MS_W}{n_B}}$$

$$HSD_B = 3.41\sqrt{\frac{3.37}{20}} = \mathbf{1.40}$$

The q value was found using a table on the Internet (Table A.6 in Appendix A is incomplete). Once an *HSD* value is determined, the difference between each pair of means in Factor *B* can be compared (Behavioral Therapy versus Psychoanalysis, –2.1; Behavioral Therapy versus Control, 1.95; and Psychoanalysis versus Control, 4.05). The findings suggest evidence exists for all three

Table 13.15 The ANOVA summary table based on the data in Table 13.14.

Source	*SS*	*df*	*MS*	*F*	*p*
Factor *A* (Disorder)	6.67	1	6.67	1.98	n.s.
Factor *B* (Treatment)	164.10	2	82.05	24.35	$p < .01$
$A \times B$	10.03	2	5.01	1.49	n.s.
Within groups	181.80	54	3.37		
Total	362.60	59			

tested differences. Looking at the means, we can say statistical evidence suggests psychoanalysis worked better than both behavioral therapy and the control, while behavioral therapy worked better than the control.

Now we will perform multiple comparisons using another procedure, Fisher's *LSD* test. The formula for Fisher's *LSD* is:

Formula for Fisher's *LSD* test, two-way ANOVA

$$t = \frac{M_i - M_j}{\sqrt{MS_W\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \quad \text{(Formula 13.21)}$$

where

M_i, M_j = the means for the two levels or cells being compared
n_i, n_j = the number of participants for each level or cell being compared

If we are testing the difference between two levels of a factor, M_i and M_j refer to the means of the two levels. Similarly, n_i and n_j refer to the number of participants in each level. If we are making a comparison between two cells, M_i and M_j refer to the two cell means and n_i and n_j are the number of participants in each cell.

Fisher's *LSD* test relies on the *t* distribution. The null hypothesis for each comparison is that M_i and M_j come from the same population. The critical value is the same for each pairwise comparison and is found in the *t* table (Table A.2). The critical value is found by entering the column for a two-tailed test at the desired alpha level and entering the row corresponding to the degrees of freedom. We will set alpha at .05. The degrees of freedom is df_W or $N - k$. From the ANOVA summary table, df_W is 54. The critical value is approximately ±2.01 (Table A.2 in Appendix A is incomplete). Each protected *t* value is compared to this critical value of ±2.01.

Behavioral therapy versus psychoanalysis

$$t = \frac{5.65 - 7.75}{\sqrt{3.37\left(\frac{1}{20} + \frac{1}{20}\right)}} = \frac{-2.10}{0.58} = \mathbf{-3.62}$$

Behavioral therapy versus control

$$t = \frac{5.65 - 3.70}{\sqrt{3.37\left(\frac{1}{20} + \frac{1}{20}\right)}} = \frac{1.95}{0.58} = \mathbf{3.36}$$

Psychoanalysis versus control

$$t = \frac{7.75 - 3.70}{\sqrt{3.37\left(\frac{1}{20} + \frac{1}{20}\right)}} = \frac{4.05}{0.58} = \mathbf{6.98}$$

Since all of the obtained t values fall outside of ±2.01, we have statistical evidence for all three differences. This is the same result found when using Tukey's *HSD*. As was mentioned in Chapter 12, there are numerous multiple comparison tools; each is designed to be used under different research situations. Which one to use in a given situation is a very complex topic and beyond the scope of this text. However, two different and frequently used follow-up tests have been presented here for our use. Please consult advanced behavioral statistics books or websites for further information.

Box 13.2 Next Steps with ANOVA

This text allocates three chapters to ANOVAs, Chapters 12–14. (Chapter 14 will introduce us to the repeated-measures ANOVA, similar to the one-way ANOVA but one in which the participants are repeatedly measured across all conditions.) However, there is much more that could be said about this extremely important family of analyses. This box is going to introduce us briefly to other procedures in the ANOVA family that researchers can use to analyze more complicated research designs.

A *mixed-design ANOVA* is similar to a two-way ANOVA. It can be used to analyze data from a research design where one factor is between groups (just like the two-way ANOVA presented in Chapter 13) but the other factor is repeated measures (see Chapter 14). Imagine a situation in which depressed and nondepressed individuals are measured for their ability to use two different mnemonic techniques. The factor "depression state" would be a between-groups factor with participants in one or the other group, but "mnemonic device" could

be used as a repeated-measures factor; where each participant is measured once using each of the two techniques.

A *multivariate analysis of variance (MANOVA)* is a form of ANOVA that analyzes more than one dependent variable. Imagine, in the previous study, if we measured not only "memory success" but also "time needed for recall." A MANOVA allows us to analyze both dependent variables in the same analysis.

Analysis of covariance (ANCOVA) is a form of ANOVA that allows the researcher to look for effects after a covariate (a variable that is suspected to covary with a factor) has been removed. Think of this technique as a way to remove the possible effects of a confounding variable. For instance, what if we wanted to remove variance associated with the participant's energy level (perhaps we feared that nondepressed people might perform better simply because they were more energetic during the procedure)? If "energy level" can be measured, an ANCOVA provides us with a mechanism to remove this variance.

We can also combine the features of the MANOVA with the features of an ANCOVA, a *multivariate analysis of covariance* (MANCOVA). This ANOVA analyzes multiple dependent variables and also factors out a concerning covariate. MANOVAs, ANCOVAs, and MANCOVAs can be used with between-groups, repeated-measures, and mixed designs. As this box shows, there are many ANOVA possibilities available to the researcher.

13.8 Interpreting the Factors in a Two-Way ANOVA

Factorial designs frequently use a participant variable as one of the factors. Personality type, biological sex, age, and psychodiagnosis are examples of participant variables that might be used as one factor in a two-way ANOVA design. It is important to remember that when a participant variable is used as a factor, no cause–effect statement can be made regarding the relationship between this variable and the dependent variable. This aspect of the design is correlational because participants are selected based on their standing on that variable. In other words, the experimenter does not manipulate the participant variable. The fact that participants are representative of the population does not alter this situation.

In the foregoing study, one factor used "disorder type" as a factor: anxiety versus depression. The experimenter did not randomly assign participants and then create the anxiety or depression. Participants were selected into the study because they were already either anxious or depressed. If the experimenter had used an experimental operation to induce anxiety or depression, then the factor

"disorder type" would not be a participant variable; it would have been an independent variable.

Suppose one factor is a participant variable and the second factor is manipulated by the experimenter. Only a main effect for the second factor can be interpreted in the language of cause–effect. Now suppose that there is an interaction between a participant variable and an experimental variable. Can we make a cause–effect statement, or is the interaction correlational in nature? We must be very careful. For example, suppose we find that Type A people are more conforming than Type B people, but only when there is a clear payoff for conforming. When there is no identifiable payoff, Type A's are less conforming than Type B people. (Imagine a graph with crossing lines forming an "×.") The payoff versus no payoff is experimentally manipulated; but Type A and B participants are not randomly assigned to conditions. In this situation, we could suggest the payoff situation *causes* differential behavior effects among Type A and Type B participants. The dependent variable change is being explained by the action of the independent variable for two different populations (Type A and Type B personalities). However, we should not conclude that personality type *causes* differential behavior as the payoff situation changes. There are any number of other participant variables correlated with personality type (e.g. competitiveness, hostility, time urgency, dominance) that could explain the relationship. There is simply no way to nail down a causal connection between a participant variable and the dependent variable.

A researcher must always determine the methodological status of each factor in a factorial design. There are also two-way factorial designs in which both factors are participant variables. In this instance, the entire study is correlational. Remember the causal interpretation of research results resides in the *design* of the study, not the type of statistical analysis used to analyze the data.

13.9 How to Present Formally the Conclusions for a Two-Way ANOVA

The proper reporting of two-way ANOVA findings is similar to what is presented in Section 12.11 regarding the reporting of one-way ANOVA findings. Remember to give priority to significant interactions prior to presenting information about main effects. The reader needs to be aware of both the presence of a rejected $F_{A \times B}$ and any follow-up tests that help to interpret this finding prior to being informed of main effect findings.

Many other principles common to the proper reporting of all types of statistical findings were first laid out in Section 8.8. Please consult this portion of the text for more general information about the proper reporting of statistical findings.

Summary

A factorial design combines at least two factors. (These factors can be referred to as "independent variables" if the study is experimental in design.) This arrangement allows an investigator to examine the effect of each factor separately (called main effects) and the joint effect of the factors (called an interaction). Each factor in a factorial design has at least two levels. A two-way factorial design with two levels of each factor is a 2×2 design. If one of the variables has three levels and the other has two levels, it is a 2×3 factorial design and so on. A cell mean is the mean of the scores for a single group or combination of levels of factors. A marginal mean is the mean of the scores from one level of a factor.

When conducting a two-way ANOVA, there are three separate null hypotheses. This means three F ratios will be calculated to test them. The null hypotheses for Factors A and B state there are no differences in the population means for the levels of Factors A and B, respectively. The null hypothesis for the interaction is that there is no interaction.

The structure and logic of the two-way ANOVA is an extension of the one-way ANOVA. Total variance is due to between-group variance plus within-group variance. However, a second stage is introduced. The between-group variance is further partitioned into the variances due to Factor A, Factor B, and the interaction. The second stage of the model yields the F ratios that are used to test null hypotheses: F_A, F_B, and $F_{A \times B}$.

The assumptions of the two-way ANOVA are the same as the assumptions of the one-way ANOVA. Samples should be representative of populations, observations must be independent of each other, interval or ratio data must be used, the data must be normally distributed, and the variances of the populations sampled must be homogeneous. The F test is robust with respect to the last two assumptions.

The F tests in the two-way ANOVA allow us to test the null hypothesis for each factor and the interaction. However, the F values do not provide information about the *size* of each effect. Omega-squared and eta-squared are statistics that quantify the relationship between the variance due to Factor A, Factor B, and the interaction with the overall variance in the study.

If the factorial design has a factor with more than two levels, a significant main effect for that factor does not tell us which levels are driving the effect. In addition, the two-way ANOVA does not interpret a found interaction. Various follow-up comparisons can be used to locate the sources of significance between conditions and between cells. This chapter presents two such tests, Tukey's *HSD* and Fisher's *LSD*.

Finally, when interpreting the results of a two-way ANOVA, consider whether a participant variable is used as a factor. Participant variables maintain the status of correlated variables and cannot be used for a causal interpretation.

Using Microsoft® Excel and SPSS® to Run a Two-Way ANOVA

Excel

General instructions for data entry into Excel can be found in Appendix C.

Data Entry

Choose one of the two factors and enter all of the scores from the samples into adjacent columns (the number of columns equaling the number of conditions for that factor), one sample in each column. Label the columns appropriately. Within the columns, organize the other factor so that all of the data for the first level of this second factor comes first, then all of the data for the second level, and so on. Leave the first column blank so it can be used to label the levels of the second factor (see Figure 13.9 for a visual example).

Data Analysis

1) Excel has built-in programs for many inferential tests, including the two-way ANOVA test. To access it, click on the Data tab on the top menu and then click **Data Analysis**.
2) With the Data Analysis box open, select **Anova: Two Factor with Replication**. (Yes, the title is confusing; this is not a repeated-measures ANOVA.)
3) Input the data range by dragging over the entire data set, including the labels, and placing those coordinates into the **Input Range** box.
4) Determine how many rows there are per condition of the second factor. Excel requires an equal number of rows for each condition of both factors. It also does not allow for empty cells (except the upper left one). Input this value into the **Rows per sample** box. (Our example has four rows per sample.)
5) Decide on an alpha value. The default is .05.
6) Decide on an Output option. The default is to place it on a separate worksheet.
7) Click **OK**.
8) Several output tables are produced. The first tables contain summary data involving the counts, sums, averages, and variances for all levels of the column factor – one table for each level of the row factor. Additionally, there is a table with these descriptive totals across the entirety of the row factor levels. The last table is the ANOVA summary table (labeled "ANOVA"); it looks very similar to the ANOVA summary table described earlier in the chapter (see, for example, Table 13.10). However, it uses the term "sample" for the row factor and "column" for the column factor. There is also an additional column identifying the F_{crit} value for each F in the design. (See Figure 13.9 for a worked example showing evidence of a main effect for the column factor.)

	Behavioral	Group	Control
	4	8	2
	5	6	3
	6	4	4
Anxiety	7	7	1
	3	6	3
	4	7	4
	3	8	2
Depression	2	5	2

Anova: Two-factor with replication

Summary	Behavioral	Group	Control	Total
Count	4	4	4	12
Sum	22	25	10	57
Average	5.5	6.25	2.5	4.75
Variance	1.6 666 667	2.9 166 667	1.6 666 667	4.568 182
Count	4	4	4	12
Sum	12	26	11	49
Average	3	6.5	2.75	4.083 333
Variance	0.6 666 667	1.6 666 667	0.9 166 667	4.083 333

	Total		
Count	8	8	8
Sum	34	51	21
Average	4.25	6.375	2.625
Variance	2.7 857 143	1.9 821 429	1.125

ANOVA

Source of variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	2.6 666 667	1	2.6 666 667	1.684 211	0.210 751 124	4.413 873
Columns	56.583 333	2	28.291 667	17.86 842	5.30888E-05	3.554 557
Interaction	10.083 333	2	5.0 416 667	3.184 211	0.06 5 464 527	3.554 557
Within	28.5	18	1.5 833 333			
Total	97.833 333	23				

Figure 13.9 A worked example using Microsoft Excel to calculate a two-way ANOVA.

SPSS

General instructions for inputting data into SPSS can be found in Appendix C.

Data Entry

In SPSS, each row of the data file represents a participant. Since all samples in a two-way ANOVA test have different participants, all of the dependent variable data from all samples will need to be placed in one column. Within **Variable View**, label this variable appropriately. However, also create a second and third

variable that will allow the user to identify which data goes with which condition of each of the two factors. Label these factor columns appropriately. Then, go to **Data View**. Input the sample data to the appropriate column, and use nominal variables in the factor columns to distinguish between the cells of the factorial design. For example, data in the first level of both factors would get a [1, 1] in the two factor columns; data in the second level of the first factor and the first level of the second factor would get a [2, 1] in the two factor columns. See Figure 13.10 for an example.

	Incidents	disordertype	treatmenttype
1	4	1	1
2	5	1	1
3	6	1	1
4	7	1	1
5	3	2	1
6	4	2	1
7	3	2	1
8	2	2	1
9	8	1	2
10	6	1	2
11	4	1	2
12	7	1	2
13	6	2	2
14	7	2	2
15	8	2	2
16	5	2	2
17	2	1	3
18	3	1	3
19	4	1	3
20	1	1	3
21	3	2	3
22	4	2	3
23	2	2	3
24	2	2	3

Figure 13.10 An example of entered data for a two-way ANOVA in SPSS.

Data Analysis

1) Click **Analyze** on the tool bar, select **General Linear Model**, and then click **Univariate** (univariate here reflects the number of dependent variables, not the number of factors).
2) Highlight the dependent variable column label in the left box, and click the arrow to move it into the **Dependent Variable** box. Move both of the factor variables into the **Fixed Factor(s)** box.
3) If we want to make specific group comparisons at the time of the ANOVA, click on the **Post Hoc** tab to the left, move the factor variables into the **Post Hoc Tests for** box, and make the appropriate selections. If not, simply skip this step.
4) If we want to get basic descriptive statistics, click on **Options** and then **Descriptive**. If not, simply skip this step.

Univariate analysis of variance

Between-subjects factors

		N
disordertype	1	12
	2	12
treatmenttype	1	8
	2	8
	3	8

Tests of between-subjects effects

Dependent variable: incidents

Source	Type III sum of squares	*df*	Mean square	*F*	Sig.
Corrected model	69.333[a]	5	13.867	8.758	.000
Intercept	468.167	1	468.167	295.684	.000
disordertype	2.667	1	2.667	1.684	.211
treatmenttype	56.583	2	28.292	17.868	.000
disordertype * treatmenttype	10.083	2	5.042	3.184	.065
Error	28.500	18	1.583		
Total	566.000	24			
Corrected total	97.833	23			

[a] *R* squared = .709 (Adjusted *R* squared = .628)

Figure 13.11 An output table from a worked example using SPSS to calculate a two-way ANOVA.

5) Click **OK**.
6) The output will generate an expanded ANOVA summary table compared with the one described in the text. It is labeled **Tests of Between-Subjects Effects**. The first row (*Corrected Model*) displays the between-groups row in a one-way ANOVA. Since this variance is split up between the three explored effects, it is dropped from our ANOVA summary table. The second line (*intercept*) can also be dropped. The following three rows reflect the three effects of interest, the two main effects, and the interaction. Once again, we are looking to see if the calculated *F*'s falls in the outermost 5% of the null *F* distribution. If the value found under **Sig**. is .05 or less, we have evidence to reject the null hypothesis. Please note: Use the "Corrected Total" as the total, not what is labeled as the total. See Figure 13.11 for a worked example. In this example, there is evidence for a main effect for "treatment type." The interaction is close to the rejection threshold, but did not reach the necessary p of ≤.05.

Key Formulas

Definitional formula for SS_T

$$SS_T = \Sigma(X - M_G)^2 \quad \text{(Formula 13.1)}$$

Computational formula for SS_T

$$SS_T = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \quad \text{(Formula 13.2)}$$

Definitional formula for SS_{BG}

$$SS_{BG} = \Sigma n_k (M_k - M_G)^2 \quad \text{(Formula 13.3)}$$

Computational formula for SS_{BG}

$$SS_{BG} = \frac{(\Sigma X_{A_1B_1})^2}{n_{A_1B_1}} + \frac{(\Sigma X_{A_1B_2})^2}{n_{A_1B_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left(\frac{(\Sigma X)^2}{N}\right) \quad \text{(Formula 13.4)}$$

Definitional formula for SS_W

$$SS_W = \Sigma(X_{A_1B_1} - M_{A_1B_1})^2 + \Sigma(X_{A_1B_2} - M_{A_1B_2})^2 + \Sigma(X_{A_2B_1} - M_{A_2B_1})^2 + \cdots + \Sigma(X_k - M_k)^2 \quad \text{(Formula 13.5)}$$

Computational formula for SS_W

$$SS_W = \Sigma X^2 - \left[\frac{(\Sigma X_{A_1B_1})^2}{n_{A_1B_1}} + \frac{(\Sigma X_{A_1B_2})^2}{n_{A_1B_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k}\right] \quad \text{(Formula 13.6)}$$

Definitional formula for SS_A

$$SS_A = n_{A_1}(M_{A_1} - M_G)^2 + n_{A_2}(M_{A_2} - M_G)^2 + \cdots + n_k(M_k - M_G)^2 \quad \text{(Formula 13.7)}$$

Computational formula for SS_A

$$SS_A = \frac{(\Sigma X_{A_1})^2}{n_{A_1}} + \frac{(\Sigma X_{A_2})^2}{n_{A_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right] \quad \text{(Formula 13.8)}$$

Definitional formula for SS_B

$$SS_B = n_{B_1}(M_{B_1} - M_G)^2 + n_{B_2}(M_{B_2} - M_G)^2 + \cdots + n_k(M_k - M_G)^2 \quad \text{(Formula 13.9)}$$

Computational formula for SS_B

$$SS_B = \frac{(\Sigma X_{B_1})^2}{n_{B_1}} + \frac{(\Sigma X_{B_2})^2}{n_{B_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right] \quad \text{(Formula 13.10)}$$

Computational formula for $SS_{A \times B}$

$$SS_{A \times B} = n_k[(M_{A_1B_1} - M_{A_1} - M_{B_1} + M_G)^2 + (M_{A_2B_1} - M_{A_2} - M_{B_1} + M_G)^2 + (M_{A_1B_2} - M_{A_1} - M_{B_2} + M_G)^2 + (M_{A_2B_2} - M_{A_2} - M_{B_2} + M_G)^2] \quad \text{(Formula 13.11)}$$

Formulas for omega-squared, ϖ_A^2, ϖ_B^2, $\varpi_{A \times B}^2$

$$\varpi_A^2 = \frac{SS_A - (df_A)MS_W}{SS_T + MS_W} \quad \text{(Formula 13.12)}$$

$$\varpi_B^2 = \frac{SS_B - (df_B)MS_W}{SS_T + MS_W} \quad \text{(Formula 13.13)}$$

$$\varpi_{A \times B}^2 = \frac{SS_{A \times B} - (df_{A \times B})MS_W}{SS_T + MS_W} \quad \text{(Formula 13.14)}$$

Formulas for eta-squared, η_A^2, η_B^2, $\eta_{A \times B}^2$

$$\eta_A^2 = \frac{SS_A}{SS_A + SS_W} \quad \text{(Formula 13.15)}$$

$$\eta_B^2 = \frac{SS_B}{SS_B + SS_W} \quad \text{(Formula 13.16)}$$

$$\eta_{A \times B}^2 = \frac{SS_{A \times B}}{SS_{A \times B} + SS_W} \quad \text{(Formula 13.17)}$$

Formulas for Tukey's *HSD* values, HSD_A, HSD_B, $HSD_{A \times B}$

$$HSD_A = q_A \sqrt{\frac{MS_W}{n_A}} \quad \text{(Formula 13.18)}$$

$$HSD_B = q_B \sqrt{\frac{MS_W}{n_B}} \quad \text{(Formula 13.19)}$$

$$HSD_{A \times B} = q_{A \times B} \sqrt{\frac{MS_W}{n_{A \times B}}} \quad \text{(Formula 13.20)}$$

Formula for Fisher's *LSD* test, two-way ANOVA

$$t = \frac{M_i - M_j}{\sqrt{MS_W \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \quad \text{(Formula 13.21)}$$

Key Terms

Interaction
Factorial (or complex) designs
Cell
Factor
Main effect
Two-way ANOVA
Higher-order Effect

Questions and Exercises

1 Describe the necessary design conditions needed to run a two-way ANOVA.

2 Specify the first and second partitioned stages of the two-way ANOVA.

3 Differentiate between a main effect and an interaction.

4 Use cells to draw the following research designs. Create meaningful labels for the factors. Make some of them independent variables and identify them as such.
a 2×2
b 3×2
c 4×4
d 3×7

5 What is the advantage, when interested in studying two different factors, to placing them in one design versus two?

6 How many types of *F*'s does a two-way ANOVA generate, and how many of each type?

7 What are the null and alternative hypotheses for:
- **a** Factor A
- **b** Factor B
- **c** Interaction

8 Complete these ANOVA summary tables. Test the null hypotheses with $\alpha = .05$.

a

Source	*SS*	*df*	*MS*	*F*	*p*
Factor A	147.00	2	73.50		
Factor B	27.44	2	13.72		
$A \times B$	12.22	4	3.06		
Within groups	95.33	45	2.12		
Total	281.99	53			

b

Source	*SS*	*df*	*MS*	*F*	*p*
Factor A	3.36		3.36		
Factor B	66.67	2			
$A \times B$	56.89				
Within groups					
Total	238.75	35			

c

Source	*SS*	*df*	*MS*	*F*	*p*
Factor A		1	.45		
Factor B	6.05				
$A \times B$	84.05	1			
Within groups		16	1.28		
Total					

9 Draw a graph for each of the following data sets. Place Factor B on the X axis and the means on the Y axis. The two lines on the graphs correspond to the

levels of Factor *A*. For each graph, indicate if there is visual evidence of an $A \times B$ interaction.

a

Factor *B*

Factor *A*			
	35	5	A_1
	55	25	A_2
	B_1	B_2	

b

Factor *B*

Factor *A*			
	30	30	A_1
	10	50	A_2
	B_1	B_2	

c

Factor *B*

Factor *A*			
	25	10	A_1
	10	30	A_2
	B_1	B_2	

10 A researcher conducts a two-factor study with three levels of factor *A* and three levels of factor *B* using a sample size of 10 participants per cell.
- **a** What are the *df* for the resulting F_A value?
- **b** What are the *df* for the resulting F_B value?
- **c** What are the *df* for the resulting $F_{A \times B}$ value?

11 Why does the size of the *F* not necessarily reflect the size of a treatment effect?

12 How do the values of effect size measures (such as ω^2 and η^2) reflect?

13 Why are multiple comparisons not needed to explore main effects in a 2×2 design?

14 Which of the two presented multiple comparison tools is more flexible in terms of different sample sizes within the cells of a study? Why?

15 Which of the two presented multiple comparison (follow-up) tools is the more commonly used test in behavioral and social science research publications?

16 An experimental psychologist is interested in how performance is affected by reinforcement and amount of food deprivation. Performance is measured by the time, in seconds, it takes a rat to run down an alley to a food box. Twenty rats are randomly assigned to four treatment conditions: High Incentive–High Deprivation, High Incentive–Low Deprivation, Low Incentive–High Deprivation, and Low Incentive–Low Deprivation. Deprivation level is manipulated by maintaining one group of rats at 85% of their normal weight and a second group at 95% of their normal weight. Incentive is manipulated by the size of the reward at the end of the alley. In the Low-Incentive condition, a 45-mg food pellet is waiting. In the High-Incentive condition, a 260-mg food pellet is waiting. The raw data for this hypothetical experiment are presented in the following 2 × 2 matrix. Set alpha at .05 and perform a two-way ANOVA.

- **a** Summarize the results in an ANOVA table.
- **b** Provide a graph with incentive on the X axis.
- **c** Calculate ω^2 for any rejected null hypotheses.
- **d** If necessary, run either type of multiple comparisons to aid in interpretation.
- **e** What do these results tell us about the effect of incentive and deprivation on performance?

		Factor *B*: Deprivation (Body Weight)	
		85% (high)	95% (low)
Factor *A*: Incentive	45 mg (low)	7 8 6 7 7	10 7 6 8 6
	260 mg (high)	5 4 4 5 6	9 9 6 7 7

17 An educational psychologist is interested in the effect of delayed feedback on learning and if delayed feedback operates differently as a function of educational level. All participants, composed of freshmen and seniors, are administered a 15-question test; after answering the questions, the participants are given the correct answers at various intervals, depending on which experimental condition they are assigned. All participants are given the same test four days later. The dependent variable is the number of correctly answered questions. Set alpha at .05. For the following data set:

a Provide an ANOVA summary table.
b Graph the results with delayed feedback on the X axis.
c Calculate η^2 for any rejected null hypothesis.
d Conduct multiple comparisons if appropriate.
e Interpret the findings.

		Factor *B*: Feedback Delay		
		No Delay	2-H Delay	1-D Delay
Factor *A*: Education Level	Freshmen	15	7	4
		12	9	6
		13	5	7
		10	8	7
		11	8	7
	Seniors	13	6	8
		15	5	5
		13	6	5
		10	9	6
		10	6	7

18 A clinical psychologist is interested in the effects of cognitive therapy alone, medication alone, and the combined effects of therapy and medication for depression. The following table provides the summary statistics for each cell, the marginal means, and the grand mean.

a Summarize the results in an ANOVA table.
b Calculate η^2 for any rejected null hypothesis.
c Perform whatever secondary tests are deemed necessary to help interpret any findings from the ANOVA.

		Factor *B:* Medication		
		Yes	*No*	
Factor *A:* Cognitive Therapy	Yes	$M_{A_1B_1} = 8.6$ $\Sigma X_{A_1B_1} = 43$ $\Sigma X^2_{A_1B_1} = 375$ $n_{A_1B_1} = 5$	$M_{A_1B_2} = 5.2$ $\Sigma X_{A_1B_2} = 26$ $\Sigma X^2_{A_1B_2} = 238$ $n_{A_1B_2} = 5$	$M_{A_1} = 6.9$
	No	$M_{A_2B_1} = 3.8$ $\Sigma X_{A_2B_1} = 19$ $\Sigma X^2_{A_2B_1} = 75$ $n_{A_2B_1} = 5$	$M_{A_2B_2} = 2.2$ $\Sigma X_{A_2B_2} = 11$ $\Sigma X^2_{A_2B_2} = 27$ $n_{A_2B_2} = 5$	$M_{A_2} = 3.0$
		$M_{B_1} = 6.2$	$M_{B_2} = 3.7$	$M_G = 4.95$

19 Romano and Bordiere (1989) conducted a study to determine if the physical attractiveness of a professor influences students' perceptions of how much they think they will learn from the professor. The design was a 2×2 factorial; one factor is the physical attractiveness of the professor (Attractive/Unattractive), and the other factor is the biological sex of the student (Male/Female). Students provided ratings on a 9-point scale, which reflected how much they thought they would learn, with higher numbers reflecting more learning. Slides of professors were used to obtain the ratings. The following data set is hypothetical, but is constructed so that we arrive at the same results as the investigators.

a Provide an ANOVA summary table.
b Calculate ω^2 for any effects found.
c Interpret the findings.
d Should the researchers be particularly concerned with any assumption violations?

		Factor *B*: Professor	
		Attractive	Unattractive
Factor *A*: Biological Sex of Student	Male	8 6 6 7 5	4 5 7 4 6
	Female	9 7 5 7 7	3 7 4 4 4

20 Suppose we wanted to explore the effects of wearing cologne on interpersonal attraction. Since we are also interested in the potential interaction effects with physical attractiveness, we have chosen to include that variable as well. We select 12 individuals as stimuli: 6 individuals who have been deemed ahead of time by independent raters as "very attractive" and 6 who are deemed "average looking." Within each group of six, three will be wearing cologne and three will not. Interpersonal attraction is measured by the amount of time (in minutes) that unknown students standing in line with our confederates will engage in conversation with them as they wait in line to gain their university ID cards. The average number of minutes each of the 12 participants was conversed with is recorded below. Set $\alpha = .05$ and run a two-way ANOVA.

a Provide an ANOVA summary table.
b Calculate η^2 for any effects found.
c Run any needed post hoc comparisons and interpret the findings.

		Factor *B*: (Physical Attractiveness	
		Very Attractive	Average-Looking
Factor *A:* Cologne	Cologne	7 9 10	8 7 4
	No Cologne	4 5 6	3 6 4

21 A psychologist is interested in whether African American (AA) defendants draw stiffer sentences than Caucasian American (CA) defendants; whether AA judges give stiffer sentences than CA judges; and if there is an interaction between the ethnicity of the judge and the ethnicity of the defendant when it comes to sentencing. A hypothetical data set was constructed. The first table shows the group means and the second table is a partial ANOVA table. Set alpha at .05. (Higher means reflect stiffer sentences.)

a Complete the ANOVA summary table.
b Interpret the findings.

		Factor *B*: Defendant	
		AA	*CA*
Factor *A*: Judge	*AA*	27.33	20.00
	CA	29.50	23.67

Source	*SS*	*df*	*MS*	*F*	*p*
Factor *A* (Ethnicity of Judge)	51.04	1			
Factor *B* (Ethnicity of Defendant)	260.04	1			
$A \times B$	3.38	1			
Within groups					
Total	1396.63	23			

22 The Type A personality is defined in part by a sense of time urgency and a hard-driving, competitive approach in achievement situations. The Type B individual takes a more relaxed approach to achievement-oriented tasks. The Type X personality is a mixture of Type A and Type B characteristics. An organizational psychologist is interested in whether there is an interaction between personality type and an incentive program on sales production. Factor *B* is personality type, and Factor *A* is the manner in which the salesperson is paid: salary or commission. The first table presents the cell means and sample sizes. The second table is a partial ANOVA table. Set alpha at .05.

a Complete the ANOVA table.

b Conduct Fisher's *LSD* tests to locate the sources of significance.

c Interpret the findings.

		Factor *B*		
		Type A	Type B	Type X
Factor *A*	Salary	17.33 $n = 6$	14.83 $n = 6$	12.17 $n = 6$
	Commission	25.0 $n = 6$	17.17 $n = 6$	17.0 $n = 6$

Source	SS	df	MS	F	p
Factor *A* (Incentive)					
Factor *B* (Personality)	288.17	2			
$A \times B$	42.72	2			
Within groups	731.83	30			
Total	1282.75	35			

23 In a *t* test, are we testing for a main effect or an interaction?

24 In a one-way ANOVA, are we testing for a main effect or an interaction?

25 Why should we use caution when interpreting a main effect when there is an interaction?

26 Suppose a constant were added to each score in a 2×3 factorial design. What effect would this have on the main effects and interaction? How would MS_W be affected?

Computer Work

27 A psychologist is interested in the following research questions.

- **a** Can cognitive strategies increase the delay of gratification among children?
- **b** Is there a difference between the biological sexes in the ability to delay gratification?
- **c** Is there an interaction between biological sex and the effectiveness of cognitive strategies?

The experimental task required the child participant to sit in front of a bowl of marshmallows placed on a table. Each child was told, "You can eat as many of the marshmallows as you want, but I would like you to try and wait until I return. If you can't wait, that's OK, but please try. To help you not eat any marshmallows I am going to give you something to think about." In the cognitive transformation condition, participants were taught to imagine that the marshmallows were white, fluffy clouds. In the self-talk condition, participants were told to repeat to themselves, "Don't eat the marshmallows." In the control condition, participants were not provided with any cognitive technique. The experimenter left the room and observed the participant through a one-way mirror,

recording the number of seconds elapsed before the child ate a marshmallow. Use $\alpha = .05$ to test for main effects and an interaction. Provide an ANOVA summary table, calculate η^2 for any found effects, and interpret the findings.

		Factor *B*: Cognitive Strategy		
		Transformation	Self-Talk	Control
Factor *A*: Biological Sex of Children	Males	15 12 13 15 15 30 10 19 32 20 25 29 30 35 19 75 60 25 40 50	14 10 15 16 17 28 11 22 32 19 18 29 29 47 25 74 55 18 45 60	30 12 17 19 22 25 14 43 32 10 18 39 25 45 16 70 50 20 40 62
	Females	65 89 61 45 22 49 50 78 35 53 74 74 75 99 77 64 77 82 55 43	72 90 65 35 18 49 60 75 39 43 74 72 75 89 60 68 85 80 50 48	85 80 60 25 30 59 60 70 25 70 90 75 67 80 80 63 82 75 50 42

28 An experimental psychologist hypothesizes that a High-drive state will increase errors on a mental arithmetic task in comparison with a Low-drive state. Drive state is experimentally manipulated by telling half the participants that performance on the task is related to intelligence (High-drive state). Participants in the Low-drive condition are told that their answers to the problems are to be used as normative data for a future study. The researcher also hypothesizes that drive state will interact with the difficulty of the task. More specifically, participants experiencing high drive will not perform as well when the task is difficult compared with easy. Participants in the Difficult condition receive more complicated problems than those in the Easy condition. The researcher is predicting a main effect for drive and an interaction between task difficulty and level of drive state. The dependent variable is the number of errors made over a long series of mental arithmetic problems. Perform a two-way ANOVA on the following data, with alpha set at .05. We will find that there are main effects for both factors, in addition to a significant interaction. Generate an ANOVA summary table, perform any useful follow-up tests, and interpret the findings.

		Task Difficulty: Factor *B*					
		Easy			Difficult		
Driver state: Factor *A*	High drive	18	12	15	28	20	19
		10	16	18	30	15	15
		19	15	20	35	30	27
		15	12	17	37	37	29
		20	22	17	25	29	30
	Low drive	16	14	15	15	17	18
		12	29	20	20	25	16
		10	27	20	10	16	25
		22	30	25	18	13	11
		20	16	19	19	12	16

29 Kirschner and Karpinski (2010) found evidence that college students who are on Facebook (or have it running in the background) performed more poorly on academic assessments than students who did not. We would like to see if there is evidence that the use of Facebook interferes differently when students are studying different types of material, namely, scientific material, the classics, and the arts. Seventy-two students from a liberal arts college are randomly sampled (36 self-professed "users" of Facebook and 36 who claimed not to use social media when studying), and each one is assigned to one of three academic conditions – the performance in the general education class corresponding to the academic category being assessed by the registrar. For consistency reasons, the registrar's office was asked to classify student academic performance by using a 7-point Likert scale (higher numbers reflecting better performance). Below are some hypothetical data. Generate an ANOVA summary table, perform any useful follow-up tests, and interpret the findings. Additionally, comment on any assumptions we should be concerned about for our analysis.

		Factor *B*: Type of Academic Material								
		Sciences			Classics			Arts		
Factor *A*: Facebook	User	3	7	4	4	6	2	4	5	3
		5	3	5	3	2	5	4	2	3
		6	6	2	5	2	2	5	3	1
		4	4	1	2	3	3	6	5	4
	Nonuser	6	5	2	6	6	4	4	5	5
		4	7	5	2	6	5	5	4	5
		5	3	4	4	4	7	3	5	7
		3	6	6	5	5	6	4	2	6

30 Now we will look at a more traditional form of distracted studying, listening to music. We would like to study the effects of listening to music while studying on academic performance. We would also like to see if the type of material being studied might have an effect. Students are selected from an Introduction to Shakespeare class and a Beginning German class (both 100-level courses). The amount of study time is controlled. After six weeks, all students are given a 50-point test in their respective class. Assume that independent judges have found these tests to be approximately equal in difficulty. Hypothetical data is presented below. Generate an ANOVA summary table, perform any useful follow-up tests, and interpret the findings.

		Factor *B*: Music	
		Music	No Music
Factor *A*: Class	Shakespeare	19 37 14 24	46 43 38 30
	German	25 14 14 39	33 46 35 48

14

Repeated-Measures Analysis of Variance

14.1 The Research Context

Repeated-Measures Designs

In a between-groups design, each participant receives one and only one treatment. This is true whether the design has one or more factors. In a **repeated-measures design** (also called a *within-participants design*), every participant is exposed to *each* of the treatment conditions.[1] Since we can obtain information about the effect of each treatment condition by using the same group of participants, we can eliminate some of the error due to random factors. This makes repeated-measures designs more statistically efficient than between-groups designs, requiring fewer participants to achieve the same statistical power. Chapter 10 used the dependent-samples t test for repeated-measures designs with two experimental conditions. In this chapter, the number of conditions can be greater than two. The appropriate statistical analysis is called a repeated-measures ANOVA.

The main advantage of a repeated-measures design is that there is greater control over participant variables. Chapter 1 explained that participant variables are fixed attributes of a person; that is, fixed at the time the participant enters the experiment. Intelligence, biological sex, psychiatric diagnosis, and personality traits are examples of participant variables. When using a between-groups design, participant variables can present a problem. Suppose a researcher is interested in a teaching technique to enhance learning. If, by chance, the intelligence differences between the participants are not evenly distributed across the conditions, the variance due to "intelligence" will add to the random factors variance and make it harder to find evidence of primary variance. How is this

1 As was the case in previous ANOVA chapters, the repeated-measures ANOVA will be presented using experimental terminology even though this analysis can be used on data gathered correlationally. Most repeated-measures designs, however, are experimental.

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

problem circumvented when using a repeated-measures design? By using the *same* participants in every condition, it is *impossible* for one condition to have more or less of a participant variable than another condition. As a result, the variance due to random factors is reduced.

Suppose we are interested in testing the effectiveness of three different studying strategies on examination performance. In a between-groups design, we would randomly assign, for instance, 60 participants to the three training programs (20 participants per group). In a within-participants design, we would take 20 participants and run each of them through each program (see Table 14.1). Our measure of performance, which is the dependent variable, would be taken after completing each program. In this way, a repeated-measures design reduces random factors variance by eliminating individual differences between conditions and also economically generates 60 data points while using only 20 individuals.

The problems that arise when using a within-participants design have to do with methodology, namely, the effects of previous measurements on future measurements, so-called *carryover* effects. The question becomes, "how sure are we that the number we gain for a participant's second and subsequent measures is only due to the condition they are in and not the fact that they have already been measured?" For instance, some participants might improve the second time around on a task not because of the new treatment condition but because they are now more familiar with the task; this is termed a *practice* effect. Other experimental situations might work the other way around. Participants might perform more poorly merely because they are being measured a second or third time; this is termed a *fatigue* effect. Both of these situations introduce confounding variance into the design and interfere with an interpretation of the findings.

These problems can sometimes be addressed by carefully managing the order through which participants experience the various conditions. If all participants

Table 14.1 A within-participants design: 20 Participants are exposed to each treatment condition.

Treatment I	Treatment II	Treatment III
Participant 1	Participant 1	Participant 1
Participant 2	Participant 2	Participant 2
Participant 3	Participant 3	Participant 3
⋮	⋮	⋮
Participant 20	Participant 20	Participant 20

experience the same order, then there will be no way to determine if group differences are due to the treatment differences or the order of progression through the research design. To avoid **order effects**, researchers often employ a strategy called counterbalancing. **Counterbalancing** involves the presentation of experimental conditions in a different order for different participants. Treatment I would be presented first for some participants, second for some participants, and third for other participants. The same would hold true for Treatments II and III. Confounding variance due to order effects can be controlled if a technique such as counterbalancing can be argued to have equally dispersed order effects across all conditions. If the technique is effective, the unwanted variance simply becomes error variance. There is much to be discussed regarding the methodological problems associated with repeated-measures designs. However, this is not a book on research methodology. We will simply note the inherent methodological challenges with repeated-measures designs and suggest that techniques such as counterbalancing offer potential solutions.[2] The statistical advantage of repeated-measures designs is so great that researchers are often highly motivated to find methodological solutions to the inherent problems of the design. Please consult a research methodology resource for a more elaborate discussion of this important topic.

Not all research questions lend themselves to a within-participants design. For instance, it is very unusual to find a psychotherapy study that uses a repeated-measures design. However, drug treatments are often investigated with a repeated-measures design. Since drugs typically do not leave a lasting effect on behavior and learning, different drugs can be administered to the same participants. Investigators control for carryover effects by allowing the first drug to clear the patient's system before beginning the second drug trial. However, any treatment that produces a relatively permanent change in the measured variable is best evaluated using an independent-samples design.

Examples of Repeated-Measures Designs

► **Example 14.1** An experimental psychologist is interested in whether a participant's level of physiological arousal influences olfactory sensitivity. Participants are asked to detect the presence of unwashed sports socks in an open hamper. The distance between the participant and the hamper is the measure of olfactory sensitivity. The independent variable is arousal. Three levels of arousal are used: low, medium, and high. Low arousal is induced by having participants relax, medium arousal is induced by having participants listen to

2 Whenever a repeated-measures design is used in this chapter, assume the experimental conditions are counterbalanced and the introduction of confounding variance has been avoided.

an annoying sound, and high arousal is created by threatening participants with an electric shock. The same sample of participants is exposed to each arousal condition. Olfactory sensitivity is assessed during each level of physiological arousal. ◄

► **Example 14.2** A marketing psychologist wants to see if people can tell the difference in the smoothness of the ride among three cars. A sample of participants is blindfolded and given a ride in a Lexus, a Cadillac, and a Rolls Royce. Ratings of smoothness are obtained after each ride. ◄

► **Example 14.3** A cognitive psychologist is interested in the effects of caffeine on memory. The same sample of participants is asked to memorize and recall nonsense syllables under three levels of caffeine intake. ◄

The type of repeated-measures design and analysis discussed in this chapter is limited to the case in which only *one* independent variable is used. However, factorial designs that have a repeated-measures factor are common in the behavioral sciences; these are called *mixed designs* (see Box 13.2 for more information). For instance, any study that has a (between-group) treatment factor and a pretest/posttest second factor qualifies as a mixed design (see Box 14.1 for more information). Most advanced statistics textbooks cover the analysis of mixed designs.

14.2 The Logic of the Repeated-Measures ANOVA

The Null Hypothesis

A repeated-measures ANOVA generates one F ratio; it tests for population mean differences among the levels of the independent variable. Therefore, there is only one null hypothesis:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_k$$

The alternative hypothesis is that at least two of the group means come from different populations:

$$H_1: \text{ the null hypothesis is false}$$

Partitioning Variability

The basic logic of the repeated-measures ANOVA is the same as in other ANOVAs. A ratio is created where the numerator contains variance due to a treatment effect (also called "primary variance") as well as random factors and the denominator contains variance due only to random factors. In the

context of a repeated-measures design, there is only one sample of participants; they form different groups only in the sense that they are exposed to different treatment conditions at different times. Figure 14.1 illustrates the sources of variability in the repeated-measures ANOVA.

Between-Group Variability

In the one-way ANOVA, between-group variability is due to three factors: variance due to the treatment, variance due to individual differences, and variance accounted for by experimental error. Individual differences and experimental error are included in the term *error* or *error variance*. The critical difference in the repeated-measures ANOVA is that variance due to individual differences is absent from between-group variation. Any difference among the treatment means cannot be due to individual differences because each condition is composed of the same participants. Therefore,

$$\textit{between-group variance} = \textit{treatment effect} + \textit{experimental error}$$

Within-Group Variability

In a one-way ANOVA, the variability within a treatment condition is due to two factors: individual differences and experimental error. This is also true in a repeated-measures ANOVA. However, the repeated-measures ANOVA allows for the partitioning of the within-group variability into variance due to individual differences and variation due to experimental error. This is possible because individuals are being measured multiple times. Since the variability due to individual differences and experimental error can be separated, the variability due to individual differences can be removed. This means we can create a denominator for the F ratio that only includes experimental error:

$$F = \frac{\textit{treatment effect} + \textit{experimental error}}{\textit{experimental error}}$$

By removing the variance due to individual differences from the denominator, the size of the denominator decreases. As the denominator decreases, the F ratio increases. As the F ratio increases, so does the likelihood that the null hypothesis will be rejected. Consequently, the repeated-measures ANOVA has greater power than the between-groups ANOVA. Recall from Chapter 11 that power is influenced by several factors, one of them being the amount of variation in the data. Repeated-measures designs remove the variance due to individual differences and correspondingly increase the power.

Refer to Figure 14.1. Note that in the second stage of the model, within-group variability has been partitioned. A new term has been introduced: *between-participants variability*. Do not confuse this with *between-group variability*. Between-group variability is the variation among group means and is due to treatment effect plus error. Between-participants variability is the variation in scores due only to individual differences. Since the denominator of F ratios is often referred to simply as *error*, note that the denominator in the repeated-measures ANOVA refers *only* to experimental error.

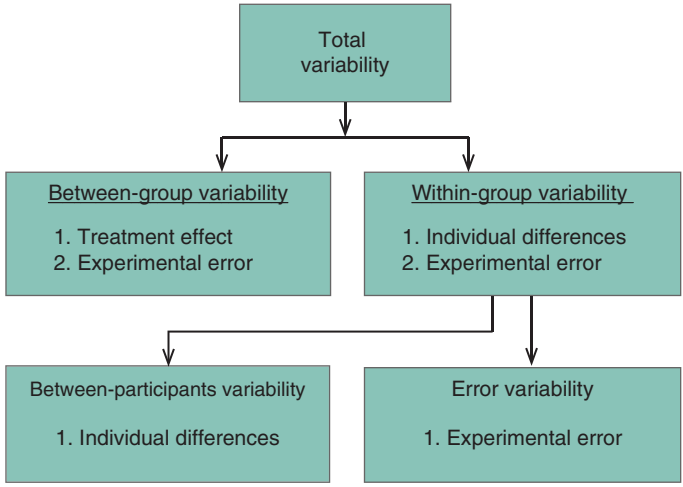


Figure 14.1 Partitioning the total variation in the repeated-measures ANOVA.

Does Removing the Effect Due to Individual Differences Really Matter?

In a one-way ANOVA, the effect due to individual differences shows up in the numerator and the denominator. The numerator also includes the treatment effect and, if large enough, yields an F ratio sufficient to reject the null hypothesis. However, if the variation due to individual differences is removed from both the numerator and the denominator, what is gained? A simple illustration answers this question. Suppose two experiments are contrasted, one using an independent-groups design and another using a repeated-measures design. Further, let us assign some units of variation to each component of the F ratio:

$$\begin{aligned} \textit{treatment effect} &= 200 \textit{ units} \\ \textit{individual differences} &= 300 \textit{ units} \\ \textit{experimental error} &= 50 \textit{ units} \end{aligned}$$

***F* ratio for independent-groups design**

$$F = \frac{treatment + individual\ differences + experimental\ error}{individual\ differences + experimantal\ error}$$

$$F = \frac{200 + 300 + 50}{300 + 50} = \frac{550}{350}$$

$$F = 1.57$$

Now what happens to the *F* ratio if *only* the variation due to individual differences is removed? (Notice, the treatment variance and experimental error values are not altered.)

***F* ratio for repeated-measures design**

$$F = \frac{treatment + experimental\ error}{experimental\ error}$$

$$F = \frac{200 + 50}{50} = \frac{250}{50}$$

$$F = 5.00$$

The result in this example is dramatic. Removing the variability due to individual differences does not always make this big of a difference. The impact of using a repeated-measures design is determined by the relative size of all three forms of variance, especially the variance due to individual differences. If this variance accounts for only a small amount of the total variation, there will only be a small increase in the size of the *F* ratio.

Since larger *F* ratios are, obviously, more likely to direct researchers to reject null hypotheses, a repeated-measures ANOVA has more *power* than an ANOVA conducted using a between-groups design, all other things being equal. If the methodological challenges of a repeated-measures design can be addressed, we should always opt for it over a between-groups design.

14.3 The Formulas for the Repeated-Measures ANOVA

The formulas for the repeated-measures ANOVA are presented first, followed by a worked problem illustrating the computational steps in the repeated-measures ANOVA. For brevity purposes, only computational formulas will be presented.

The Sums of Squares

The Total Sum of Squares, SS_T

The repeated-measures ANOVA begins with the calculation of SS_T. This value is stated in the ANOVA summary table. It can be used as a computational check

and may be needed for follow-up analyses. The SS_T measures the total variability among all the scores in the study. Formula 14.1 is identical to the computational formula for the SS_T in the one-way and two-way ANOVA.

Computational formula for SS_T

$$SS_T = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \quad \text{(Formula 14.1)}$$

where

ΣX^2 = the sum of *all* squared scores
$(\Sigma X)^2$ = the sum of *all* scores, quantity squared
N = the total number of participants

The Sum of Squares Between Groups, SS_{BG}

In the first stage of the partitioning of the total variability, the total variation among all the scores is partitioned into between-group variability and within-group variability (refer to Figure 14.1).

Computational formula for SS_{BG}

$$SS_{BG} = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right] \quad \text{(Formula 14.2)}$$

where

$(\Sigma X_1)^2$, $(\Sigma X_2)^2$, $(\Sigma X_k)^2$ = the sum of the scores in the first experimental condition, the second experimental condition, and so on, quantity squared

n_1, n_2, n_k = the number of scores in the first experimental condition, the second experimental condition, and so on (see Table 14.1)— these should all be equal

The Sum of Squares Within Groups, SS_W

Recall that SS_W refers to variability within the treatment conditions.

Computational formula for SS_W

$$SS_W = \Sigma X^2 - \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k}\right] \quad \text{(Formula 14.3)}$$

Computational Check

Since SS_T is partitioned into SS_{BG} and SS_W, it must be the case that

$$SS_T = SS_{BG} + SS_W$$

The Sum of Squares Between Participants, SS_{BP}

The first stage of the partitioning of the total variation has been completed. What remains is to further partition the within-group variability SS_W. The SS_W is partitioned into SS_{BP} and SS_{error}. Instead of summing scores within columns, we sum scores within rows. Note that a new symbol, P, is introduced.

Computational formula for SS_{BP}

$$SS_{BP} = \frac{(P_1)^2}{k} + \frac{(P_2)^2}{k} + \cdots + \frac{(P_n)^2}{k} - \left[\frac{(\Sigma X)^2}{N}\right] \quad \text{(Formula 14.4)}$$

where

$(\Sigma P_1)^2$, $(\Sigma P_2)^2$, $(\Sigma P_n)^2$ = each *participant's* score in every experimental condition is summed; all the first participant's scores are summed, all the second participant's scores are summed, and so on, until the last participant's scores are summed; each quantity is squared

$(\Sigma X)^2$ = the sum of all the scores in the study, quantity squared

k = the number of experimental conditions

The Sum of Squares Error, SS_{error}

Recall that within-group variability is partitioned into between-participants variation (individual differences) and variation due to error (experimental error). Therefore,

$$SS_W = SS_{BP} + SS_{error}$$

The SS_{error} is found by subtraction.[3] Rearranging the foregoing equation gives

$$SS_{error} = SS_W - SS_{BP}$$

Partitioning the Degrees of Freedom

Partitioning the total degrees of freedom follows the same form as partitioning the total variation of scores. Figure 14.2 illustrates this fact. Table 14.2 lists the various degrees of freedom and their computation.

3 In some textbooks, SS_{error} is referred to as SS_{res} (residual sum of squares).

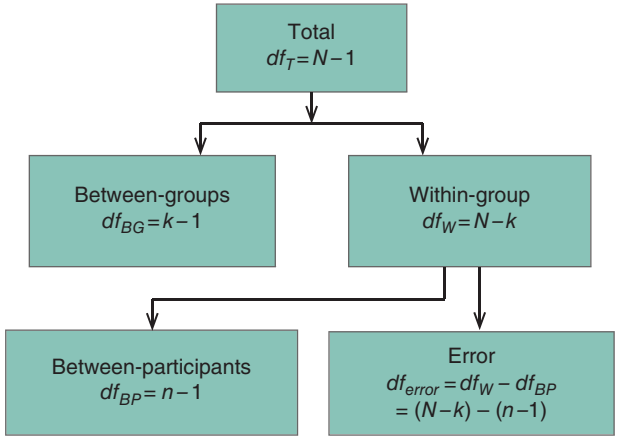


Figure 14.2 Partitioning the degrees of freedom in the repeated-measures ANOVA.

Table 14.2 The degrees of freedom for a repeated-measures ANOVA.

Source	Computation	Degrees of Freedom
Total	$N - 1$	df_T
Between groups	$k - 1$	df_{BG}
Within groups	$N - k$	df_W
Between participants	$n - 1$	df_{BP}
Error	$(N - k) - (n - 1)$	$df_{error} = df_w - df_{BP}$

The Mean Squares and *F* Ratio

The last step in the repeated-measures ANOVA is to calculate the various mean squares used in the *F* ratio. Since the type of repeated-measures ANOVA in this chapter has only one independent variable, there is only one *F* ratio:

$$F = \frac{MS_{BG}}{MS_{error}}$$

where

$$MS_{BG} = \frac{SS_{BG}}{df_{BG}}$$

$$MS_{error} = \frac{SS_{error}}{df_{error}}$$

Remember that the denominator of the F ratio in the repeated-measures ANOVA is MS_{error} rather than MS_W. With the effect of individual differences removed, using MS_{error} preserves the basic structure of the F ratio:

$$F = \frac{treatment + experimental\ error}{experimental\ error}$$

Worked Problem

Clinical psychologists have noted that anxious people sometimes have difficulty concentrating. Suppose a researcher is interested in the effects of a drug, administered at different dosages, on cognitive performance. Ten anxious participants are selected and exposed to four treatment conditions. In one treatment condition, the participants are administered 2.5 mg of Valium. A second and third treatment condition involves the administration of 5 and 10 mg of Valium. Another condition is added as a control; here the participants receive a placebo. Participants are asked to solve a series of mental arithmetic problems, with the dependent variable being the number of errors committed. The treatment conditions are presented one week apart to assure that the drug has cleared from the participants' systems before the effect of a new dosage is assessed. In addition, the order in which the treatments are delivered is counterbalanced among the participants to remove order as a confounding variable. Table 14.3 presents the design of this experiment. In the body of the table, each

Table 14.3 The repeated-measures design for the hypothetical worked problem.

	Treatment			
	Placebo	2.5 mg	5 mg	10 mg
Participants	*(1)*	*(2)*	*(3)*	*(4)*
P_1	X_{11}	X_{12}	X_{13}	X_{14}
P_2	X_{21}	X_{22}	X_{23}	X_{24}
P_3	X_{31}	X_{32}	X_{33}	X_{34}
P_4	X_{41}	X_{42}	X_{43}	X_{44}
P_5	X_{51}	X_{52}	X_{53}	X_{54}
P_6	X_{61}	X_{62}	X_{63}	X_{64}
P_7	X_{71}	X_{72}	X_{73}	X_{74}
P_8	X_{81}	X_{82}	X_{83}	X_{84}
P_9	X_{91}	X_{92}	X_{93}	X_{94}
P_{10}	X_{101}	X_{102}	X_{103}	X_{104}

X symbolizes a score. The first subscript of X identifies the participant; the second subscript identifies the experimental condition. For example, X_{23} refers to the obtained score for participant number 2 under experimental condition number 3. Note that although the total number of participants is 10, the total number of scores is 40. This point will become important when we analyze degrees of freedom. The data for this study are presented in Table 14.4.

Table 14.4 Raw data for the worked problem.

	Treatment				
Participants	**Placebo**	**2.5 mg**	**5 mg**	**10 mg**	
P_1	14	5	9	15	$\Sigma P_1 = 43^a$
P_2	12	3	6	11	$\Sigma P_2 = 32$
P_3	10	2	7	9	$\Sigma P_3 = 28$
P_4	7	4	5	7	$\Sigma P_4 = 23$
P_5	9	4	6	8	$\Sigma P_5 = 27$
P_6	9	1	3	7	$\Sigma P_6 = 20$
P_7	10	3	4	9	$\Sigma P_7 = 26$
P_8	5	0	0	13	$\Sigma P_8 = 18$
P_9	6	4	4	6	$\Sigma P_9 = 20$
P_{10}	8	6	6	7	$\Sigma P_{10} = 27$
	$\Sigma_1 = 90$	$\Sigma X_2 = 32$	$\Sigma X_3 = 50$	$\Sigma X_4 = 92$	$\Sigma X = 264$
	$M_1 = 9.0$	$M_2 = 3.2$	$M_3 = 5.0$	$M_4 = 9.2$	

[a] P is the symbol used for a participant in a repeated-measures design.

Summary of the Computational Steps

Compute:

1) SS_T
2) SS_{BG}
3) SS_W
4) SS_{BP}
5) SS_{error}
6) Compute the df for each of the foregoing SS
7) MS_{BG}
8) MS_{error}
9) F ratio

Calculating the Sums of Squares

Step 1. Compute the total sum of squares, SS_T.

$$SS_T = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

Therefore,

$$SS_T = 2236 - \frac{(264)^2}{40}$$

$$SS_T = 2236 - 1742.40$$

$$SS_T = \mathbf{493.60}$$

Step 2. Compute the sum of squares between groups, SS_{BG}.

$$SS_{BG} = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right]$$

Placing the numbers in the formula,

$$SS_{BG} = \frac{(90)^2}{10} + \frac{(32)^2}{10} + \frac{(50)^2}{10} + \frac{(92)^2}{10} - \left[\frac{(264)^2}{40}\right]$$

$$SS_{BG} = 810 + 102.40 + 250 + 846.40 - 1742.40$$

$$SS_{BG} = 2008.80 - 1742.40$$

$$SS_{BG} = \mathbf{266.40}$$

Step 3. Compute the sum of squares within groups, SS_W.

We are about to complete the first stage of the partitioning of the total variability. Recall that the SS_W refers to variability within the treatment conditions.

The formula for SS_W is

$$SS_W = \Sigma X^2 - \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k}\right]$$

For the worked problem,

$$SS_W = 2236 - \left[\frac{(90)^2}{10} + \frac{(32)^2}{10} + \frac{(50)^2}{10} + \frac{(92)^2}{10}\right]$$

$$SS_W = 2236 - 810 + 102.40 + 250 + 846.40$$

$$SS_W = 2236 - 2008.80$$

$$SS_W = \mathbf{227.20}$$

Step 4. Perform a computational check.

Since SS_T is partitioned into SS_{BG} and SS_W,

$$SS_T = SS_{BG} + SS_W$$

Therefore,

$$493.60 = 266.40 + 227.20$$

Step 5. Compute the sum of squares between participants, SS_{BP}.

$$SS_{BP} = \frac{(P_1)^2}{k} + \frac{(P_2)^2}{k} + \cdots + \frac{(P_n)^2}{k} - \left[\frac{(\Sigma X)^2}{N}\right]$$

$$SS_{BP} = \frac{(43)^2}{4} + \frac{(32)^2}{4} + \frac{(28)^2}{4} + \frac{(23)^2}{4} + \frac{(27)^2}{4} + \frac{(20)^2}{4} + \frac{(26)^2}{4} + \frac{(18)^2}{4} + \frac{(20)^2}{4} + \frac{(27)^2}{4} - \left[\frac{(264)^2}{40}\right]$$

$$SS_{BP} = 1861 - 1742.40$$

$$SS_{BP} = \mathbf{118.60}$$

Step 6. Compute the sum of squares error, SS_{error}.

Within-group variability is partitioned into between-participants variation (individual differences) and variation due to error (experimental error). Therefore,

$$SS_W = SS_{BP} + SS_{error}$$

As noted previously, the SS_{error} is found by subtraction.

Rearranging the terms,

$$SS_{error} = SS_W - SS_{BP}$$

From the data,

$$SS_{error} = 227.20 - 118.60 = \mathbf{108.60}$$

We have now completed stage two of the partitioning of the total variation of scores. Here is a summary of the calculations:

$$SS_T = \mathbf{493.60}$$

$$SS_{BG} = \mathbf{266.40}$$

$$SS_W = \mathbf{227.20}$$

$$SS_{BP} = \mathbf{118.60}$$

$$SS_{error} = \mathbf{108.60}$$

To arrive at the F ratio, we need to compute MS_{BG} and MS_{error}.

Step 7. Compute MS_{BG}. The $df_{BG} = k - 1 = 4 - 1 = 3$.

$$MS_{BG} = \frac{SS_{BG}}{df_{BG}} = \frac{266.40}{3} = \mathbf{88.80}$$

Step 8. Compute MS_{error}. The $df_{error} = df_W - df_{BP} = 36 - 9 = 27$.

$$MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{108.60}{27} = \mathbf{4.02}$$

Step 9. Compute the F ratio. The F ratio is

$$F = \frac{MS_{BG}}{MS_{error}} = \frac{88.80}{4.02} = \mathbf{22.09}$$

The remaining steps of the repeated-measures ANOVA use the F ratio to test the null hypothesis.

14.4 Using the *F* Ratio to Test the Null Hypothesis

The null hypothesis for the worked problem is $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$. The alternative hypothesis is that H_0 is false. If the null hypothesis were rejected, we would conclude that at least two of the sample means come from different populations. To reject the null hypothesis, the obtained F ratio must be larger than the critical value of F found in the F table (Table A.5). The degrees of freedom for F are the degrees of freedom associated with the numerator and denominator of the F ratio, df_{BG} and df_{error}, respectively. The degrees of freedom for this study are $df_{BG} = 3$ and $df_{error} = 27$. Entering the F table, the critical value when $\alpha = .05$ is 2.96. The critical value when $\alpha = .01$ is 4.60. The obtained F value is 22.09. At either alpha value, we have statistical evidence to reject the null hypothesis.

14.5 Interpreting the Findings

What does a rejected null hypothesis mean for the study? The independent variable has four levels: placebo, 2.5, 5, and 10 mg of Valium. A value of F that exceeds F_{crit} means we have statistical evidence suggesting that at least two of the group means differ on the dependent variable. In other words, the number of errors committed during the mental arithmetic test seems to be affected

by dosage level. At this point we have the same dilemma as when we reject an F ratio in a one-way ANOVA. There is no way to know which groups are the ones that are differing from one another. To locate the differences, pairwise comparisons among the means are required. This topic will be discussed shortly. We turn now to the form of the summary table in the repeated-measures ANOVA.

14.6 The ANOVA Summary Table

The general form of the summary table for a repeated-measures ANOVA is shown in Table 14.5. The presentation of the findings is shown in Table 14.6. This is not the only way researchers can choose to create a summary table for repeated-measures ANOVAs, however; it is what we will be using in this text.

Table 14.5 General form of the summary table for a repeated-measures ANOVA.

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	SS_{BG}	df_{BG}	MS_{BG}	F	
Within groups	SS_W	df_W			
Between participants	SS_{BP}	df_{BP}			
Error	SS_{error}	df_{error}	MS_{error}		
Total	SS_T	df_T			

Table 14.6 The ANOVA summary table for the hypothetical study.

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups (dose level)	266.40	3	88.80	22.09	$p < .01$
Within groups	227.20	36			
Between participants	118.60	9			
Error	108.60	27	4.02		
Total	493.60	39			

Box 14.1 Next Steps for Repeated-Measures ANOVAs: Mixed Designs and Quasi-Experimentation

We have already learned that the term "mixed design" refers to studies with multiple factors where at least one factor is between groups and at least one is repeated measures (within participants). Sometimes we will find the longer term "mixed between–within-participants design" used, or when speaking of the analysis of this design, the term "split-plot ANOVA" (or SPANOVA). What has not been talked about previously is the term "quasi-experimentation." A quasi-experiment is a study design that takes on some of the features of an experiment, but not all of them. There are many versions of quasi-experiments, and discussions of them can be found in most research methodology resources (e.g. Cook & Campbell, 1979 and Shadish, Cook, & Campbell, 2002 are widely regarded as excellent resources).

Of particular interest here is a frequently used version of quasi-experimentation where order effects are purposefully not counterbalanced; in fact, the order is of particular importance. The best example of this would be multiple-group pretest/posttest designs (of course these designs are not limited to two measurements; oftentimes participants are measured multiple times after an initial premeasure). In these designs, participants are assigned to different groups where a treatment is administered after a premeasure but before a postmeasure (or an ongoing treatment occurs and various postmeasures are taken). The question to be answered is, "which treatment type creates the most change over time?" For example, suppose a claim is made that a certain type of treatment works initially but fades in effectiveness as time passes, while a different treatment has a less dramatic initial effect, but creates change that is more lasting. A design could be set up to test this claim where participants are assigned to one of two treatment conditions (between-groups factor) and measured at three different times: premeasure, postmeasure 1, and postmeasure 2. This would be quasi-experimental because there may be threats to internal validity that come with the passage of time. Changes in participants' scores between the measures may be due to the effect of the treatment but may also be due to other factors, some of which may be varying somewhat systematically with the treatment conditions. (For example, imagine that one group of participants must travel to point A to get to their treatment site and road construction has made the commute much longer; participants in the other group who are traveling to point B for treatment incur no such added frustration.) Quasi-experimentation designs are susceptible to these types of problems. Nonetheless, they are popular and very helpful for tracking change across time.

The 2×3 quasi-experimental design described above would be analyzed using a hybrid of the ANOVA presented in Chapter 13 (two-way ANOVA) and the ANOVA presented in this chapter (repeated-measures ANOVA). The denominator reflecting the error of random factors would be calculated more than once, and different versions would be used for different F's depending on whether the numerator is a between-groups factor, a within-groups factor, or the interaction.

These types of designs are well represented in the behavioral and social science literature and statistical software programs such as SPSS can be used to analyze them.

14.7 Assumptions of the Repeated-Measures ANOVA

The assumptions for the repeated-measures ANOVA are the same as those for the one-way and two-way ANOVA, with the exception of the second and last assumption:

1) The samples are representative of the populations from which they come.
2) Observations within each condition are independent of one another.
3) Gathered data comes from an interval or ratio scale.
4) The populations from which the data come are normally distributed.
5) The variances of the population of difference scores are homogeneous.

The difference in the second assumption reflects the fact that the same participants are used for each condition. This qualification to the assumption of independence was also made for the dependent-samples t test.

The fifth assumption requires a bit more explanation. We may find it described elsewhere as the *assumption of sphericity*. Since each participant provides a score within each level of the independent variable, it is possible to arrange the data as pairs of scores for any two treatments. In a study with four treatment conditions, there would be six pairs of scores (Treatments 1 and 2; 1 and 3; 1 and 4; 2 and 3; 2 and 4; 3 and 4). By subtracting the two scores for each of the participants in a given pair, a variance can be calculated on the difference scores. With four treatment conditions, we would calculate six variances. It is assumed that these variances, at the population level, are roughly equal. It is possible to test this assumption and make corrections in the ANOVA if this assumption is violated (see Keppel and Wickens, 2004).

Just as with the previous ANOVAs, as n_p increases the F test is robust to minor violations of the last two assumptions and can be performed even when they are not *strictly* met. However, gross violations of these assumptions require the use of statistical tests (nonparametric tests), which do not require the populations to be normally distributed with equal variances.

14.8 Measuring Effect Size for Repeated-Measures ANOVA

In the chapters covering the one-way and two-way ANOVA, it was emphasized that the size of the F ratio does not indicate the degree to which the levels of the independent variable influence the dependent variable. The strength of association between the independent and dependent variables, also called the effect size, can be determined with the omega-squared, ω^2, or eta-squared, η^2, statistic.

Repeated-measures omega-squared, ω^2

$$\varpi^2 = \frac{SS_{BG} - df_{BG}(MS_{error})}{SS_T + MS_{error}} \quad \text{(Formula 14.5)}$$

Using the data from the hypothetical worked problem,

$$\varpi^2 = \frac{266.40 - 3(4.02)}{493.60 + 4.02}$$

$$\varpi^2 = \frac{254.34}{497.62}$$

$$\varpi^2 = \mathbf{0.51}$$

This statistic estimates 51% of the dependent variable variation is due to dosage level.

Now we will look at how eta-squared measures the effect size.

Repeated-measures eta-squared, η^2

$$\eta^2 = \frac{SS_{BG}}{SS_T} \quad \text{(Formula 14.6)}$$

Using the data from the hypothetical worked problem,

$$\eta^2 = \frac{266.40}{493.60}$$

$$\eta^2 = \mathbf{0.54}$$

Here again, we see that η^2 seems to overestimate the effect size relative to ω^2.

14.9 Locating the Source(s) of Statistical Evidence

A rejected null hypothesis in a repeated-measures design that has more than two levels of the independent variable requires follow-up comparisons to locate the source(s) of the statistical evidence. To remain consistent with the preceding two chapters, Tukey's *HSD* and Fisher's *LSD* or protected *t* test are used. The formula for Tukey's *HSD* is similar to previous versions. Be sure to see that MS_{error} is now used as the error term and df_{error} is needed to find q. Here is the formula.

Formula for Tukey's *HSD*, repeated measures

$$HSD = q\sqrt{\frac{MS_{error}}{n}} \quad \text{(Formula 14.7)}$$

where

q = the Studentized range statistic (Table A.6)
n = the number of scores in each group

In the worked problem, there are four levels to the independent variable, but only one *HSD* value is needed to make all comparisons[4]:

$$HSD = q\sqrt{\frac{MS_{error}}{n}}$$

$$HSD = 3.87\sqrt{\frac{4.02}{10}}$$

$$HSD = \mathbf{2.45}$$

Applying this value to the six different comparisons, we find statistical evidence suggesting participants performed better under the 2.5 and 5 mg conditions compared with the 10 mg and placebo conditions.

Fisher's *LSD* test is also very similar to versions presented in previous chapters. However, MS_{error} instead of MS_W serves as the error term.

Formula for Fisher's *LSD*, repeated measures

$$t = \frac{M_i - M_j}{\sqrt{MS_{error}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \quad \text{(Formula 14.8)}$$

where
$M_i,\ M_j$ = the means for the two groups being compared
$n_i,\ n_j$ = the number of scores in each of the two groups being compared

The critical value is found in the *t* table (Table A.2). The *df* used to find the critical value is taken from the MS_{error} term ($df_{error} = dfw - df_{BS}$).

In the worked problem, there are four levels to the independent variable. To make all possible pairwise comparisons require six *t* tests. Only one comparison is made to illustrate how the formula works.[5]

Placebo versus 5 mg

$$t = \frac{M_i - M_j}{\sqrt{MS_{error}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

4 To find q, an online table was used. The one found in Table A.6 is incomplete.

5 The protected *t* test is presented in this worked problem to maintain consistency with Chapters 12 and 13. However, when using Fisher's *LSD* for more than three pairwise comparisons, the probability of a Type I error increases.

$$t = \frac{9 - 5}{\sqrt{4.02\left(\frac{1}{10} + \frac{1}{10}\right)}}$$

$$t = \frac{4}{\sqrt{0.804}} = \frac{4}{0.90}$$

$$t = \mathbf{4.44}$$

The critical value in the *t* table that corresponds to a *df* of 27 is ±2.05. Since the obtained value of 4.44 falls outside of the critical values, reject the null hypothesis that these two means come from the same population. We would interpret this statistically significant difference by saying that statistical evidence has been found suggesting the mental concentration of anxious participants is improved when they are administered 5 mg of Valium compared with a pill that they *think* will help their concentration (placebo).

Box 14.2 presents a study that examines the influence of physiological arousal on olfactory sensitivity. The investigator uses a repeated-measures design by exposing the same group of participants to three arousal conditions.

Box 14.2 The Inverted U Relationship Between Arousal and Task Performance

Most students in an introductory psychology class learn of the "inverted U" relationship (∩) between arousal and task performance. Optimal performance in problem solving is observed during a moderate level of arousal, with poorer performance found during low and high levels of arousal (e.g. Obrist, 1962). The "inverted U" relationship is an example of a *curvilinear* relationship because a graph of the relationship describes a curved line. Another example of a curvilinear relationship would be if the line were not drawn as ∩, but rather, simply as a ∪.

Halpin (1978) wondered if perhaps increasing levels of physiological arousal would produce a curvilinear relationship with olfactory sensitivity. In other words, are people better able to detect a smell under a moderate level of arousal than under a low or high level of physiological arousal? To answer this question, Halpin used a repeated-measures design.

Experimental Procedure

Thirty-six university students were exposed to each of three experimental conditions: low, medium, and high arousal. Arousal was manipulated in the following manner. In the Low-arousal condition, participants listened to a relaxation tape. In the Medium-arousal condition, participants were exposed to loud, continuous white noise (white noise sounds like static). In the High-arousal condition, participants heard intermittent bursts of loud, white noise and were led to expect periodic electric shocks.

During each experimental condition, participants were presented with an odorant of 1-propanol in distilled water. Several presentations of the odorant

were administered in each experimental condition. With each successive presentation, the strength of the solution was increased. The participants were asked to indicate at what point they could smell the substance. In this way, for each participant, under each experimental condition, an olfactory sensitivity threshold was identified and served as the dependent variable. If a curvilinear relationship holds for arousal and olfactory sensitivity, then the lowest thresholds should be observed when participants are experiencing a medium level of arousal. Higher and similar thresholds should be found during low and high levels of arousal.

The dependent variable was the percentage of concentration of 1-propanol present when the participants signaled that they detected a smell. The following table presents the mean thresholds, in percentage of concentration, for each of the experimental conditions. The standard deviations are also shown. We may want to draw a graph of the relationship between arousal and olfactory sensitivity to depict the curvilinear relationship. The author conducted a repeated-measures ANOVA, and a significant difference among the conditions was found, $F(2, 70) = 8.80$, $p < .05$. To locate the source(s) of the statistical evidence, multiple comparisons were conducted among the three means. The mean for the Medium-arousal condition was significantly different from the means of both the Low- and High-arousal conditions. No difference between the Low- and High-arousal conditions was found. These findings allow Halpin to extend the generality of the arousal-performance curvilinear phenomenon to olfactory perception.

Arousal

	Low	Medium	High
M	.66	.21	.60
s	.008	.01	.07

14.10 How to Present Formally the Conclusions for a Repeated-Measures ANOVA

The proper reporting of repeated-measures ANOVA findings is similar to what is presented in Section 12.11, regarding the reporting of one-way ANOVA findings. When reporting a significant F, we must include the df_{BG} and df_{error}, the F value, and the alpha level used to make our decision. For instance, "Statistical evidence suggests the type of therapy administered influenced recovery, $F(2, 14) = 13.18$, $p < .05$. Further analysis found evidence suggesting Therapy C performed better than Therapy A, $t(14) = 4.71$, $p < .05$; and Therapy B, $t(14) = 3.15$, $p < .05$. No difference was found between Therapy A and Therapy B, $t(14) = 2.21$, *n.s.*" A failure to reject might read, "There was no statistical evidence to suggest the

type of therapy used influenced recovery, $F(2, 14) = 1.95$, *n.s.*" Measures of effect size can be added at the end of the sentence when appropriate.

Many other principles common to the proper reporting of all types of statistical findings were first presented in Section 8.8. Please consult this portion of the text for more general information about the proper reporting of statistical findings.

Summary

In a repeated-measures design, also called a within-participants design, every participant is exposed to each of the treatment conditions. Since we can obtain information about the effect of each treatment condition by using the same group of participants, this type of design requires fewer participants than a between-groups design and, as a result, is more statistically efficient. The efficiency occurs because a repeated-measures design eliminates individual differences as a potential explanation for the results of the study. By using the same participants in all conditions, this variance is logically eliminated from the numerator of the F ratio and can be mathematically partitioned out of the denominator. However, not all research questions lend themselves to a within-participants design, and many methodological problems may be introduced that must be addressed.

When conducting a repeated-measures ANOVA, one F ratio is produced. It tests the null hypothesis associated with the levels of the independent variable.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_k$$

The alternative hypothesis is that at least two of the group means come from different populations.

$$H_1: \text{the null hypothesis is false}$$

The total variation of scores can be partitioned into between-group variability and within-group variability. Between-group variability is due to treatment effect and experimental error. Within-group variability is partitioned into variability due to individual differences and variability due to experimental error, with the variability due to individual differences removed. The repeated-measures F ratio has the form

$$F = \frac{\textit{treatment effect} + \textit{experimental error}}{\textit{experimental error}}$$

The assumptions of the repeated-measures ANOVA are:

1) The samples are representative of the populations from which they come.
2) Observations within each condition are independent of one another.
3) Gathered data comes from an interval or ratio scale.

4) The populations from which the data come are normally distributed.
5) The variances of the population of difference scores are homogeneous.

An *F* that leads to rejecting the null hypothesis does not provide direct information about the size of the effect between the independent and dependent variables. Omega-squared and eta-squared are statistics that estimate the effect size. Furthermore, a rejected null hypothesis requires follow-up tests where pairwise comparisons can be made among the cell means to locate the source(s) of statistical evidence. Tukey's *HSD* and Fisher's *LSD* are two of the procedures that can accomplish that task.

Using Microsoft® Excel and SPSS® to Run a Repeated-Measures ANOVA

Excel

General instructions for data entry into Excel can be found in Appendix C.

Data Entry

Enter all of the scores into adjacent columns (one column per condition), with each row assigned to a participant. For example, participant 1 will have all of their data entered on the same row, then participant 2's data, and so on. Label the columns appropriately. (See Figure 14.3 for an example.)

Participant	TherapyA	TherapyB	TherapyC
1	12	15	16
2	14	17	18
3	14	14	16
4	12	14	16
5	8	14	20
6	9	16	17
7	13	12	21
8	12	11	15

Figure 14.3 An example of an encoded data set in preparation for running a repeated-measures ANOVA in Excel.

Data Analysis

1) Excel has built-in programs for many inferential tests, including the repeated-measures ANOVA test. To access it, click on the Data tab on the top menu and then click **Data Analysis**. (Some versions of Excel have a "Tools" tab. The Data Analysis function may be under this tab.) If this option is not found, the Data Analysis ToolPak needs to be installed. See Excel instruction materials for how to install this feature.
2) With the Data Analysis box open, select **Anova: Two-Factor Without Replication**. (Yes, this is a confusing title for a repeated-measures ANOVA.)
3) Input the data range by dragging over the entire data set and placing those coordinates into the **Input Range** box. (If we included the labels in the data range, make sure to click the **Labels** box to exclude those cells.)

4) Decide on an Output option. The default is to place it on a separate worksheet.
5) Click **OK**.
6) The first output box will present summary data, the count, sum of all values, means (average), and variance for all conditions. The second output box will be an ANOVA summary table (labeled "ANOVA") similar but not identical to the one found in this chapter. The *F* of importance (between groups) is associated with the "Columns" row. The row labeled "Rows" is the "Between-participants" row. (Ignore the *F* on this row.) The ANOVA table presented in this chapter can be fully constructed once this is realized (simply add data from the "Rows" row and the "Error" row to determine the missing "Within-groups" line.) Note also the addition of an F_{crit} value. (See Figure 14.4 for a worked example.)

Anova: two-factor without replication

Summary	*Count*	*Sum*	*Average*	*Variance*
1	3	43	14.33333	4.333333
2	3	49	16.33333	4.333333
3	3	44	14.66667	1.333333
4	3	42	14	4
5	3	42	14	36
6	3	42	14	19
7	3	46	15.33333	24.33333
8	3	38	12.66667	4.333333
TherapyA	8	94	11.75	4.785714
TherapyB	8	113	14.125	3.839286
TherapyC	8	139	17.375	4.553571

ANOVA

Source of variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	24.5	7	3.5	0.723247	0.655321	2.764199
Columns	127.5833	2	63.79167	13.18204	0.000604	3.738892
Error	67.75	14	4.839286			
Total	219.8333	23				

Figure 14.4 A worked example using Microsoft Excel to calculate a repeated-measures ANOVA.

SPSS

General instructions for inputting data into SPSS can be found in Appendix C.

Data Entry

In SPSS, each row of the data file represents a participant. Since all samples in a repeated-measures ANOVA have the same participants, multiple columns will be needed to house the data. Within **Variable View**, create a series of variables corresponding to the various conditions in the study. Then, go to **Data View**, and input the data, be careful to keep data from each participant within a given row. (See Figure 14.5 for an example.)

	TherapyA	TherapyB	TherapyC
1	12	15	16
2	14	17	18
3	14	14	16
4	12	14	16
5	8	14	20
6	9	16	17
7	13	12	21
8	12	11	15

Figure 14.5 An example of entered data for a repeated-measures ANOVA in SPSS.

Data Analysis

1) Click **Analyze** on the tool bar, select **General Linear Model**, and then click **Repeated Measures**.
2) In the **Within-Subject Factor Name** box, assign a label for the independent variable (perhaps "TherapyType" for this example). In the **Number of Levels** box, type the number of conditions present (in this example, 3), and click **Add**.
3) Once this is done, the **Define** box at the bottom will become active. Click it.
4) Using the arrow button, move the names of the conditions we wish to investigate into the **Within-Subjects Variable** box.
5) If we want to run post hoc (follow-up) tests, select the **Options** box. Use the arrow button to move the independent variable label (in our example, TherapyType) into the **Display Means for** box. Then click **Compare Main Effects** and select the test of choice. (Tukey's HSD is not available here.) Then click **Continue**.
6) If we want to have the descriptive statistics associated with each condition and/or effect size, click on **Options** and then **Descriptive** and/or **Estimates of effect size**, as the case may be, and then **Continue**.
7) The output will generate multiple tables and more if we asked for descriptives, estimates of effect size, or any post hoc tests. A repeated-measures ANOVA summary like the one presented earlier in the text can be constructed from the available information. Find the **Tests of Within-Subjects**

Effects box. Assuming sphericity, we find the Sum-of-Squares, *df*, Mean Square, *F*, and probability of *F* for both the Between-groups row and the Error row. To see that the *F* is correct, divide the mean square of the independent variable (TherapyType) by the Mean Square Error, and we will find the *F* value presented. (The other rows in this box reflect more sophisticated analyses designed to adjust for the violation of some assumptions. This material goes beyond the scope of our text.) To complete the ANOVA summary table as presented in the chapter, we need to also find the Between-participants line. Go to the **Tests of Between-Subjects Effects** box (usually the very last box unless post hoc tests are run). The row labeled "Error" is what we need. Now that we have both the Error row (from the **Tests of Within-Subjects Effects** box) and Between-participants row (from the **Tests of Between-Subjects Effects** box), we are able to construct fully

General linear model

Tests of within-subjects effects

Measure: MEASURE_1

Source		Type III sum of squares	*df*	Mean square	*F*	Sig.
factor1	Sphericity assumed	127.583	2	63.792	13.182	.001
	Greenhouse-Geisser	127.583	1.935	65.924	13.182	.001
	Huynh-Feldt	127.583	2.000	63.792	13.182	.001
	Lower bound	127.583	1.000	127.583	13.182	.008
Error(factor1)	Sphericity assumed	67.750	14	4.839		
	Greenhouse-Geisser	67.750	13.547	5.001		
	Huynh-Feldt	67.750	14.000	4.839		
	Lower bound	67.750	7.000	9.679		

Tests of between-subjects effects

Measure: MEASURE_1

Transformed variable: Average

Source	Type III sum of squares	*df*	Mean square	*F*	Sig.
Intercept	4988.167	1	4988.167	1425.190	.000
Error	24.500	7	3.500		

Figure 14.6 An output table from a worked example using SPSS to calculate a repeated-measures ANOVA.

the table. Recall that SS_W can be found by adding SS_{BP} and SS_{error} and df_W can be found by adding df_{BP} and df_{error}. Finally, SS_T can be found by adding SS_{BG} and SS_W, and df_T can be found by adding df_{BG} and df_W. (See Figure 14.6 for a worked example.)

Key Formulas

Computational formula for SS_T

$$SS_T = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \quad \text{(Formula 14.1)}$$

Computational formula for SS_{BG}

$$SS_{BG} = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right] \quad \text{(Formula 14.2)}$$

Computational formula for SS_W

$$SS_W = \Sigma X^2 - \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k}\right] \quad \text{(Formula 14.3)}$$

Computational formula for SS_{BP}

$$SS_{BP} = \frac{(P_1)^2}{k} + \frac{(P_2)^2}{k} + \cdots + \frac{(P_n)^2}{k} - \left[\frac{(\Sigma X)^2}{N}\right] \quad \text{(Formula 14.4)}$$

Repeated-measures omega-squared, ω^2

$$\varpi^2 = \frac{SS_{BG} - df_{BG}(MS_{error})}{SS_T + MS_{error}} \quad \text{(Formula 14.5)}$$

Repeated-measures eta-squared, η^2

$$\eta^2 = \frac{SS_{BG}}{SS_T} \quad \text{(Formula 14.6)}$$

Formula for Tukey's *HSD*, repeated measures

$$HSD = q\sqrt{\frac{MS_{error}}{n}} \quad \text{(Formula 14.7)}$$

Repeated-measures Fisher's *LSD*

$$t = \frac{M_i - M_j}{\sqrt{MS_{error}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \quad \text{(Formula 14.8)}$$

Key Terms

Repeated-measures Design
Order Effects
Counterbalancing

Questions and Exercises

1 A repeated-measures ANOVA can be seen as an extension of the dependent-samples t test into what situations?

2 What is the methodological difference between a repeated-measures design and a between-groups design?

3 What are "order effects" and why are they a concern?

4 What is "counterbalancing" and how can it be helpful?

5 When should repeated-measures designs not be used?

6 In a repeated-measures design, what accounts for between-group variation, and what accounts for within-group variation?

7 How does the error term (denominator) in the F ratio differ between an independent-groups design and a within-groups design?

8 What values comprise the F ratio for a repeated-measures ANOVA?

9 How is it that individual differences are removed from a repeated-measures design?

10 Statistically speaking, in which of the following experimental situations would a repeated-measures design be most advantageous compared with a between-groups design?
- **a** Access to many participants and large individual differences
- **b** Access to many participants and small individual differences
- **c** Access to few participants and large individual differences
- **d** Access to few participants and small individual differences

11 Suppose we have access to 15 participants and we need to measure performance across three conditions. Compare the *df* values if we ran a one-way ANOVA with the *df* values if we ran a repeated-measures ANOVA. Compare F_{crit} values ($\alpha = .05$) as well.

12 If we add together df_{BP}, df_{error}, and df_{BG} – what does this equal?

13 If we conduct a repeated-measures study with 5 treatment conditions and 20 participants, what would be the *df* for the *F* ratio?

14 Suppose repeated-measures ANOVA results are reported as $F(3, 24) = 4.25$, $p < .05$. How many participants were involved in the study?

15 Fill in the missing values in the following repeated-measures ANOVA summary table. The study has three experimental conditions; five participants are run under each condition. Use $\alpha = .05$ to test the null hypothesis.

Source	*SS*	*df*	*MS*	*F*	*p*
Between-groups				11.51	
Within-groups	20				
Between-participants					
Error			2.30		
Total					

16 Complete the following repeated-measures ANOVA summary tables. Test the *F* (use $\alpha = .05$) to see if the null hypothesis can be rejected.

a

Source	*SS*	*df*	*MS*	*F*	*p*
Between-groups	80	2			
Within-groups					
Between-participants		4			
Error			2.38		
Total	110				

b

Source	*SS*	*df*	*MS*	*F*	*p*
Between-groups		4			
Within-groups	63.82				
Between-participants		8			
Error	51.42				
Total	184.37				

17 A study is conducted to examine the sales performance associated with different incentive programs. Over a one-month period, three incentive programs are used at a local car dealership. The number of cars sold by each salesperson, under each incentive program, is presented in the following table.

a Summarize the results in an ANOVA summary table.
b Calculate a measure of effect size (either one), if appropriate.
c If necessary, run Tukey's *HSD* tests to help clarify the findings.
d Interpret the findings.
e Are there any methodological issues that need to be addressed?

Incentive *A*	Incentive *B*	Incentive *C*
8	7	4
6	6	4
9	5	3
7	4	3
7	5	5

18 A psychologist is interested in the effects of subliminal messages on problem solving. A repeated-measures design is used. Simple arithmetic problems are presented on a computer screen; the participants are told to work as quickly as possible. In the positive condition, the phrase "Good Work" is flashed just below recognition threshold, every 30 seconds. In the negative condition, the phrase "Don't Fail" is flashed. In the control condition, no subliminal phrase is projected. The number of problems correctly solved is presented in the following table.

a Summarize the results in an ANOVA summary table.
b Calculate a measure of effect size (either one), if appropriate.
c If necessary, run Fisher's *LSD* tests to help clarify the findings.
d Interpret the findings.
e Are there any methodological issues that need to be addressed?

Positive	Negative	Control
45	20	30
56	18	29
59	10	24
48	15	25

19 A wine manufacturer would like to know which of three hors d'oeuvres goes best with their white Chardonnay. Participants are asked to take a bite

of an hors d'oeuvre, sip the wine, and provide a taste rating from 1 – atrocious to 10 – fantastic. Taste ratings are provided in the following table. Test the null hypothesis when $\alpha = .05$.

a Summarize the results in an ANOVA summary table.
b Calculate a measure of effect size (either one), if appropriate.
c If necessary, run Tukey's *HSD* tests to help clarify the findings.
d Interpret the findings.
e Are there any methodological issues that need to be addressed?

Feta Cheese	Caviar	Popcorn
1	1	3
1	2	4
2	2	6
2	3	5
1	3	6

20 We observe that people seem to be happier when they are wearing a new article of clothing. We would also like to test whether level of happiness depends on the particular type of new clothing worn. To test this, we provide a random sample of five of our classmates with new T-shirts and new shoes and instruct them to wear each article of new clothing for one day and to wear only one new article each day. Order of wearing the articles is counterbalanced across participants. At the end of the day, we ask these participants to rate, on a 10-point scale, how happy they are. On another day, when they are not wearing a new article of clothing, we also ask for a happiness rating. Ratings for each participant are reported below. Higher scores indicate greater happiness.

a Summarize the results in an ANOVA summary table.
b Calculate a measure of effect size (either one), if appropriate.
c If necessary, run Fisher's *LSD* tests to help clarify the findings.
d Interpret the findings.
e Are there any methodological issues that need to be addressed?

New T-shirt	New Shoes	Control
6	8	4
6	7	6
7	9	5
5	7	3
8	10	5

21 A psychologist is interested in the effects of distraction on pain tolerance. Three different slide shows, varying in distraction, are projected on a screen while participants have their hands immersed in ice-cold water. A 20 minute interval is used between experimental conditions to allow participants to recover from the preceding hand immersion. The number of seconds participants kept their hands in the water is presented in the following table.

a Summarize the results in an ANOVA summary table.
b Calculate a measure of effect size (either one), if appropriate.
c If necessary, run Fisher's *LSD* tests to help clarify the findings.
d Interpret the findings.
e Are there any methodological issues that need to be addressed?

Distraction		
Low	**Medium**	**High**
56	32	120
76	65	90
60	55	69
72	70	100
50	57	111

Computer Work

22 An educational psychologist is interested in comparing three visual scanning techniques on reading speed. The reading speeds of six participants are recorded (in seconds) after training in each technique. Conduct a repeated-measures ANOVA on the following data.

a Summarize the results in an ANOVA summary table.
b Calculate a measure of effect size (either one), if appropriate.
c If necessary, run Fisher's *LSD* tests to help clarify the findings.
d Interpret the findings.
e Are there any methodological issues that need to be addressed?

Participant	Technique *A*	Technique *B*	Technique *C*
P_1	450	250	500
P_2	426	300	456
P_3	399	170	300

(*Continued*)

(Continued)

Participant	Technique *A*	Technique *B*	Technique *C*
P_4	400	227	310
P_5	420	225	250
P_6	350	270	350

23 Where students study may be as important as how much they study. Studying regularly in a quiet setting may lead to a different performance than studying regularly in a noisy setting or studying different locations on different days. To study this, a random sample of six introductory psychology students was asked to study their psychology material for one hour every day, first in a special quiet room in the university library, next in the dining hall, and last rotating among a classroom, dorm room, the quad, dining hall, and library quiet room. Students spent two weeks in each condition. They were tested with a 25-point quiz at the end of each treatment condition. Order of presentation of study environments was counterbalanced across participants. Individual test scores are listed below. Conduct a repeated-measures ANOVA on the following data.

a Summarize the results in an ANOVA summary table.
b Calculate a measure of effect size (either one), if appropriate.
c If necessary, run Tukey's *HSD* tests to help clarify the findings.
d Interpret the findings.
e Are there any methodological issues that need to be addressed?

Participant	*Quiet Room*	*Dining Hall*	*Various Sites*
P_1	12	10	9
P_2	20	16	18
P_3	19	16	20
P_4	25	20	22
P_5	18	15	13
P_6	14	10	14

24 Dion, Berscheid, and Walster (1972) examined the stereotypes we hold about attractive people. Participants looked at three types of photographs: one of a physically attractive person, one of a person of average attractiveness, and a photograph of an unattractive person. Participants supplied ratings along various dimensions, including occupational success, marital and parental competence, happiness, and their social desirability as a person. Even though the photographs were of people unknown to the participants,

attractive people were viewed as superior to unattractive people, whether the target person was a biological male or a biological female. The following data set is hypothetical, providing scores for the social desirability of the person's personality. Higher scores reflect greater social desirability. The scores have been generated so that the results of our analysis will be consistent with those of the authors. Set alpha at .05, and test the null hypothesis that there is no difference in ratings among the three conditions.

	Target Person		
Participant	**Unattractive**	**Average**	**Attractive**
P_1	35	60	65
P_2	39	59	74
P_3	45	59	47
P_4	50	45	65
P_5	30	58	72
P_6	37	56	74
P_7	50	59	69
P_8	44	63	55
P_9	59	60	49
P_{10}	65	49	57
P_{11}	51	58	65
P_{12}	49	48	62
P_{13}	47	47	67
P_{14}	53	48	66
P_{15}	36	43	59
P_{16}	39	57	70

25 A social psychologist is interested in the stereotyping of masculinity among biological females. Fifteen biological females examine three pictures of biological males and provide ratings on how likely the target person is to achieve success in the corporate world (1 – very unlikely to 10 – very likely). In the traditional condition, a biological male is pictured possessing the traditional physical characteristics of masculinity (e.g. broad jaw, some facial hair). In the nontraditional condition, participants view a biological male with softer facial features and even some basic makeup to cover facial blemishes. In the control condition, the target person has both traditional and nontraditional facial characteristics. Conduct a repeated-measures ANOVA. Perform *pairwise* comparisons using Fisher's *LSD*, if warranted.

Participant	Traditional	Nontraditional	Control
P_1	10	7	8
P_2	9	4	7
P_3	7	3	4
P_4	8	6	5
P_5	6	3	6
P_6	9	2	4
P_7	10	6	7
P_8	5	10	7
P_9	3	5	10
P_{10}	4	4	4
P_{11}	8	3	6
P_{12}	7	2	3
P_{13}	2	8	6
P_{14}	9	3	5
P_{15}	7	1	3

26 A relatively recent research topic concerns the effect of electronic pedometers on health (e.g. Jackson & Howton, 2008). Below are constructed data from 10 participants who were given a pedometer and told to record the number of steps they took in a week. The study lasted 12 weeks. Data from weeks 1, 6, and 12 are below. Conduct a repeated-measures ANOVA. Use Fisher's *LSD* to perform follow-up comparisons if warranted.

Number of Steps (×1000)			
	Week		
Participant	1	6	12
1	6	7	8
2	3	4	7
3	6	3	6
4	1	6	5
5	1	5	6
6	5	8	9
7	4	6	8
8	5	10	7
9	3	5	10
10	4	7	5

27 Busseri, Choma, and Sadava (2009) compared optimists with pessimists for Past, Present, and Future life satisfaction judgments regarding their own life experience. Not surprisingly, pessimists were less satisfied with their current life experiences compared with optimists; however, although both sets of people expected brighter futures, pessimists did so even more than optimists. What if we wanted to use a different personality characteristic to explore life satisfaction for the Past, Present, and Future? Another interesting personality trait that might have a bearing on perceptions of life experiences would be the introvert/extrovert dimension. Suppose we take 8 extreme extroverts and ask them to answer questions about past, present, and projected future life satisfaction (higher scores reflect greater life satisfaction). Below are constructed data. Conduct a repeated-measures ANOVA. Use Fisher's *LSD* to perform follow-up comparisons if warranted.

	Life Satisfaction		
Participant	**Past**	**Present**	**Future**
1	17	17	18
2	13	11	19
3	7	14	16
4	15	8	11
5	18	7	14
6	8	15	15
7	9	10	19
8	15	12	15

Part 5 Review

Analyses of Variance

Review of Concepts Presented in Part 5

The purpose of this brief review section is to revisit both the similar concepts that hold Chapters 12–14 together and the concepts that distinguish them one from another. First let us look at the similarities. All three of the statistical tests presented in these chapters (one-way ANOVA, two-way ANOVA, repeated-measures ANOVA) are grounded on the same basic logic. That is, each one is designed to test a null hypothesis of no difference between population means by comparing a measure of variance *between* the samples based on primary variance (or "treatment effect") and random factors with another measure of variance *within* the samples based merely on random factors alone. This comparison is done in ratio form such that if there is no primary variance present, the measures of variance due to random factors should roughly equate, yielding a resultant value close to 1. If, however, primary variance is present, the resulting ratio will increase corresponding to the amount of primary variance present. The statistic generated in all ANOVAs is an F (named for Sir Ronald Fisher). As with the t tests presented earlier, a table of critical values based on null distributions of various design forms can be found in Table A.5. If the observed F equals or exceeds the critical F, statistical evidence has been found suggesting the null hypothesis of no differences between population means is false. (Of course, the cautiousness associated with drawing probabilistic conclusions presented in previous chapters holds true for these decisions as well.)

Another point of similarity between the chapters concerns what can be done if an overall null hypothesis is rejected. In all three chapters we are introduced to

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

two different ways to estimate the size of an effect (using omega-squared, ω^2, or eta-squared, η^2). The formulas vary slightly depending on the features of the design, but each one creates a ratio comparing a measure of the amount of primary variance with a measure of the amount of the total variance in the system. Additionally, in an effort to locate the source(s) of statistical evidence, all three ANOVAs introduce tools for the comparison of means between pairs of cells, namely, Tukey's *HSD* and Fisher's *LSD*. (A proper discussion exploring which post hoc tools should be used in which situation is beyond the scope of this resource. As a result, Tukey's *HSD* and Fisher's *LSD* are introduced as two general-purpose comparison tools.) Once again, the formulas between the three chapters vary slightly depending on the design form, but each one allows the investigator to see if the mean difference between any two cell means is large enough to suggest a difference at the population level. Since numerous tests can be run in designs with 3 or more cell means, the tests selected are conservative to keep the accumulated alpha value low.

The differences between the ANOVA stem from the number of factors used and the way participants are assigned to conditions. Both the one-way and repeated-measures ANOVA make use of only one factor. Each one can accommodate numerous conditions, but these conditions must vary across a single dimension. The two-way ANOVA is used for designs where two factors are being used. Because of this, two different types of effects have been introduced: main effects and interactions. A main effect is an effect due to the action of only one factor. The main effect for a given factor can be imagined by collapsing across the conditions of the other factor. The main effect for therapy style, for example, does not take into account that participants are also given one of two different medicines. This second condition is collapsed, and all that matters for the main effect for therapy style is a difference in the dependent measure between the various therapy conditions. Of course, the main effect for medicines investigates dependent measure differences based on medicinal condition, regardless of the therapy type individuals are receiving. Since two-way designs have two factors, there are two potential main effects for each two-way factorial design. The interaction is the effect caused by the combination of factors and not reducible to either main effect. This effect is considered a higher-order effect than main effects and must be prioritized in terms of analysis, if present.

One-way and repeated-measures ANOVAs, although both dealing with single-factor designs, are distinguished by the way participants are assigned to the various conditions. One-way ANOVAs are used to analyze between-group (or independent-group) designs; each participant is assigned to one condition and measured only once. Repeated-measures ANOVAs are used to measure within-group (or repeated-measures) designs; each participant is exposed to each level of the factor (usually an independent variable because these designs are usually experimental). Because of this difference, the numerator in a repeated-measures

F ratio does not include variance due to individual differences; they are logically eliminated when each condition is comprised of the same participants. To restore the logic of the *F* ratio, the variance due to individual differences needs to be partitioned out of the denominator. In the end, this creates a more statistically efficient mechanism for detecting the presence of primary variance; it is a more statistically powerful test than the one-way ANOVA.

The problems with the repeated-measures design are methodological and surround the issue of repeatedly measuring a participant. Confounding variance can easily be introduced if carryover effects from these repeated exposures to the dependent measure are not controlled. Also, some treatment types are irreversible, or if not irreversible, leave long-lasting effects on the participants. Research situations using these types of variables make repeated-measures designs inappropriate to use. However, if the methodological issues can be appropriately addressed, the statistical advantage of increased power that is gained when using a repeated-measures ANOVA can be rather dramatic.

Since real-world research problems do not come with a label informing the researcher of which test to use for analysis, it is important for us to work on our diagnostic skills. Understandably, the exercises at the end of each particular chapter require use of only the tests found and studied within that chapter for solution. They are designed for us to gain familiarity with using the tools just described in that chapter to solve a statistical problem. They are not designed to challenge our diagnostic skills (i.e. knowing which test to use for a given situation). However, the following review section has been created to help us develop these skills.

The exercises below will help us review the statistical differences between the various ANOVAs introduced in Chapters 12–14 and the *t* tests introduced in the preceding chapters. The hypothesis testing exercises will not identify which test is appropriate for the described scenario. We will need to use the available information presented in the exercise to make that determination. (Note: Most of the exercises below can be solved either with or without the use of statistical software.)

Questions and Exercises

1 Which pairs of tests theoretically fit together well (more than one can be selected)? Why?
- **a** Single-sample *t* test; one-way ANOVA.
- **b** Single-sample *t* test; two-way ANOVA.
- **c** Independent-samples *t* test; one-way ANOVA.

d Independent-samples *t* test; repeated-measures ANOVA.
e Dependent-samples *t* test; two-way ANOVA.
f Dependent-samples *t* test; repeated-measures ANOVA.

2 For designs with only two cells (either independent groups or repeated measures), is there an advantage of one type of analysis over another (i.e. a *t* test compared with an ANOVA)?

3 Which two types of ANOVAs presented in this text are needed to understand the term "mixed design"?

4 A clinical psychologist who works with alcoholics is interested in the effects of therapy and medication in preventing relapse. Sixteen patients at an outpatient treatment center volunteer to participate in a study. Participants are randomly assigned to either a "talk" therapy or "relaxation" therapy condition and to one of two medication conditions, receiving either imipramine (an antidepressant) or vitamin B_1. Total alcohol-free days in a one-month period are used as the dependent measure. Individual data are presented below. Select and run the appropriate statistical analysis and provide a general interpretation of the findings.

Talk therapy		Relaxation therapy	
Imipramine	Vitamin B_1	Imipramine	Vitamin B_1
28	8	18	11
21	9	18	9
20	13	11	8
27	12	16	10

5 A friend wants to see if a background color influences unconscious perceptions of biological female attractiveness for biological males. Our friend has biological male participants that look at a series of 200 pictures of individuals; many different types of backgrounds are used. Unbeknown to the participants, the picture of one individual is repeated in the series, once with a red background (believing this color to be unconsciously associated with sexuality) and once with an off-white background. The dependent variable is the amount of time (in seconds) the participants look at the target images before moving on to the next one. The presentation order of the two images is counterbalanced across the participants. The gathered data are presented below. Select and run the appropriate statistical analysis and provide a general interpretation of the findings.

	Background color	
	Red	Off-white
P_1	4.3	3.5
P_2	2.0	2.0
P_3	2.7	2.2
P_4	3.4	3.0
P_5	3.9	3.3
P_6	5.1	5.1
P_7	1.8	1.5

6 Another friend does not believe that the effect found in Exercise #5 is due to the supposed sexual nature of the color red but rather to the fact that it is bright in comparison to the boring off-white color. This friend constructs another similar study but this time introducing a bright green background as well. So, now one particular individual is shown three times, once each with a red, green, and off-white background. Counterbalancing is once again used. The data follow. Select and run the appropriate statistical analysis and provide a general interpretation of the findings.

	Background color		
	Red	Green	Off-white
P_1	4.6	4.4	3.5
P_2	2.5	2.2	2.0
P_3	1.5	1.1	1.2
P_4	4.4	4.2	3.4
P_5	2.7	2.9	2.3
P_6	4.1	4.0	4.1
P_7	5.0	3.9	3.5

7 A preschool teacher would like to make sure students rest during quiet time. The teacher wonders if the children will relax more quickly if a story is read to them, soft music is played, or they drink a glass of milk. Children are randomly assigned to one of three treatment conditions. For one week, the teacher records the average number of minutes it takes each child to fall asleep. The data are shown below. Select and run the appropriate statistical analysis and provide a general interpretation of the findings.

Sleep inducer		
Story	**Music**	**Milk**
6	4	2
6	8	6
9	7	5
8	6	4
8	10	7
10	6	5
12	5	3

8 Where students study may be as important as how much they study. Studying regularly in a quiet setting may lead to a different performance than studying regularly in a noisy setting or studying different locations on different days. To study this, a random sample of 18 introductory psychology students was asked to study their psychology material for one hour every day, some in a special quiet room in the university library, some in the dining hall, and some rotating among a classroom, dorm room, dining hall, and library quiet room. Students spent two weeks studying in this manner. They were tested with a 25-point quiz at the end of the two weeks. Individual test scores are listed below. Select and run the appropriate statistical analysis and provide a general interpretation of the findings.

Quiet room	**Dining hall**	**Various sites**
12	10	9
20	16	18
19	16	20
25	20	22
18	15	13
14	10	14

9 A frequently pondered topic of many university students concerns the broad area of gender differences. Suppose, in particular, we are interested in exploring differences in preparation time for a formal social gathering (e.g. a campus dance). Timing devices are randomly given to 15 classmates. They are asked to start the device when they begin to get themselves ready for the dance and to stop the device when they are ready to leave their dorm room. The times, in minutes, are recorded below. Select and run the appropriate statistical analysis and provide a general interpretation of the findings.

Time	Stated gender
14	M
35	F
17	M
28	M
20	F
7	M
22	F
38	F
21	M
33	M
19	M
27	F
24	F
30	F
22	M

10 An instructor of a creative writing course wonders if there are changes in creativity due to the weather, specifically the outdoor temperatures. Since the instructor teaches a yearlong course, there is an opportunity to sample poems written by students in hot, warm, and cold weather. The works of seven students are randomly selected and rated by other professors for their creativity. The average of those ratings is presented in the table below. Select and run the appropriate statistical analysis and provide a general interpretation of the findings.

	Outdoor temperature		
	Warm	Hot	Cold
P_1	2	4	6
P_2	5	3	6
P_3	4	7	8
P_4	8	7	9
P_5	3	4	10
P_6	8	6	5
P_7	7	8	9

11 Over the years much research has been conducted on the psychology of negotiations (see Loschelder, Swaab, Trötschel, and Galinsky (2014) for a recent example). A couple variables of particular interest to many researchers are the first mover (i.e. whether the seller or buyer initially states the conditions of transaction) and the biological sex of the seller. In this experiment a raffle ticket for a new mountain bike is initially gifted to each participant. An acquaintance of the researcher who is ignorant of the hypothesis and is only told to buy back each ticket using as little money as possible, then, engages in a negotiation with each participant to buy back the ticket prior to the raffle. In one condition the participant (seller) is told to make the opening bid; in another condition the buyer is asked to make the opening bid. The biological sex of the seller is also noted. The data are presented below. Select and run the appropriate statistical analysis and provide a general interpretation of the findings.

		Biological sex	
		Male	Female
First mover	Seller	\$8.50	\$6.25
		\$6.00	\$2.75
		\$4.50	\$4.00
		\$3.75	\$2.25
	Buyer	\$4.75	\$2.75
		\$5.00	\$5.00
		\$6.75	\$3.00
		\$3.50	\$1.75

Part 6

Inferential Statistics

Bivariate Data Analyses

15

Linear Correlation

15.1 The Research Context

This chapter discusses correlational analysis. A **correlation coefficient** is a measure of the strength of association between two variables. A correlation coefficient can range from −1 to +1. The larger the absolute value of the correlation, the stronger the association between two variables. Measuring the strength of association between two variables has a very broad and useful function for scientific investigations. In this chapter, the focus is on the use of correlations in the behavioral and social sciences, especially psychology. Keep in mind, however, that correlations can be computed to answer questions from many different fields including economics (is there a relationship between gross national product and the value of the dollar?), meteorology (is there a correlation between rainfall and number of trees per acre of land?), sociology (is there a correlation between household incomes and religiosity?), and medical epidemiology (is there a relationship between the size of the local deer population and the incidence of Lyme disease?).

In psychology, correlational analyses are often applied to two attributes or an attribute and an overt behavior. Two scores are obtained from each participant. Together, the two scores define a *pair* of scores. The distribution of pairs of scores is called a **bivariate distribution** (*bi* meaning two, *variate* meaning variable). IQ and academic performance, anxiety and fine motor movements, depression, and self-reinforcement are all examples of bivariate data for which correlations have been found. This chapter addresses the application of a correlational analysis to data measured on an interval or ratio scale. Chapter 18 discusses two other correlational analyses to be used when data come from an ordinal or even a nominal scale.

The statistical methods for calculating the correlation were invented by Sir Francis Galton (1822–1911). However, the precise formula for the statistic

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

discussed in this chapter was derived by Karl Pearson (1857–1936) and is called, more formally, the **Pearson product-moment correlation coefficient**. (See Spotlight 15.1 for more information about Karl Pearson.) The Pearson r is the most powerful and most frequently used version of the correlation measure. The Pearson r relies on the same statistical and methodological assumptions as the independent-samples t test, when applied to both variables, namely, representativeness, independent observations, interval or ratio scale of measurement, and normality. All sample correlation coefficients are symbolized using r; all population correlation coefficient are symbolized using ρ (*rho*).

Spotlight 15.1 Karl Pearson

Karl Pearson was born in London in 1857, two years before Darwin published *Origin of Species,* a work that would shape Pearson's entire academic life. It is reasonable to consider Pearson the father of modern statistics. Pearson completed the work on correlation that Galton had started, arriving at the coefficient that bears his name (the Pearson product-moment correlation coefficient). He subsequently devised formulas for computing correlations for variables that are noncontinuous (see Chapter 18). Pearson is responsible for many of the concepts and statistical terms that were introduced earlier in this book: the histogram, mode, and standard deviation. He also invented the chi-square test (see Chapter 17).

Karl Pearson believed himself to have been a careful thinker from the beginning of his life. He claimed his earliest memory has him sitting in a high chair, sucking his thumb when someone urged him to stop, so his thumb would not wither away. Upon examining both thumbs he thought, "I can't see that the thumb I suck is any smaller than the other; I wonder if she could be lying to me" (Walker, 1968, p. 497).

Pearson's abiding belief in the importance of observation, if not his rejection of authority, guided his lifelong pursuit: the development of mathematical tools that could be used to test the theory of Darwinian evolution. Over his lifetime, he published more than 500 works. When asked how he found the time to publish so much, he offered, "You Americans would not understand, but I never answer a telephone or attend a committee meeting" (Stouffer, 1958, p. 25).

After earning a degree in mathematics at King's College, Cambridge, Pearson studied law and subsequently established a private practice for three years. In 1884, he abandoned law, became a professor of mathematics at University College London, and began his illustrious career. The first major influence on Pearson's thinking was a book published by Sir Francis Galton, *Natural Inheritance*. Although Pearson was never a formal student of Galton's, he became his disciple and defender. Pearson was looking for a model of semi-determinism as an alternative to what he believed was the biological sciences' rigid adherence

to causality. He found this semi-determinism in the concept of the correlation. For Pearson, the correlation represented a fundamental paradigm shift, which, he believed, would revolutionize the biological and the social sciences. Researchers could use the correlation as an important measure of the "degree of relatedness" between two variables without having the strict determinist's burden of claiming causality. In 1896, Pearson introduced the formula that we now use to compute the correlation between two continuous measures. [As an historical aside, the Pearson formula was actually first published a year earlier by Yule (1895), a student of Pearson's who gave his mentor full credit for the formula.]

Pearson was a product of his times. Darwin's ideas about the fundamental principles that guide evolution, namely, heredity, variation, and natural selection, influenced the thinking of many scholars. Galton had introduced the term eugenics and started the eugenics movement, which was dedicated to improving the human race through selective breeding. This was a period of time when many scientists believed nature (heredity) to be far more important than nurture (environment) in determining personal qualities (like temperament, intelligence, personality, and even one's moral proclivities). Indeed, Pearson developed formulas for correlation coefficients appropriate for noncontinuous measures in his attempt to show that "…the degree of resemblance of the physical and mental characteristics in children is one and the same" (Pearson, 1903, p. 203). He became a strong advocate for eugenical action. Pearson could be quite accurately described as a Social Darwinist, an imperialist, nationalist, and a racist (Grosskurth, 1980). For instance, he believed that war was necessary to eliminate "inferior stock." He also opposed legislation to aid the oppressed. According to Pearson, "No degenerate and feeble stock will ever be converted into healthy and sound stock by the accumulated effects of education, good laws, and sanitary surroundings" (Semmel, 1958). Unfortunately, his nationalistic and racist beliefs even influenced his scientific conclusions (e.g. Delzell & Poliak, 2013). Several of his research projects serve as an example of how personal bias and preconceived notions can influence the methods used and even the conclusions drawn from scientific investigations (see Box 2.3).

The eugenics movement was birthed in England, came of age in the United States (where forced sterilization, marriage restrictions, and eugenical segregation policies were legalized to varying degrees across the country), and was adopted as national policy in Nazi Germany (e.g. Kühl, 1994). The death of millions of physically, mentally, racially, and socially "inferior" people was the result.

Pearson died in 1936, just before the outbreak of World War II. Moving forward we must take only the best of Karl Pearson, namely, his insistence on the importance of quantifying social phenomena and the numerous correlational techniques he developed to achieve this goal. Additionally, we can recognize that his vision to accomplish a paradigm shift in the social sciences has led to many innovative, multivariate methods that bridge the gap between experimental and correlational designs.

The correlation coefficient is a measure that reflects the degree to which two variables are associated. However, it cannot be inferred from a mere association that a causal relationship in either direction exists between the two variables. For example, psychologists know that depressed persons are less likely to reward themselves for achieving a goal. However, this could be because depression leads to reward minimization (*A* causing *B*), minimizing rewards leads to depression (*B* causing *A*), or some third variable is responsible for both depression and a low rate of self-reward (*C* causing both *A* and *B*). The methodological context within which the correlation between two variables is found determines the most reasonable interpretation of the correlation.[1]

The Distinction Between a Correlational Design and the Correlation Coefficient

We have probably heard the motto "correlation does not imply causation." This phrase can be misleading if we do not keep in mind the difference between how data are collected and the statistical analysis used to interpret the data. If we conduct a study in which the researcher manipulates one variable and the other is observed and measured, the study is called an experiment. Under these conditions, we may be warranted in making causal statements about the relationship between independent and dependent variables, even if we use the correlation statistic for analysis. If we conduct a study and do not exert control over an independent variable, then we are using a correlational design, and a causal analysis is inappropriate. Whether causal language can be used to explain a relationship is entirely determined by the nature of the research design, not the statistical analysis. A much better motto would be, "correlational *design* does not imply causation." The following examples remind us of the difference between a correlational design and an experiment.

Some people believe that a person's mood is affected by the temperature outside. Suppose we tested this hypothesis by recording the daily temperature and obtaining mood ratings every day from several people. After analyzing the data, we discover that more positive mood ratings are associated with warmer temperatures. Is there a causal connection? Well, obviously mood did not affect the weather; so must it be that the changes in temperature directly caused the changes in mood? Not necessarily. It could be that when it is cold, people do not socialize as often; their mood is depressed due to the loss of social contact. In other words, a third variable (socializing) may account for the observed relationship between temperature and mood.

1 For review, please go back and read Section 1.6.

If we wanted to determine if there is a direct, causal relationship between temperature and mood, we would have to find a way to systematically manipulate the temperature, control for socialization, and then examine its effect on mood. Perhaps we could alter the temperature in a room, holding all other conditions constant (including socialization), and then see if the relationship between temperature and mood holds. Now the design is experimental and a causal interpretation is allowed.

Suppose we want to test the hypothesis that anxious individuals finish their exams faster than calm individuals. Based on a standard assessment technique, we classify everyone in the class as either anxious or calm. Without the students' knowledge, we time how long it takes each student to complete their exam. We then analyze the data by conducting a t test between the time-to-completion means of the two groups. Even though we have used a t test, an analysis that is usually associated with an experiment, can we make a causal statement? No, because we have not manipulated any variables. Whenever a participant variable is used to create groups (e.g. personality trait, biological sex, psychiatric diagnosis), the design is correlational (see Chapter 1 for a discussion of participant variables).

It is true that correlational designs often use a correlation coefficient (r) for analysis and experiments often use t tests and F tests for analysis. However, this is not always the case. For example, suppose we are interested in understanding the relationship between the use of practice imagery and bowling proficiency. To do this, we randomly assign participants to various amounts of imagery practice time where they visualize themselves using proper bowling form. Subsequently, bowling scores are recorded. In this situation, a correlational analysis could be used to identify the strength of association between the amount of time visualizing proper bowling form and bowling scores. If we found a strong correlation between these two variables, could we state that the visualization technique *caused* an increase in bowling performance? Yes, because we manipulated an independent variable. We directed participants to practice imagery for differing stretches of time. In this context, we would be justified in connecting correlation and causation. The basis for making causal statements *never* resides with the type of statistical analysis, but rather with the methodology used to collect the data.[2] Cronbach (1967, p. 27) offers a nice metaphorical distinction between the experimental and correlation approaches, "… the experimentalist [is] an expert puppeteer, able to keep untangled the strands to half-a-dozen independent variables. The correlational psychologist is a mere observer of a play where Nature pulls a thousand strings."

2 For an introductory yet in-depth treatment of the use of correlational techniques as applied to experimental designs, refer to Keppel and Zedeck (1989).

15.2 The Correlation Coefficient and Scatter Diagrams

This section addresses the statistical aspects of correlation and hypothesis testing. Recall that a correlation coefficient is represented by a number that ranges from −1 to +1; the higher the coefficient's absolute value, the stronger the association between the two variables. An r of −.80 reflects an association as strong as an r of +.80. A correlation of 0 reflects no relationship between the two variables. If higher values of one variable are associated with higher values of the other variable (as with IQ and academic performance), the correlation is said to be *positive*. If higher values of one variable are associated with lower values of the second variable, the correlation is said to be *negative*. For instance, educational achievement and number of children are negatively correlated. That is, the more educated a person is, the fewer children they tend to have.[3]

Chapter 2 showed a few ways in which a distribution of scores can be displayed. For instance, polygons and histograms are common ways of displaying data from a univariate distribution. When a bivariate distribution is plotted on a graph, it is called a **scatter plot** (or *scatter diagram*). Consider a study in which a child psychologist is interested in developing a measure of aggression (see Box 15.1 for more information on scale development). A rating scale is given to each of the ten children in a third-grade class; they are asked to rate the aggressiveness of their peers (on a scale from 1, not aggressive at all, to 10, extremely aggressive). To ascertain if the perceptions of aggression measured by the rating scale correspond to observed aggression, the psychologist conducts behavioral observations of interactions among the children during recess. In Table 15.1, the X column lists the average peer ratings gathered for each of the 10 children in the class. The Y column lists the number of observed instances of aggression, over three days, for each child. The designation of the X and Y variables is arbitrary. The X scores for all the children constitute one univariate distribution, and the Y scores constitute another univariate distribution. Since 2 scores are recorded for each child (peer ratings and behavioral observations), the 10 pairs of scores define 1 bivariate distribution.

The bivariate distribution can be represented visually by plotting each participant's X and Y score on a graph. For this data set, there are 10 points; each point corresponds to a participant's X and Y score. Figure 15.1 is the scatter plot of the data in Table 15.1.

To draw a scatter plot, place the X variable on the horizontal axis and the Y variable on the vertical axis. To plot the data point for participant 1's score, follow the X axis to the number 8. Imagine a line drawn vertically, parallel to the Y axis. Now locate the participant's Y score along the Y axis. The Y score for participant 1 is 14. Imagine drawing a horizontal line, parallel to the X axis.

3 This observation, by the way, frustrated many eugenicists (e.g. Riddle, 1947).

Table 15.1 Hypothetical data depicting peer ratings of aggression (*X*) and observed aggression (*Y*).

	Peer ratings of aggression: *X*	Observed aggression: *Y*
P_1	8	14
P_2	10	12
P_3	4	9
P_4	1	4
P_5	5	11
P_6	6	10
P_7	3	1
P_8	9	12
P_9	7	10
P_{10}	2	4

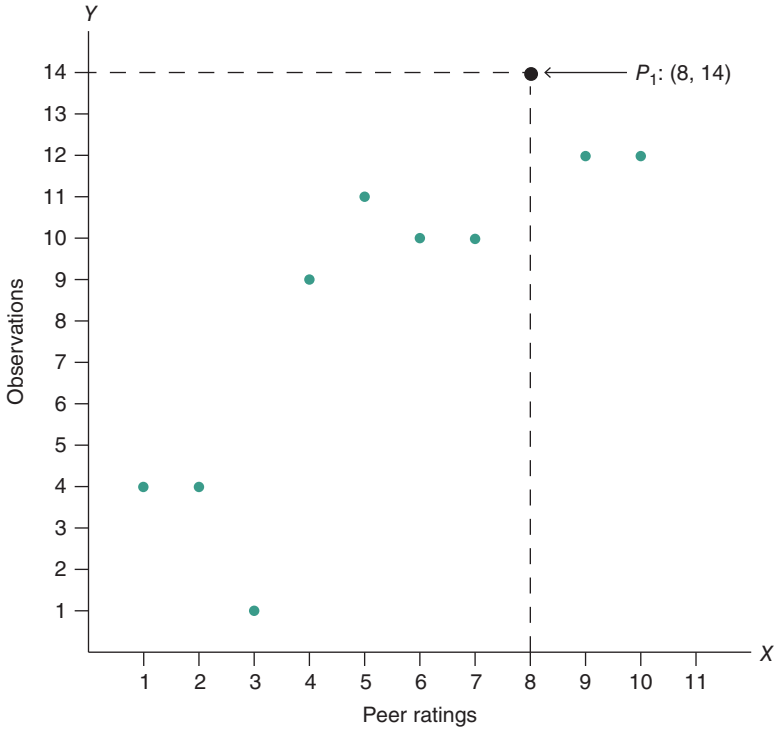


Figure 15.1 The scatter plot of the data from Table 15.1.

Where the two imaginary lines intersect is where we plot the participant's data point. As Figure 15.1 is examined, make sure that all the pairs of scores in Table 15.1 have been accurately plotted.

Interpreting the Scatter Diagram

The scatter plot provides a wealth of information about the relationship between two variables. Figure 15.2 shows the scatter plot diagrams for several correlations. The magnitude of the correlation can be estimated by looking at the general shape formed by the data points. The more narrow the width of the oval enveloping the data, the stronger the correlation. The more the data take the shape of a circle, the weaker the correlation. Compare the correlation of +.70

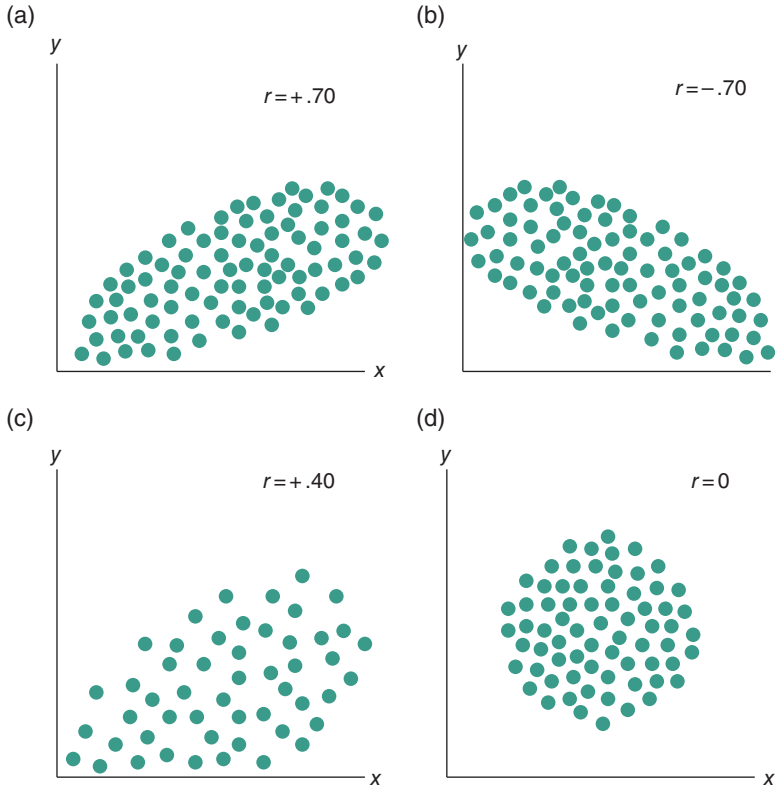


Figure 15.2 Several scatter diagrams that depict the magnitude and direction of the correlation. (a) is a strong positive correlation; (b) is a strong negative correlation; (c) is a weak positive correlation; and (d) reflects no correlation.

in Figure 15.2a with the correlation of +.40 in Figure 15.2c. Figure 15.2d illustrates a plot of unrelated variables.

Not only does the scatter plot indicate the strength of association between *X* and *Y,* but also it reveals the direction of the correlation. If the oval containing the majority of the points slopes from the lower left to the upper right, the correlation is positive. As the *X* scores tend to have higher values, the *Y* scores are apt to be larger (see Figure 15.2a and c). If the plot slopes from the upper left to the lower right, then the correlation is negative. As the value of the *X* score increases, the value of the *Y* score tends to decrease (see Figure 15.2b). Please note that the degree of the slope (e.g. gradual versus steep) is *not* indicative of the strength of the correlation. The scatter plots of two correlations of the same magnitude, but with different signs, are shown in Figure 15.2a and b.

Linear and Nonlinear Correlations

The scatter diagram provides a graphic representation of the relationship between the distributions of both variables. Viewing the plot's approximate shape allows us to make a crude estimate of the strength of association of the variables. However, a researcher would *never* report a correlation merely based on looking at the shape and direction of the oval. So why construct a scatter plot? Well, a visual analysis can be very helpful, especially for detecting nonlinear relationships.

In the examples of correlations provided so far, the higher values of one variable are associated with higher values of the second variable (positive correlation), or the higher values of one variable are associated with lower values of the second variable (negative correlation). Whether the correlation is positive or negative, the scatter plots illustrated in Figure 15.2(a–c) show linear relationships between *X* and *Y*. When *X* and *Y* have a linear relationship, the correlation is called a linear correlation. In a **linear relationship**, each time the value of one variable increases, the value of the other variable shows a constant change. In other words, the relationship between *X* and *Y* can be represented by a straight line, thus the term "linear." On the other hand, what if we observed that lower scores on *X* were associated with lower scores on *Y,* medium *X* scores were associated with medium *Y* scores, *but* higher *X* scores were found to be associated with *lower Y* scores? (See Figure 15.3.) If we were shown a scatter plot of the left half of this bivariate distribution, we would estimate a positive correlation. However, we would assume a negative correlation between *X* and *Y* if only the right portion of the distribution were illustrated. Figure 15.3 depicts the overall relationship between arousal and task performance. In this example, the data are distributed like an arch (∩). The variables are obviously associated (notice the narrowness of the arch), but any straight line will fail to capture this association. This is an example of a **curvilinear relationship.** How would we interpret the curvilinear relationship between arousal and performance, depicted in Figure 15.3?

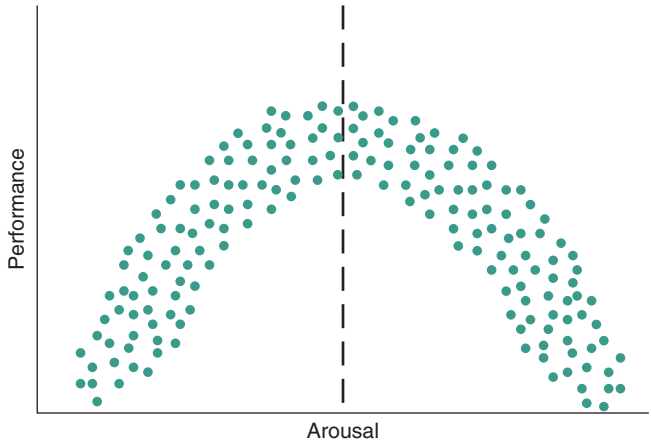


Figure 15.3 A plot of a curvilinear relationship between *X* and *Y*. As arousal increases, performance improves until arousal becomes too high, and then performance declines.

Given a certain level of task difficulty, performance is optimal when the person performing a task is experiencing a moderate level of arousal. A low or high level of arousal is associated with poor performance.

The formulas for calculating a linear correlation are different from those used to calculate a curvilinear correlation. In fact, if a linear formula were applied to the data of Figure 15.3, the correlation would be close to 0, suggesting *no* association between performance and arousal. However, by viewing the scatter plot, we can see that in actuality, there is a strong correlation between the variables. Indeed, we should *always* inspect the scatter plot of our data *before* we interpret the correlation coefficient. If the plot shows an arch or an "s" shape, the formula for a linear correlation should not be used. This chapter only addresses linear correlation since it is the most common type of correlated relationship (and the easiest to compute). Consult an advanced statistics resource to understand the analysis of curvilinear relationships.

Now that we are familiar with how to plot a bivariate distribution and interpret a scatter plot, we can turn to the statistical basis and computational methods of the correlation coefficient.

The *z* Score Formula for the Correlation Coefficient

The *z* score formula for the following correlation is not the preferred method for computing the correlation. It is much too tedious. However, an examination of the *z* score formula is a very good way to introduce the statistical conceptualization of the correlation coefficient.

Recall that a z score locates a raw score in a univariate distribution. A z score is the number of standard deviations a raw score is from the mean of the distribution. All raw scores above the mean transform to positive z scores, while all raw scores below the mean transform to negative z scores. Table 15.2 presents the X and Y scores for five participants, as well as each score's z score. In a bivariate distribution, it is important to note that when transforming a raw score of variable X to a z score (z_X), the mean and standard deviation of variable X is used in the z score formula. In Table 15.2, the mean and standard deviation for the X scores is 17 and 2.83, respectively. The same point holds for the Y scores: the transformation to z scores (z_Y) uses the mean and standard deviation of the Y distribution.

The z score formula for the correlation coefficient of a *population* is given in Formula 15.1.

The z score formula for the population correlation

$$\rho = \frac{\Sigma(z_X z_Y)}{N_p} \quad \text{(Formula 15.1)}$$

where

ρ = rho, the symbol for the population correlation
$\Sigma(z_X z_Y)$ = sum of the cross products of z scores
N_p = number of *pairs* of scores

Table 15.2 Computing the population correlation with the z score formula.

Participant	X	μ	$z_X = (X - \mu_X)/\sigma_X$	Y	μ	$z_Y = (Y - \mu_Y)/\sigma_Y$	$z_X z_Y$
P_1	13	17	−1.41	23	30	−1.52	+2.14
P_2	15	17	−0.71	28	30	−0.43	+0.31
P_3	17	17	0	30	30	0	0
P_4	21	17	+1.41	32	30	+0.43	+0.61
P_5	19	17	+0.71	37	30	+1.52	+1.08
							$z_X z_Y = +4.14$

Summary values

$\mu_X = 17$; $\sigma_X = 2.83$; $\mu_Y = 30$; $\sigma_Y = 4.60$; $N_p = 5$

$$\rho = \frac{\Sigma(z_X z_Y)}{N_Y}$$
$$\rho = \frac{+4.14}{5}$$
$$\rho = +\mathbf{.83}$$

The term $z_X z_Y$ is called a cross product. A cross product is a given participant's z_X score multiplied by the corresponding z_Y score. In Table 15.2, the cross product for participant 1 is (−1.41)(−1.52) = +2.14. The numerator of the z score formula is the sum of all individual cross products.

There are two important points to make about the z score formula for the correlation. First, under what condition would the z score formula yield a positive correlation? Remember that a positive correlation results when higher scores on one variable are associated with higher scores on the second variable. In terms of z scores, participants who score relatively high on the X variable will receive z_X scores that are positive. If they also tend to score high on the Y variable, their z_Y scores will also be positive. This will produce many positive cross products. Furthermore, those participants with X and Y scores below the means of the X and Y distributions will have negative z_X and negative z_Y scores. Two negative z scores multiplied together also produce a positive cross product. Refer to Table 15.2. The correlation between X and Y is positive (+.83). As we inspect the z_X and z_Y columns, note that negative z scores on X are associated with negative z scores on Y, and positive z scores on X are associated with positive z scores on Y. Although, in this example, every cross product is positive, it is not necessary for *all* cross products to be positive for the correlation to be positive. As long as the sum value of all cross products is positive, the correlation will be positive.

Given this explanation of how a positive correlation can arise, it should be easy to determine how a negative correlation can occur. A negative correlation occurs when the higher scores on one variable are associated with lower scores on the second variable. In terms of z scores, positive z scores would tend to be associated with negative z scores. This would yield many cross products with a negative sign. If the sum value of all cross products is negative, the correlation will be negative.

The second important point about the z score formula has to do with the magnitude of the correlation. A z score not only specifies whether a raw score is above or below the mean, but it also states how far the score is from the mean (in standard deviations). Refer to Table 15.2. Think about the relative rankings of the participants' z_X and z_Y scores. Notice that participant 1 has the lowest X score in the X distribution and therefore has the largest negative z_X score. Participant 1 also has the lowest Y score in the Y distribution and, accordingly, the largest negative z_Y score in the distribution. Participant 1 ranks at the bottom of each distribution. Note that participant 2 ranks second from the bottom of each distribution. Participant 3 ranks at the middle of each distribution. However, participant 4 ranks the highest in the X distribution but the second highest in the Y distribution. Participant 5 ranks the second highest in the X distribution but the highest in the Y distribution. Considering the entire bivariate distribution in Table 15.2, the rankings of the z_X scores are very similar to the way in which their corresponding z_Y scores are ordered. When the rankings of the X

scores have a high degree of correspondence to the rankings of the Y scores, the correlation will be large. The bivariate distribution in Table 15.2 shows a good deal of correspondence; therefore, the correlation is high: +.83. If the correlation were high and negative, there would still be a good deal of correspondence among the ranks of X and Y. However, high rankings on X would be associated with low rankings of Y.

The Computational Formula for the Correlation Coefficient

The z score formula is instructive because it provides a way to conceptualize the statistical basis of the correlation coefficient. However, since each raw score has to be transformed into a z score, using the z score formula to calculate the correlation is an arduous task. Formula 15.2 is much easier to use when working with raw scores. It is the *computational formula* for computing the correlation. In addition, the symbol r indicates that the correlation is being derived from a sample of scores, not a population. This is much more typical of social and behavioral science research.

Computational formula for Pearson *r*

$$r = \frac{n_p(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{\left[n_p(\Sigma X^2) - (\Sigma X)^2\right]\left[n_p(\Sigma Y^2) - (\Sigma Y)^2\right]}} \quad \text{(Formula 15.2)}$$

where

ΣX^2 = the sum of all the squared X scores
ΣY^2 = the sum of all the squared Y scores
XY = the cross product of an X and Y score
ΣXY = the sum of the cross products
$(\Sigma X)^2$ = the sum of all the X scores, quantity squared
$(\Sigma Y)^2$ = the sum of all the Y scores, quantity squared
n_p = number of *pairs* of observations

At first glance, this appears to be an imposing formula. If we accidentally turned to this page the first day of class, we might have thought, "I'll never be able to do this." However, we are already familiar with all of these terms. Even the notion of cross products is not new; we just encountered it with the z score formula for correlation. If we carefully follow the next worked problem, we should have no trouble using the computational formula for hand calculations.

■ **Question** *An investigator wants to know the correlation between subjective ratings of discomfort and the length of time participants can keep their hands in ice water. The X variable in the following table is the amount of discomfort; higher scores indicate greater discomfort. The Y variable is the number of minutes participants kept their hands in the water. What is the correlation*

between discomfort ratings and duration of hand immersion for this sample of participants?

Solution

X	Y	X^2	Y^2	XY
3	2	9	4	6
5	3	25	9	15
4	4	16	16	16
7	5	49	25	35
10	6	100	36	60
$\Sigma X = 29$	$\Sigma Y = 20$	$\Sigma X^2 = 199$	$\Sigma Y^2 = 90$	$\Sigma XY = 132$

■

Using the Computational Formula

After we compute the summary statistics shown at the bottom of each column, proceed as follows:

$$r = \frac{n_p(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{\left[n_p(\Sigma X^2) - (\Sigma X)^2\right]\left[n_p(\Sigma Y^2) - (\Sigma Y)^2\right]}}$$

$$r = \frac{5(132) - (29)(20)}{\sqrt{\left[5(199) - (29)^2\right]\left[5(90) - (20)^2\right]}}$$

$$r = \frac{660 - 580}{\sqrt{[995 - 841][450 - 400]}}$$

$$r = \frac{80}{\sqrt{(154)(50)}}$$

$$r = \frac{80}{\sqrt{7700}}$$

$$r = \frac{80}{87.75}$$

$$r = +\mathbf{0.91}.$$

15.3 The Coefficient of Determination, r^2

If we have already covered the ANOVA chapters, we will recall the statistic, ω^2, omega-squared. Omega-squared is a measure of effect size; it reflects the amount of variation in the scores of the dependent measure that are accounted for by the levels of the factor. The **coefficient of determination,** represented as r^2, accomplishes for the correlation coefficient what ω^2 accomplishes after an F test has been performed. An r^2 is a measure of the amount of variation of the Y variable accounted for by variation of the X variable; a measure of **shared variance** (or **common variance**). This is a bidirectional concept. Therefore, r^2 can also be stated as the *amount* of the X variable accounted for by variation in the Y variable.

Shared variance is the key concept in understanding the coefficient of determination. It is usually stated as a percentage. If the correlation between two measures is .80, then the amount of shared variance is $.80^2 \times 100 = 64\%$. What is the coefficient of determination when the correlation is –.30? Square the correlation and multiply by 100 to arrive at 9%. Researchers will use different phrases when referring to shared variance:

1) Sixteen percent of the variance of Y scores is *explained* by the variation of X scores.
2) Nine percent of X *is due to* Y.
3) Twenty-five percent of the variance of Y is *accounted for* by X.

Keep in mind that shared variance is a bidirectional notion. Whether we state it from the perspective of X or Y makes no difference. In addition, do not be confused by the terms "explained" or "accounted for." Strictly speaking, the coefficient of determination does not *explain* or *account for* anything. In other words, the coefficient of determination does not tell us *why* there is a relationship between two variables.

The coefficient of determination is a concept more difficult to understand than the correlation coefficient. However, many social and behavioral scientists believe that r^2 is a more useful measure of the relationship between X and Y than the correlation coefficient. The following examples will help further explain this important concept.

Suppose we administer two *different* tests of anxiety to a group of individuals. From previous research, we know that anxious people score higher than calm people on both tests. Now we give *both* tests to the *same* group of people. Since both scales are measuring anxiety, would we expect the correlation between them to be +1? Well, even though both tests are measuring the same concept, the correlation between them will probably not be perfect.

The score of *any* measure is based on more than what the test is designed to measure; all tests have some degree of *measurement error*. For example, in

addition to tapping anxiety, a person's score could be affected by fatigue, misunderstanding the instructions, distractions during the testing, or perhaps the tests use different ways of wording the questions.[4] Moreover, one of the tests may ask more questions aimed at the physical experience of anxiety (e.g. "How often do you feel your heart pounding?"), whereas the other test may ask more questions aimed at the cognitive aspect of anxiety (e.g. "How often do you find yourself worrying more than other people?"). Two scales designed to measure the same trait may be tapping into different aspects of that trait. In Figure 15.4, the shaded, overlapping area is what the two tests have in common. It is the shared variance of the tests. Note how the amount of shared variance (shaded area) is larger as r and r^2, increase in magnitude. The nonoverlapping areas define the extent of *unshared variance*. Measurement error and differences in the focus and wording of the questions are all factors that account for the size of the unshaded area.

Since the concept of r^2 is so important, and because it can be difficult to grasp, let us consider another more detailed example. The Wechsler Intelligence Scale for Children (WISC-R) includes dozens of questions and tasks that are used to assess various aspects of intelligence. The "block design" task, for instance, requires the child to look at a picture of a design (e.g. a diamond). Several wooden blocks, some white, some red, and some half white and half red are given to the child who must then arrange the blocks to reproduce the design in the picture. Another task, "object assembly," presents the pieces of a puzzle; the child must fit them together to form the correct figure (e.g. a horse). It is reasonable to think that some of the abilities that would lead a child to do well on the "block design" task would also lead to good performance on the "object assembly" tasks; an assessment of these abilities is what the tests have in common. Variance will be shared to the degree the measurements taken reflect

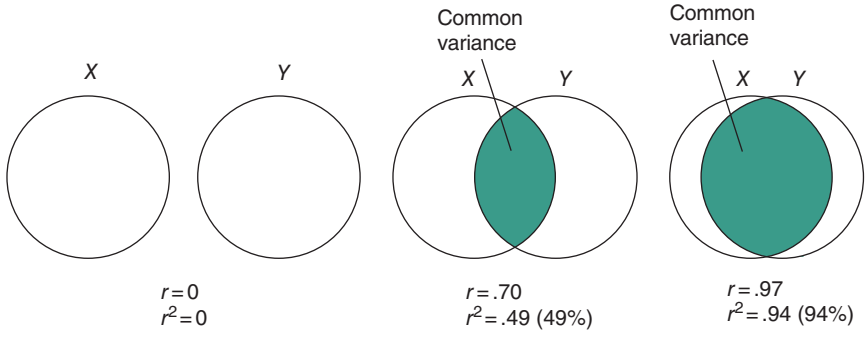


Figure 15.4 The shaded overlapping areas depict shared variance. As r increases so does r^2.

4 M*easurement error* is similar to the concept of *experimental error* discussed in Chapters 12–14.

these common abilities. Table 15.3 lists some of these common sources of variance. However, each test is also influenced by unique factors or sources of variance. Table 15.3 also lists some of these unshared sources of variance. Suppose that the correlation between performance on "block design" and "object assembly" is +.70. Refer to Figure 15.4, the middle illustration. If the correlation is +.70, r^2 is approximately 50%. The overlapping segment of the ovals, the shaded area, represents this shared variance. The nonoverlapping segments of the ovals make up the proportion of variance unique to each test. These abilities are measured by one test but not the other. (See Box 15.1 for more information about how the correlation concept is used in the development and assessment of measurement tools.)

Comparing *r* and r^2

If we compare an *r* of +.50 with an *r* of +.25, *mathematically speaking,* one is twice the size of the other. However, the *worth* of the correlations is better captured by r^2 not *r*. It is more appropriate to compare relationships in terms of their shared variance, r^2:

$$(.50)^2 \times 100 = 25\%$$
$$(.25)^2 \times 100 = 6.25\%$$
$$\frac{25\%}{6.25\%} = 4.$$

Table 15.3 Some shared and unshared abilities between the "block design" and "object assembly" tasks of the WISC-R.[a, b]

Block design and object assembly	
Shared abilities	**Abilities not shared**
(Common variance; shared variance; variance accounted for)	**(Uncommon variance; unshared variance; variance not accounted for)**
Visual–motor coordination	Analysis of whole into component parts (BD)
Spatial relations	Reproduction of models (BD)
Perceptual organization	Ability to benefit from sensory–motor feedback (OA)
Working under time pressure	

[a] See Kaufmann (1979).
[b] Those abilities in common will contribute to shared variance, r^2.
BD, block design; OA, object assembly.

Box 15.1 Next Steps with Correlations: Scale Development

A common activity for many academic psychologists is the construction of measuring tools. There are literally hundreds of different psychological traits, tendencies, and abilities that psychologists are interested in measuring, from commonly used concepts like extroversion and neuroticism to less frequently referenced concepts like humility (e.g. Rowatt et al., 2006) and right-wing authoritarianism (e.g. Mirels & Dean, 2006). The scales used to measure these attributes, however, need to be created. They do not appear out of thin air.

Scale development is usually an extensive process. First, the concept is carefully defined, with a lot of thought given to identifying various subcomponents of the concept (e.g. is extroversion marked by being very talkative, striving to be the center of attention, enjoying meeting new people, being very physically demonstrative, all four?). Second, researchers typically start to gather data to see if their understanding of the concept fits well with how people answer questions about themselves. This usually involves the generation of numerous response items, oftentimes presented in the form of a question or a statement to be agreed or disagreed with by participants using a Likert scale. These responses are then statistically analyzed.

This is where the correlation concept comes in. By looking at the size and nature of the relationships between pairs of items, researchers can gain feedback information regarding the nature and scope of the concept they are studying. For instance, if high correlations were found between individuals' responses regarding questions related to how talkative a person is, how much they enjoy being the center of attention, and how much they like meeting new people, this interrelatability would provide statistical evidence, in the form of shared variance, that the concept of "extroversion" encapsulates all of these subcomponents. If, on the other hand, the responses of individuals to questions related to being physically demonstrative do not tend to correlate highly with the responses regarding these other components, this would provide statistical evidence that this component is not necessarily a part of the concept "extroversion." Once data has been gathered and analyzed, this two-step process of concept definition/clarification and data gathering/analysis can repeat itself, often several times.

The process, as one might imagine, is actually much more sophisticated than what has been presented here. It is described in simple terms to show the relationship between the correlation concept and this important professional activity. Correlations, by the way, are also the base concept behind other sophisticated analytical techniques (see Box 16.3).

Table 15.4 The relationship between r and r^2.

	$r(\pm)$	r^2	
	.00	.00	
Small r's	.10	.01	
	.20	**.04**	4% } **Difference of 5%**
	.30	**.09**	9% }
Moderate r's	.40	.16	
	.50	.25	
	.60	.36	
	.70	.49	
Large r's	**.80**	**.64**	64% } **Difference of 17%**
	.90	**.81**	81% }
	1.00	1.00	

Correlations of small, moderate, and large are indicated on the left.

In terms of r^2, a correlation of .50 is actually four times as large as a correlation of .25. The coefficient of determination, r^2, is not *directly* related to r. In fact, as Table 15.4 illustrates, we have to go all the way up to an r of ±.70 to find shared variance equal to 50%. To give an account of shared variance from 50 to 100% requires correlations from ±.70 to 1. Another way to clarify that there is not a direct relationship between changes in r and changes in r^2 is to compare two sets of adjacent correlations. In Table 15.4, the adjacent correlations of .20 and .30 translate into a mere 5% change in the amount of shared variance. Yet, the adjacent correlations of .80 and .90 translate into a difference of 17% of the shared variance. In other words, the difference between a correlation of .20 and .30 is not as meaningful as the difference between a correlation of .80 and .90.

15.4 Using the Pearson *r* for Hypothesis Testing

The Null and Alternative Hypotheses

The statistic, r, is based on a *sample* of bivariate scores. It is an estimate of ρ, a *population* of bivariate scores. The statistic, r, can be used to make a number of inferences about ρ. For example, confidence intervals for ρ can be established,

two *r*'s can be compared to see if they have been drawn from populations having different ρ's, a given *r* can be tested to see if it is different from a specified value of ρ (e.g. +.50), and *r* can be tested to determine if the population correlation is different from 0. This section only addresses the most common form of hypothesis testing: Is ρ different from 0? When asking the question, "Is the population correlation different from 0?," the null and alternative hypotheses are stated as

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

Sampling Distributions

Whenever statistically testing a null hypothesis, whether it is concerning means, variances, or a correlation, a sampling distribution is required. Recall that a sampling distribution is a theoretical distribution made up of the statistic that is being tested (see Chapter 7). If we are testing to see whether a sample mean differs from a specified population mean, then the appropriate sampling distribution is made up of means. If we are testing the difference between two sample means, then the relevant sampling distribution is a distribution of differences between means. Although a sampling distribution is never actually constructed, the characteristics of the appropriate sampling distribution are known, provided certain assumptions are met.

A statistical test of significance asks the question, "What is the probability that we would obtain this sample statistic by chance when the null hypothesis is true?" If the probability is low (e.g. less than five times out of 100), then the null hypothesis is rejected, and we cautiously suggest something other than chance is involved. Given this line of reasoning, testing the null hypothesis that $\rho = 0$ involves computing *r* from a sample of scores and determining how unlikely it is that the obtained *r* would occur if the population correlation was 0.

The appropriate statistical model for testing the significance of an *r* is the *t* distribution. However, we do not need to compute a *t* ratio. A direct method for testing the null hypothesis is to use a table of critical *r's* (see Table A.7). This table is derived from the sampling distributions of *t*; it allows us to make a direct comparison between the size of the sample correlation and the critical value of *r* in the table. As with *t* tests, we will need to determine both the size and the placement of the rejection region, using either a two-tailed or one-tailed test. The same issues discussed in Chapter 8 present themselves here. The size of the rejection region determines the degree of Type I error risk we wish to tolerate. In terms of placement, it is usually considered prudent to split the

rejection region evenly between the two tails of the distribution, that is, to use a two-tailed test.

The following worked problem illustrates how to use the table of critical values to test the null hypothesis that there is no correlation between X and Y in the population. Note: The degrees of freedom for a Pearson r correspond to the number of paired scores minus 2 ($n_p - 2$). Assuming each individual contributes a pair of scores, the number of paired scores equals the number of participants.

■ **Question** *An educational psychologist hypothesizes a relationship between trait anxiety and GPA. A sample of 20 students is randomly selected; the correlation between anxiety and GPA is found to be –.50. Is there statistical evidence that ρ is different from 0?*

Solution

Step 1. State the null and alternative hypotheses. In this case, H_0: $\rho = 0$ and H_1: $\rho \neq 0$.
Step 2. Establish an alpha level. Use $\alpha = .05$.
Step 3. Compute r_{obt}. It is given as –.50.
Step 4. Locate r_{crit} in Table A.7, using $df = n_p - 2 = 18$, two-tailed test. The critical value equals ±.444.
Step 5. Compare r_{obt} with r_{crit}. If the sample correlation falls outside the r_{crit} values, reject the null hypothesis. Since –.50 does fall outside of ±.444, reject H_0: $\rho = 0$.
Step 6. Interpret the findings. "There is statistical evidence to suggest a negative correlation exists between trait anxiety and GPA, $r(18) = -.50$, $p < .05$. Students who are anxious tend to have lower GPAs."

Note that we cannot make a causal statement that trait anxiety *leads* to lower grades. It could be the case that having lower grades leads to a chronic feeling of anxiety. Moreover, there is always the possibility that a third, unmeasured variable, accounts for this correlation. ■

The following worked problem also shows the steps involved in testing H_0: $\rho = 0$. However, this problem begins with raw data. Follow the steps to ensure understanding of the procedural flow in computing r_{obt} and testing H_0: $\rho = 0$.

■ **Question** *A clinical psychologist hypothesizes a correlation between a personality dimension, extroversion/introversion, and depression. Two questionnaires are administered. One measures the personality trait, with higher scores indicating more extroversion. The other questionnaire measures depression, with higher scores reflecting greater depression. Conduct a two-tailed test of r using an alpha of .05.*

Extroversion/introversion: *X*	Depression: *Y*
16	22
14	18
15	20
6	9
3	10
5	3
10	10
2	4
13	15

Solution

Step 1. State the null and alternative hypotheses. In this case, H_0: $\rho = 0$ and H_1: $\rho \neq 0$.

Step 2. Establish an alpha level. Use $\alpha = .05$.

Step 3. Compute r_{obt}.

$$r_{obt} = \frac{n_p(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{\left[n_p(\Sigma X^2) - (\Sigma X)^2\right]\left[n_p(\Sigma Y^2) - (\Sigma Y)^2\right]}}$$

$$r_{obt} = \frac{9(1306) - (84)(111)}{\sqrt{\left[9(1020) - (84)^2\right]\left[9(1739) - (111)^2\right]}}$$

$$r_{obt} = \frac{2430}{\sqrt{(2124)(3330)}}$$

$$r_{obt} = \frac{2430}{\sqrt{7\,072\,920}}$$

$$r_{obt} = \frac{2430}{2659.50}$$

$$r_{obt} = +\mathbf{0.91}.$$

Step 4. Locate r_{crit} in Table A.7, using $df = n_p - 2$ (9 – 2 = 7), two-tailed test. The critical value equals ±.666.

Step 5. Compare r_{obt} with r_{crit}. Since .91 falls outside of ±.67, reject H_0: $\rho = 0$.

Step 6. Interpret the findings. "There is statistical evidence suggesting a correlation between extroversion and depression, $r(7) = +.91$, $p < .05$. Those people who are depressed are more likely to be extroverts; those people who are not depressed are more likely to be introverts." ■

General Considerations in Testing H_0: $\rho = 0$

Hypothesis testing is an inferential procedure. In the context of correlational analyses, a decision is made as to whether the population correlation differs from 0 (the null hypothesis). As with any test of a null hypothesis, two types of decision errors can be made. A Type I error is committed if a researcher concludes that the population correlation ≠ 0, when, in fact, it does (i.e. a true null hypothesis is rejected). A Type II error is made when a researcher cannot conclude that the population correlation ≠ 0, when, in fact, it does not (i.e. a false null hypothesis is not rejected).

The power of a statistical test is the probability that the test will correctly reject a false null hypothesis. In the context of correlation, power is the ability a test has to detect a nonzero population correlation. For a given sample size, as ρ departs from 0, it is easier to obtain an r that leads to rejecting the null hypothesis. Moreover, increasing the sample size decreases r_{crit}, making it easier to detect a nonzero population correlation. Indeed, with very large samples, extremely small population correlations can be detected. Whether a researcher would want to detect such small correlations, however, is questionable. (This issue is further discussed below.)

Rejecting Null Hypotheses and r^2

Refer to Table A.7, which gives the critical values for a direct test of r. On the left-hand column, we will see that critical values are once again linked to the degrees of freedom (df) associated with the inferential test. Now, look down any other column. Notice that r_{crit} becomes *smaller* as the df becomes *larger*. This means that an r_{obt} of a given magnitude may direct us to reject the null hypothesis for one sample size, but not for another sample size. Although the table of critical correlations is incomplete and stops at $df = 100$, complete r_{crit} tables can be easily found online. In addition, statistical computer programs such as SPSS test the null hypothesis for any sample size. With very large samples, it is possible for a correlation to prompt the rejection of the null hypothesis even though the magnitude of the correlation is very small. Under these circumstances, it is especially important that we keep the proportion of shared variance, r^2, in mind. A *statistically* significant correlation may not be *theoretically* and/or *practically* significant. A striking example of how small correlations can achieve statistical but not theoretical significance comes from a study on antismoking attitudes and general prejudice among West Germans (Grossarth-Maticek, Eysenck, & Vetter, 1988).

Table 15.5 lists the types of general prejudice at that time, the correlation with antismoking attitudes, r^2, and the level of statistical significance attained by each correlation. The correlation between anti-Semitism and antismoking prejudice, for instance, is .06, significant at the .001 level. The r^2 is .0036, meaning that well

Table 15.5 Correlations between political prejudice and antismoking attitudes.

Prejudice	*r*	*p*	r^2	Percent shared variance (%)
Anti-Semitic	.06	.001	.0036	.36
Anti-Arab	.05	.001	.0025	.25
Racist	.11	.001	.0121	1.21
Anti-American	.05	.001	.0025	.25

under 1% of the variance in antismoking attitudes is accounted for by variation in the anti-Semitic scores!

How could such small correlations lead to a rejection of the null hypothesis? Well, the sample size was 5977! With such a large sample, the statistical test of the null hypothesis was exceptionally powerful. Obtaining data from 5977 participants is a time-consuming and expensive undertaking. We might ask the question, "Given the effort and expense, how important is it to be able to detect correlations that are so small as to be trivial?" We should not be impressed by the *level of statistical significance;* always look at the magnitude of the correlation and then square it to get r^2, the shared variance.

Box 15.2 Maternal Cognitions and Aggressive Children

Over the past several decades, studies have revealed that aggression is a relatively stable, self-perpetuating behavior (Huesmann and Eron, 1984; Juon, Doherty, & Ensminger, 2006; Olweus, 1979). Aggressive behavior in children is of substantial concern to psychologists because it is predictive of later behavior, including the number and seriousness of criminal convictions (Huesmann & Eron, 1984), substance abuse, unemployment, divorce, and psychiatric illness (Caspi, Elder, & Bern, 1987).

The processes that perpetuate aggression are still unknown. While genetic, physiological, and other constitutional factors most likely contribute to the stability of aggression, research suggests that, in most cases, environmental conditions are probably the most important source of influence. Factors from each of the major systems in which young children interact (e.g. school, family, peers, media, etc.) have been implicated as influencing the development and maintenance of aggression (Slaby & Roedell, 1982).

Miller (1990) was interested in the relationship between a particular aspect of family life and the aggressiveness of children. It was hypothesized that maternal cognitions would be correlated with children's aggression. More specifically,

Miller expected to find a correlation between how dissatisfied mothers are with their children and the aggressiveness of their children. Although Miller measured numerous maternal attitudes and behaviors to test competing models of aggression, we will consider only a couple of simple correlations for the purposes of illustrating the concepts discussed in this chapter.

To assess childhood aggression, a peer-nomination measure was used; each child's aggression score was derived from the reports of a sample of their classmates. The children were asked to name as many other children in the class as they wished who behaved in a certain way (e.g. "Who pushes or shoves children?"). The aggression score for a given child was the percentage of times they were nominated by classmates on ten aggression items.

A second variable of interest was mothers' dissatisfaction with their child. A mother who scored high on the dissatisfaction questionnaire was the one who complained that her child "is too forgetful," "doesn't follow directions," and "wastes too much time." Miller was also curious to learn if there was a difference in this relationship based on the gender of the child. The following table presents the correlations between maternal dissatisfaction and childhood aggression found by Miller.

Correlations Between Maternal Dissatisfaction and Childhood Aggression

Boys ($n = 54$)		Girls ($n = 45$)	
r	***p***	***r***	***p***
.33	<.01	.31	<.05

A positive correlation means that higher measures of maternal dissatisfaction correspond with greater aggressiveness in the child. Alternatively stated, the more aggressive children have mothers who were more dissatisfied with them. The correlations for both boys and girls are statistically significant – boys: $r(54) = .33$, $p < .01$; girls: $r(45) = .31$, $p < .05$.

Considering these two correlations, an issue arises; one that frequently arises when doing correlational research. Is there a causal connection between the variables? Given only the information that has been presented from Miller's study, there is no way to know if a child's aggression is influenced by maternal dissatisfaction, or if maternal dissatisfaction is influenced by the child's aggression. Furthermore, it is possible that some third variable accounts for the relationship between both variables. Perhaps the level of aggression *exhibited by the mothers* influences both their dissatisfaction with their children *and* the level of aggression demonstrated by their children.

15.5 Factors That Can Create Misleading Correlation Coefficients

The definition of a correlation coefficient can be deceptively simple: It is a measure of the strength of association between two variables. However, several factors can affect the size of the correlation; these factors can hide the real nature of the relationship between the variables. This section addresses some of the issues that arise when interpreting correlations.

Restricted Range

One problem that can arise when calculating correlations concerns the range of each measure. Each measure will have a potential range of scores and an obtained range of scores. For instance, a scale that measures kindness might range from 20 to 60. If, for example, due to sampling error, the range of scores obtained from a group of participants ranges from 39 to 60, the entire range of test scores is not reflected in the sample. The range of scores is restricted. If we correlate the kindness scale with some other measure, a problem arises. *Samples with restricted ranges tend to underrepresent population correlations.*

In Figure 15.5, the correlation, based on the entire plot, is quite high, about +.80. Now look at just the plot within the inset in the upper right corner.

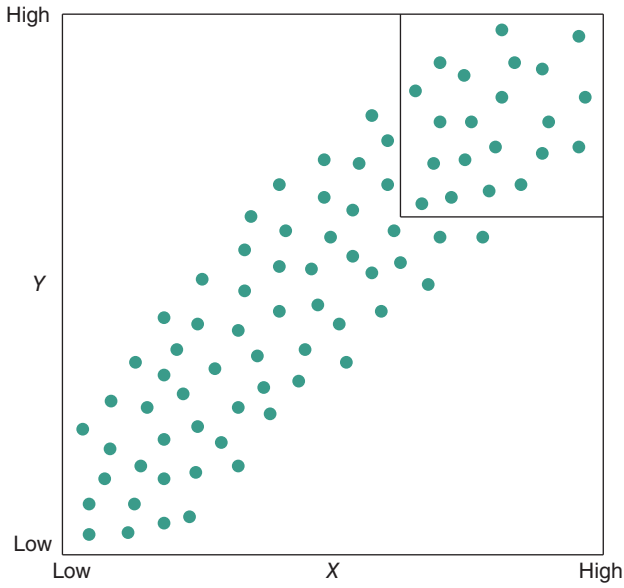


Figure 15.5 The true strength of association between *X* and *Y* is underestimated when the range of scores is restricted.

Those scores are restricted to the upper ranges of X and Y. The plot in the inset reflects a much lower correlation.

The problem of restricted ranges can arise when examining the relationship between Graduate Record Examination (GRE) scores and the GPAs obtained by students at the end of the first year of graduate school. Admissions committees obviously do not admit students to graduate school at random. They favor students who have performed very well on the GREs (in addition to other criteria). As a result, the group of successful applicants will not show much variability on GREs, at least not as much variability as the scores of all students applying to the program. Likewise, students' GPAs at the end of the first year tend to be restricted to the upper grades. What would happen if we computed a correlation between GRE scores and the GPA obtained at the end of the first year of graduate school? The correlation would be spuriously (artificially) low. Failing to consider the range restrictions on the first-year graduate student data, an admissions official could erroneously conclude that GRE scores have no relationship to subsequent GPA and therefore, should not be used as a criterion of admission. Thankfully, graduate admissions committees are aware of this issue; no one is surprised that the correlation between GRE scores and graduate student GPA is low.

A Nonlinear Relationship Between *X* and *Y*

The formula for the Pearson correlation discussed in this chapter is only appropriate when there is a linear relationship between X and Y. The use of the Pearson formula becomes less appropriate as the relationship between X and Y increasingly departs from linearity. In these situations, the Pearson r will underestimate the correlation. Other correlational techniques found in advanced statistical manuals can be used to capture the strength of a relationship when there is nonlinearity.

Figure 15.3 depicts an example in which there was a curvilinear relationship between arousal and task performance. It was pointed out that using the Pearson formula would yield a spuriously low correlation, when, in fact, the association between the two variables was high. We need to protect ourselves from misapplying the Pearson formula; fortunately, to do so is not difficult. Simply examine the scatter plot. If the data can be circumscribed with an oval, we are safe in using the Pearson formula. Figure 15.6 illustrates some nonlinear relations between X and Y. Note that nonlinearity is a matter of degree. Nonlinearity is not confined to a plot that takes the shape of a $\cup$ or $\cap$. Nonlinearity is reflected by any plot that shows a curve. The Pearson formula for r presented in this chapter should not be used with bivariate distributions that have any of the shapes shown in Figure 15.6.

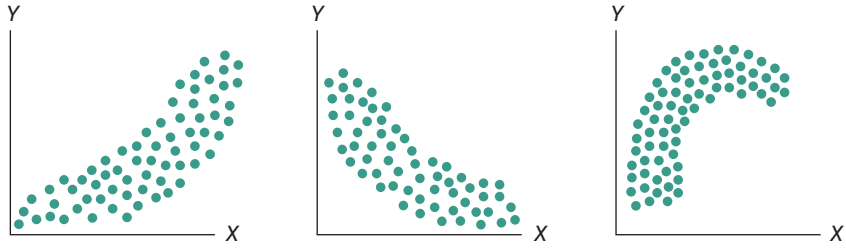


Figure 15.6 Some nonlinear relations between *X* and *Y*.

The *X* and *Y* Distributions Are Skewed

When the *X* and *Y* distributions are skewed in opposite directions, correlation computations can generate misleading, artificially low, values. Under normal circumstances, we might view a correlation of .50 as moderate. However, if the distributions are oppositely skewed, it is conceivable that the highest possible correlation calculable to be around .60. If this were the case, an *r* of .50 would be considered very strong. Consider the following two frequency distributions.

Score on *X*	0	1	2	3	4	5	6	7	8	9	10
Frequency	25	14	13	14	12	11	6	4	2	3	1
Score on *Y*	0	1	2	3	4	5	6	7	8	9	10
Frequency	0	1	3	2	2	5	3	10	15	20	44

The *X* distribution is positively skewed and the *Y* distribution is negatively skewed. To understand how this *bivariate* distribution places a ceiling on the potential size of the correlation, it is helpful to think in terms of the *z* score formula for correlation:

$$\rho = \frac{\Sigma(z_X z_Y)}{N_p}$$

With relatively few scores *above* the mean of the positively skewed *X* distribution, and so few scores *below* the mean of the negatively skewed *Y* distribution, it is impossible for every positive z_X score to be associated with a positive z_Y score. For this reason, the correlation for this sample could never approach +1 or −1. One tactic that could be taken would be to reverse the scoring for one of the variables. Now the skewness of the two variables is in the same direction.

Another way to address the problem would be to transform the raw data (see Section 4.6). Mathematical steps like these can help to normalize the shape of a distribution. An in-depth discussion of transformation techniques is beyond the scope of this introductory-level textbook; consult an advanced statistics resource for more information.

The Use of Extreme Groups

A common type of correlational design, particularly in the field of personality research, is to compare extreme groups with respect to some variable. For example, suppose we are interested in comparing Type A and Type B individuals with respect to interpersonal dominance. To do this, we administer a questionnaire that provides a continuous measure of the degree to which an individual exhibits the characteristics of the Type A personality. We later select only those individuals who scored in the top 10% (Type A's) and the bottom 10% (Type B's). Now we administer our dominance questionnaire to the members of these extreme groups and correlate the Type A/B scores with dominance.

This methodological approach will typically yield a correlation between X and Y that is larger than would be found if the entire population were used. If, instead of selecting only those participants with very high and very low scores, we included *all* of the participants that had been administered the questionnaire, the resulting correlation would most likely have been lower. The association between X and Y is oftentimes weaker in the midrange of the distributions. Figure 15.7 depicts this point.

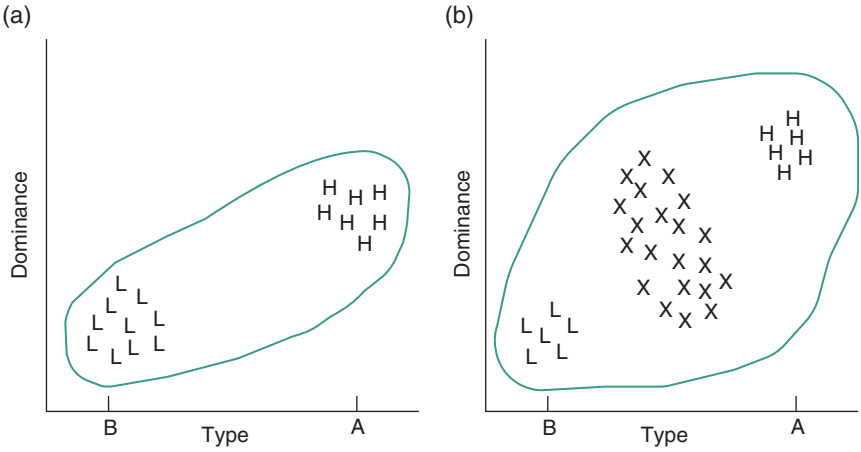


Figure 15.7 Using extreme groups increases the magnitude of the correlation between X and Y. The oval surrounding scatter plot (a) is narrower than the one surrounding scatterplot (b).

Figure 15.7a shows that the oval is fairly narrow when using only extreme groups. When the entire range of the Type A/B scores is used, the oval becomes more circular (Figure 15.7b). Because using only extreme groups can leave a misleading impression regarding the strength of a correlation across the span of a relationship, some statisticians warn against this type of methodology. However, sometimes the theory being explored is based on the extremities of a population. To be safe, we should use caution when interpreting correlations based on data drawn only from the ends of distributions.

The Effect of an Extreme Score

A data point that stands off by itself, whether it be abnormally high or low, is called an outlier. An outlier can create a spuriously low or high correlation. This problem most arises when the sample size is small. A simple example will illustrate the point. The only difference between sample *A* and sample *B* is the *Y* score of the last participant; it has been tripled in sample *B*.

Sample *A*		Sample *B*	
X	*Y*	*X*	*Y*
2	4	2	4
3	4	3	4
5	6	5	6
9	8	9	8
3	3	3	9

The *r* between *X* and *Y* in sample *A* is +.94. Higher *X* values strongly correspond with higher *Y* values. The creation of one outlier in sample *B* has dropped *r* to +.49.

Wainer and Thissen (1976) presented a bivariate distribution of heights and weights. The correlation for 25 participants was computed as +.83. However, they showed that if a transcription error was made, and the height and weight values for just one participant were switched, the correlation could change dramatically (–.26!).

Outliers present a dilemma for researchers. There is a strict rule in research that investigators should not simply discard data because it is inconsistent with their expectations. It is a good rule because it guards against biasing the outcome of a study to fit preconceived notions. Without the rule, confidence in the findings of empirical research would drop dramatically.

The conventional approach to the extreme score problem is to analyze the data with and without the outlier, noting the difference it makes in the type of conclusion one would draw from the results. Sometimes an outlier will have

only a small effect on the analysis, particularly when the sample size is sufficiently large (yet another argument for large samples). If the outlier does make a difference, the researcher may be justified in basing conclusions on the analysis conducted *without* the outlier. Since that one participant's score is so aberrant, the researcher might suspect that the score is unreliable. It could be a measurement or transcription error. On the other hand, it could be accurate data. Whatever the case, outliers need to be carefully examined. Unusual data can sometimes end up becoming very important. For example, interviewing the participant responsible for the outlier may lead to another hypothesis for a subsequent study.[5]

The correlation coefficient is one of the most useful measures for assessing the degree to which two variables are related. However, since there are numerous factors that affect the interpretation of the correlation, it is imperative that we carefully "dig into the data." Keep in mind r^2 when r is statistically significant, and pay close attention to the scatter plot to check for linearity, skewness, restricted ranges, extreme samples, and the presence of outliers.

15.6 How to Present Formally the Conclusions of a Pearson *r*

The proper reporting of Pearson r findings is similar to the proper reporting of t test findings. When reporting a significant Pearson r, we must include the degrees of freedom, the value of the obtained r (usually with the proper valence and without the "0" preceding the decimal), and the alpha level used to make our decision. For instance, "Statistical evidence suggests a relationship exists between shoe size and height, $r(6) = +.94$, $p < .05$." A failure to reject the null might read, "There was no statistical evidence found to suggest a relationship between shoe size and height, $r(6) = +.04$, *n.s.*" A measure of shared variance (r^2) can be added to a rejected null hypothesis if needed.

Many other principles common to the proper reporting of all types of statistical findings were first laid out in Section 8.8. Please consult this portion of the text for more general information about the proper reporting of statistical findings.

5 A note of caution is in order. Outliers should be examined irrespective of the implications they have on our findings. Discarding an outlier *only* when it negatively affects our experimental hypothesis potentially biases our findings. To be fair, we should attend to outliers irrespective of the magnitude of the correlation.

Summary

A correlation is a measure of the strength of association between two variables. The distribution of pairs of scores is called a bivariate distribution. The correlation based on a sample is symbolized as *r;* the population coefficient is symbolized as ρ.

Be careful not to imply causality when interpreting correlations. This is a common interpretive mistake. However, some correlational coefficients may imply causality but only if the data has been gathered experimentally. The basis for making causal statements never resides with the type of analysis, but rather with the methodology used to collect the data. "Correlational *design* does not imply causation" is a more accurate motto than "correlation does not imply causation."

A correlation coefficient is represented by a number that ranges from +1 to −1; the higher the coefficient's absolute value, the stronger the association between the two variables. An r of +.60 reflects a strength of association as strong as an r of −.60. If higher values of one variable are associated with higher values of the other variable, then the correlation is described as "positive." If higher values of one variable are associated with lower values of the second variable, then the correlation is described as "negative."

When a bivariate distribution is plotted on a graph, it is called a scatter plot or scatter diagram. The scatter plot provides a great deal of information about the relationship between two variables. The magnitude of the correlation can be estimated by looking at the general shape formed by the points. The size of the correlation is estimated by examining the width of an oval used to envelope the data: The more narrow the oval, the higher the correlation. If the points have no trend and are best contained within a circle, the correlation is close to zero; that is, the variables are unrelated.

Not only does the scatter plot indicate the strength of association between X and Y, but also it reveals the nature of the correlation. If the enveloping oval slopes from the lower left to the upper right, the correlation is positive. If the oval slopes from the upper left to the lower right, then the correlation is negative. The degree of the slope (e.g. gradual vs. steep) is *not* indicative of the correlation strength.

The scatter plot can also reveal when the oval arches or forms a $\cup$ or $\cap$. Plots shaped like this reveal a nonlinear relationship between X and Y. The formulas for calculating a linear correlation are different from those used to calculate a correlation based on a curvilinear relationship.

Raw scores above the mean of a distribution transform to positive z scores, and raw scores below the mean transform to negative z scores. If X and Y are positively correlated, then positive z scores of variable X will be paired more often with the positive z scores of variable Y; the negative z scores of the X distribution will be paired more often with the negative z scores of the

Y distribution. If *X* and *Y* are negatively correlated, then the positive *z* scores of *X* will be paired more often with the negative *z* scores of *Y*, and vice versa. This is the underlying logic of the *z* score formula for the correlation coefficient. Most often, however, when sample correlations are hand calculated, the computational formula (Formula 15.2) is used.

The coefficient of determination, r^2, is a measure of the amount of variation associated with the *Y* variable that is accounted for by variation in the *X* variable. (This can also be stated the other way around; it is a bidirectional concept.) Shared variance is usually stated as a percentage. If the correlation between two measures is .60, then the amount of variance held in common is $.60^2 \times 100 = 36\%$.

We can compare two correlation coefficients and see that one is larger than the other. However, we cannot say that an *r* of +.80 is twice as large as an *r* of +.40 or that .50 is half the size of a perfect correlation. Any effect size comparisons should be made in terms of shared variance, r^2.

The test of the null hypothesis for *r* asks the question, "Is the magnitude of *r* sufficiently large to conclude that ρ is not 0?" For any inferential test, a theoretical sampling distribution is needed. The correct sampling distribution for using *r* to test a null hypothesis is a *t* statistic, based on a transformation of *r*'s. The transformation has the effect of normalizing the sampling distribution of correlations. When testing the significance of a correlation, the most common null and alternative hypotheses are H_0: $\rho = 0$ and H_1: $\rho \neq 0$.

Several factors can affect the size of the correlation. These factors can hide the real nature of the relationship between the variables being correlated.

When the distribution of *X* and/or *Y* is restricted, then the correlation is likely to be spuriously low. Familiarity with the potential range of values is important when interpreting correlations.

The use of the Pearson formula is inappropriate as the relationship between *X* and *Y* departs from linearity. In these situations Pearson *r* will underestimate the relationship between *X* and *Y*.

When the *X* and *Y* distributions are skewed in opposite directions, correlation computations can generate misleading, artificially low, values. Various mathematical transformations can be considered to change the shape of either the *X* and/or *Y* distributions and reduce the underrepresentation of the relationship.

A common type of correlational design is to compare extreme groups with respect to some variable. This methodological approach typically yields a correlation between *X* and *Y* that is larger than what would be found if normal sampling was used. Exercise caution when interpreting correlations from designs that use extreme groups.

A data point that stands off by itself is called an outlier. Outliers can create a misleading *r*, especially when the sample size is small.

Using Microsoft® Excel and SPSS® to Calculate Pearson *r*

Excel

General instructions for data entry into Excel can be found in Appendix C.

Data Entry

Enter the bivariate data into two adjacent columns, being sure to keep the data from each participant together in the same row. Label the columns appropriately. (See Figure 15.8 for an example.)

Data Analysis

1) Excel has built-in programs for many inferential tests, including a Pearson *r*. To access it, click on the Data tab on the top menu and then click **Data Analysis**. (Some versions of Excel have a "Tools" tab. The Data Analysis function may be under this tab.) If this option is not found, the Data Analysis ToolPak needs to be installed. See Excel instruction materials for how to install this feature.
2) With the Data Analysis box open, select **Correlation**.
3) Input the data range by dragging over the entire data set and placing those coordinates into the **Input Range** box. (If we include the labels in the data range, make sure to click the **Labels** box to exclude those cells.)
4) Decide on an Output option. The default is to place it on a separate worksheet.
5) Click **OK**.
6) A correlation grid is produced. Each variable is listed down both the left-side column and across the top of the grid. (Excel can run multiple correlations at once. For example, if we had three variables and included them in the analysis, both the left-hand column and the top row would have all three variables listed.) Down the diagonal spine of the correlation grid will be the value 1, representing the correlation between a variable and itself. The correlations of interest can be found by locating the coordinate between one variable on the left-hand column and the other across the top row. The table is redundant showing each correlation from each perspective. (See Figure 15.8 for a worked example.)
7) Excel does not test the null hypothesis that $\rho = 0$. We will need to use a critical value table (e.g. Table A.7 in the Appendix A) to find r_{crit} and make our decision regarding the null hypothesis.

Shoe size	Height
72	10
66	9
74	13
68	10
63	7
70	10
73	12
67	9

	Shoe size	*Height*
Shoe size	1	
Height	0.93489	1

Figure 15.8 A worked example using Microsoft Excel to calculate a Pearson *r*.

SPSS

General instructions for inputting data into SPSS can be found in Appendix C.

Data Entry

In SPSS, each row of the data file represents a participant. Since bivariate data is used in calculating a Pearson *r*, create a series of variables within **Variable View** corresponding to the variables measured. Then, go to **Data View** and input the data, being careful to keep the values from each participant within a given row. See Figure 15.9 for an example.

Figure 15.9 An example of entered data for a Pearson *r* calculation in SPSS.

	Shoe size	Height
1	72	10
2	66	9
3	74	13
4	68	10
5	63	7
6	70	10
7	73	12
8	67	9

Data Analysis

1) Click **Analyze** on the tool bar, select **Correlate**, and then click **Bivariate**.
2) Use the arrow key to move the variables of interest into the **Variables** box.
3) The default correlation coefficient calculated is the Pearson; leave this box checked. Make a selection regarding the critical *r* value to be calculated, one tailed or two tailed.
4) If descriptive statistics are of interest, open the **Options** box in the upper right corner, and click the **Means and standard deviations** option.
5) Click **Ok**.

6) If descriptives were asked for, the first box will present the means, standard deviations, and sample size of all selected variables. The next box is the correlation grid box simply labeled **Correlations**. Each variable is listed both down the left-side column and across the top of the grid. (SPSS can run multiple correlations at once. For example, if we had three variables and moved all of them into the **Variables** box, both the left-hand column and the top row would have all three variables listed.) Down the diagonal spine of the correlation grid will be the value 1, representing the correlation between a variable and itself. The correlations of interest can be found by locating the coordinate between one variable on the left-hand column and the other across the top row. The table is redundant showing each correlation from each perspective. Within each correlation box can also be found the probability of getting a Pearson *r* of that size if $\rho = 0$ [**Sig. (2-tailed**)] as well as a count of the number of paired scores (**N**). If the significance value is equal to or less than .05, there is statistical evidence to reject the null hypothesis. (See Figure 15.10 for a worked example.)

Correlations

Correlations

		Shoe size	Height
Shoe size	Pearson correlation	1	.935**
	Sig. (2-tailed)		.001
	N	8	8
Height	Pearson correlation	.935**	1
	Sig. (2-tailed)	.001	
	N	8	8

** Correlation is significant at the 0.01 level (2-tailed).

Figure 15.10 An output table from a worked example using SPSS to calculate a Pearson *r*.

Key Formulas

The *z* score formula for the population correlation

$$\rho = \frac{\Sigma(z_X z_Y)}{N_p} \quad \text{(Formula 15.1)}$$

Computational formula for Pearson *r*

$$r = \frac{n_p(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{\left[n_p(\Sigma X^2) - (\Sigma X)^2\right]\left[n_p(\Sigma Y^2) - (\Sigma Y)^2\right]}} \quad \text{(Formula 15.2)}$$

Key Terms

Correlation coefficient
Bivariate distribution
Pearson product-moment correlation coefficient
Scatter plot
Linear relationship
Curvilinear relationship
Coefficient of determination
Shared variance (or common variance)

Questions and Exercises

1 What is the distinction between a correlational and experimental design? Why does an experimental design offer the potential for making causal statements about the relationship between two variables?

2 Which statement is most accurate?
a Correlational design does not imply causation.
b Correlation does not imply causation.
c Correlation implies causation.
d Correlation equals causation.

3 Describe what is meant by the term "bivariate distribution."

4 Provide examples of variables that are positively correlated.

5 Provide examples of variables that are negatively correlated.

6 Which of the following is the strongest legitimate correlation? Why?
a −1.14
b −.69
c 1.09
d .58

7 Describe what a scatter plot of a +1 or −1 correlation would look like.

8 What information can be gleaned by examining a scatter plot? What feature of a scatter plot is uninformative?

9 Given the following population of z scores, what is the correlation between X and Y?

z_X	z_Y
−0.32	−0.56
−0.10	0
0.42	−0.12
0	0.68

10 Given the following population of z scores, what is the correlation between X and Y?

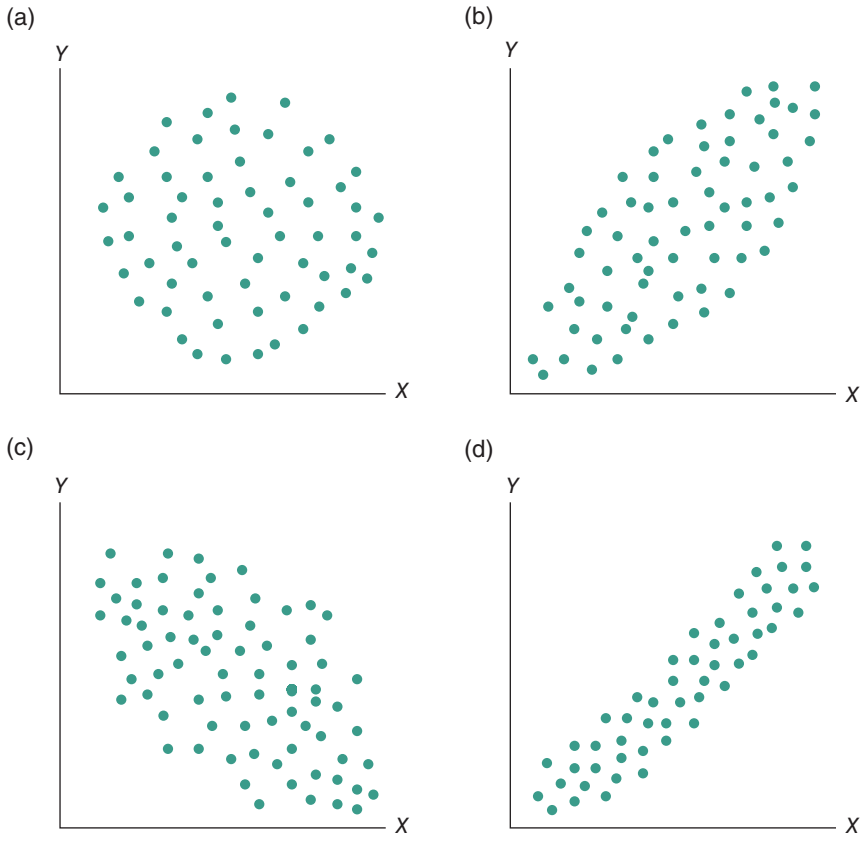
z_X	z_Y
1.48	−0.75
−0.31	−1.14
−1.62	1.55
0.45	0.34

11 For the following correlations and degrees of freedom, what are the r_{crit}'s for $\alpha = .05$ and $\alpha = .01$ (two-tailed tests)? (Note: Use an online table when Table A.7 is incomplete.) In each case, should the null hypothesis be rejected?

			r_{crit} at $\alpha = .05$	Reject H_0?	r_{crit} at $\alpha = .01$	Reject H_0?
a	$r = .39$	$df = 100$	_____	Y or N	_____	Y or N
b	$r = -.47$	$df = 21$	_____	Y or N	_____	Y or N
c	$r = -.09$	$df = 11$	_____	Y or N	_____	Y or N
d	$r = .44$	$df = 6$	_____	Y or N	_____	Y or N
e	$r = -.62$	$df = 12$	_____	Y or N	_____	Y or N
f	$r = .93$	$df = 29$	_____	Y or N	_____	Y or N

12 Draw a scatter plot reflecting the following correlations.
- **a** A moderately strong positive correlation.
- **b** A very strong negative correlation.
- **c** A very weak positive correlation.
- **d** A curvilinear correlation.

13 Estimate the *r* for each of these scatter plots.



14 For the following correlations and *df,* what are the critical *r's* when using a 5 and a 1% alpha value (two-tailed test). (Note: Use an online table where Table A.7 is incomplete.) In each case, should we reject or fail to reject the null hypothesis at the 5% level? At the 1% level?

a $r = .67$, $df = 24$
b $r = .39$, $df = 24$
c $r = .89$, $df = 12$
d $r = .74$, $df = 18$
e $r = .45$, $df = 29$
f $r = .95$, $df = 7$
g $r = .62$, $df = 13$
h $r = .24$, $df = 100$

15 For the following data set:
- **a** Calculate r and conduct a two-tailed test of the null hypothesis ($\alpha = .05$).
- **b** Specify the null and alternative hypotheses.
- **c** What is r_{crit}?
- **d** Reject or fail to reject the null hypothesis?
- **e** What percent of the variance of Y is accounted for by the variance of X?

X	Y
6	4
7	5
7	6
4	6

16 For the following data set:
- **a** Calculate r and conduct a two-tailed test of the null hypothesis ($\alpha = .05$).
- **b** Specify the null and alternative hypotheses.
- **c** What is r_{crit}?
- **d** Reject or fail to reject the null hypothesis?
- **e** What percent of the variance of Y is accounted for by the variance of X?

X	Y
6	3
9	7
7	6
10	9

17 For the following data set:
- **a** Calculate r and conduct a two-tailed test of the null hypothesis ($\alpha = .05$).
- **b** Specify the null and alternative hypotheses.
- **c** What is r_{crit}?
- **d** Reject or fail to reject the null hypothesis?
- **e** What percent of the variance of Y is accounted for by the variance of X?

X	Y
10	12
9	6
11	10
13	13

18 For the following data set:
- **a** Calculate r and conduct a two-tailed test of the null hypothesis ($\alpha = .05$).
- **b** Specify the null and alternative hypotheses.
- **c** What is r_{crit}?
- **d** Reject or fail to reject the null hypothesis?
- **e** What percent of the variance of Y is accounted for by the variance of X?

X	Y
1	3
2	4
4	5
6	5
7	7

19 Provide an example of a correlational design in which X is a personality variable and Y is a measure of observed behavior.

20 Provide an example of an experimental design in which X is a medicinal manipulation and Y is a measure of observed behavior.

21 Why is a causal interpretation prohibited for the analysis of Exercise #19 but allowed for the analysis of Exercise #20?

22 Answer these questions for the following data set.
- **a** Calculate r and conduct a two-tailed test of the null hypothesis ($\alpha = .05$).
- **b** Specify the null and alternative hypotheses.
- **c** What is r_{crit}?
- **d** Reject or fail to reject the null hypothesis?
- **e** What percent of the variance of X scores is explained by the variance of Y scores?

X	Y
1	8
3	7
3	6
5	5
6	4

23 Draw the scatter plots for the data of problems 15, 16, 17, and 22.

24 For a group of 75 participants, $\Sigma(z_X z_Y)$ is 64. What is ρ?

25 In what way might the range of scores sampled influence the size of the correlation?

26 In what way might the use of extreme groups affect the correlation?

27 If we use the Pearson r to estimate the population correlation between X and Y when X and Y are related nonlinearly, what is the likely result?

28 A sports psychologist is interested in the relationship between how many weeks people exercise and their resting heart rate. Using the following data, answer these questions.

a What is r?
b What are the null and alternative hypotheses?
c What is r_{crit} for a nondirectional test when $\alpha = .05$?
d What percent of the variance of resting heart rate scores is accounted for by the number of weeks of exercise?
e Should we reject or fail to reject the null hypothesis?

Weeks of exercise	Resting heart rate
2	82
4	78
8	72
14	66
10	66
9	70
9	69

29 A school psychologist hypothesizes a relationship between IQ and number of siblings. Please use the data below to answer the following questions.

a What is r?
b What are the null and alternative hypotheses?
c What is r_{crit} for a nondirectional test when $\alpha = .05$?

d What percent of the variance of IQ scores is accounted for by the number of siblings?

e Should we reject or fail to reject the null hypothesis?

Number of siblings	IQ
8	123
3	100
1	90
4	111
2	102
0	95

30 A psychologist is interested in the strength of association between age and performance on a certain task requiring motor skills. Plot the scatter diagram of the following data and decide on the most reasonable course of action for testing the hypothesis.

Age in years	Number of errors
6	23
7	19
8	17
9	16
10	16
11	18
12	18
13	19
14	20
15	22

Computer Work

31 A research team is interested in the relationship between smoking and illness. They randomly select a sample of 13 smokers in a large office and ask them to report the average number of cigarettes they smoke per day. They then obtain the company records that monitor the number of sick days each employee has taken over the past six months of employment. Please

calculate a Pearson r and make a decision about the null hypothesis. Interpret the finding. If warranted, indicate the value of r^2.

Number of cigarettes	Number of sick days
11	1
10	1
26	5
15	3
9	2
16	2
20	2
8	1
3	0
24	4
21	6
5	0
14	3

32 A psychologist is interested in the relationship between intelligence and word processing speed on a keyboard. Twelve university students are randomly selected and measured in both domains. Below are the gathered data. Please calculate a Pearson r and make a decision about the null hypothesis. Interpret the finding. If warranted, indicate the value of r^2.

Intelligence	Word processing speed
108	28
96	46
90	55
111	40
119	34
105	38
98	57
93	47
117	48
127	73
101	56
103	48

33 Baron, Logan, and Kao (1990) studied the relationship between student dentists' perceptions of their patients' discomfort and the patients' perceptions of their own discomfort. Discomfort was defined as a combination of anxiety, pain, and distress (with low numbers indicating low discomfort). Discomfort ratings were obtained under two conditions: during drilling and during the rubber dam placement. (The rubber dam is a thin rubber sheath attached to a metal frame. It fits around the tooth, which isolates it and prevents debris from being swallowed. Placement of the rubber dam requires more of the dentist's attention than a simple filling.)

The following data set is hypothetical. The numbers are selected so that the correlations will lead to conclusions that are consistent with what was found by the authors. Please calculate a Pearson *r* and make a decision about the null hypothesis.

Discomfort ratings during drilling

Dental students	Patients	Dental students	Patients
8	6	3	3
6	9	9	7
3	1	7	8
1	4	6	9
5	5	2	8
4	6	5	7
8	8	6	6
7	6	3	2
9	6	1	1
2	3	5	7
1	1	6	9
6	8	8	8
4	6	9	6

Discomfort ratings during rubber dam

Dental students	Patients	Dental students	Patients
8	6	3	3
6	9	9	7
3	1	7	8

(*Continued*)

(Continued)

Discomfort ratings during rubber dam			
Dental students	**Patients**	**Dental students**	**Patients**
1	4	6	9
5	5	2	8
4	6	5	7
8	8	6	6
7	6	3	2
9	6	1	1
2	3	5	7
1	1	6	9
6	8	8	1
4	6	9	4

34 Carrie (1981) investigated the relationship between a biological female's symptomatic reports during pregnancy and menstruation and the association of these reports with the general tendency to report psychological and physical symptoms. Among the findings was the fact that there is a relationship between the number of symptoms experienced during menstruation and the number of symptoms reported during pregnancy. The following raw data are hypothetical yet will give a correlation value that is consistent with what was found in the study by Carrie. Please calculate a Pearson *r* and make a decision about the null hypothesis.

Hypothetical questionnaire scores	
Last menstruation symptoms	**Last pregnancy symptoms**
93	87
75	64
34	78
23	55
76	43
34	45
21	20
34	54
60	60
45	82

(Continued)

Hypothetical questionnaire scores	
Last menstruation symptoms	**Last pregnancy symptoms**
67	67
50	48
89	72
61	68
56	45
82	75
45	34
53	55
71	50
59	90
90	56
43	62
49	32

16

Linear Regression

16.1 The Research Context

Regression is a set of statistical procedures that build on the concepts of correlation presented in Chapter 15 and allow a researcher to use information about one variable to *predict* the value of a second variable. The idea of using statistical techniques for prediction purposes is new territory and is uniquely associated with this chapter. On many occasions, social and behavioral scientists would like to make predictions. Graduate admissions committees, for example, would like to select students who will do well in their school's graduate programs. If a measure such as Graduate Record Examination (GRE) scores is found to correlate with future grade point averages, then an individual's GRE scores can be used to predict that person's subsequent GPA. If a researcher finds a correlation between the number of times prisoners get into fights while incarcerated and the number of domestic quarrels after release, a parole board may be able to predict the level of postrelease, familial fighting. A generation ago, researchers Zullow and Seligman (1990) showed that the outcome of presidential elections could be predicted by examining the content of campaign speeches. They found that the more a candidate dwelled on negative events, the less likely they were to win the election. They concluded that the American voter "places a high premium on the appearance of hope." Any time two variables are correlated, one of them can be used to predict the other.

In addition to these practical problems, researchers often use regression strategies to make behavioral predictions in order to build and test theories. For example, Zullow and Seligman's study not only had obvious practical implications for speechwriters, but also their data had theoretical interest to behavioral and social scientists. Indeed, Zullow and Seligman derived their research hypothesis about election outcomes from work in the area of depression! They noted that people who tend to dwell on negative events are more vulnerable to

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

depression (Zullow, 1984). Furthermore, they found depressed people to be relatively passive, conveying a sense of hopelessness, and more disliked by others. Consequently, Zullow and Seligman predicted that voters would react negatively to candidates who dwell on negative events in a similar manner, that is, with rejection.[1]

However, recent political results in many parts of the world suggest there are limitations to this theory. When this happens, theory construction often requires the introduction of other variables to help build a better theory of behavior. For instance, Combs and colleagues (Combs, Powell, Schurtz, & Smith, 2009), using regression analyses, found that political partisanship predicts people's type of emotional reactions to tragic events in the news. For example, an economic collapse would seem to be bad news regardless of whether a person is a republican or a democrat. Bad economic news, after all, is bad for everyone. However, when people interpreted the negative event as caused by a rival party, their levels of political partisanship predicted increased happiness about it (even while acknowledging being hurt by the negative event!). This more elaborate picture predicts those who are politically partisan will respond positively to negative events *if* they are seen as caused by the rival party. Understanding complex real-world relationships like the one surrounding partisanship, tragedy, and emotion would be next to impossible using experimental designs. For questions like this, correlational designs and regression analyses are welcome tools to the researcher.

16.2 Overview of Regression

The actual prediction is accomplished by using a **regression equation.** For linear regression to be of use to an investigator, the two variables *must* be related (correlated). This precondition makes perfect sense. If high school GPA is uncorrelated with university GPA, then there is no way to use high school GPA to predict the subsequent GPA. Not only must two variables be correlated, but also to use linear regression, the relationship between the variables must be *linear*. In a linear relationship, each time the value of one variable increases, the value of the other variable shows a constant change. If the change in the second variable is not constant, the relationship between X and Y is nonlinear. This quality will be reflected in the scatter plot. Chapter 15 presented numerous scatter plots in which the oval surrounding the majority of data points was curved, indicating a nonlinear relationship between X and Y. The Pearson formula for the correlation is not used when there is a nonlinear relationship between X and Y;

1 Technically, Zullow and Seligman (1990) used more sophisticated analyses than simple regression. However, their data set and the type of question they were researching are consistent with regression.

likewise, the regression methods presented in this chapter cannot be used when *X* and *Y* are related in a nonlinear fashion. Inspecting the scatter plot is very helpful when deciding whether linearity exists.

This chapter considers the use of simple regression in which information about only one variable (called the predictor or independent variable) is used to predict a second variable (called either the predicted, criterion, or dependent variable). The terms *independent* and *dependent* variable have different meanings in the context of regression. In a regression context, the experimenter typically does *not* manipulate the independent variable. It is simply the variable used for prediction purposes; it is the *X*. The dependent variable, that is, the value of *Y* that is predicted for a given *X*, is said to *depend* on the value of *X*. This should not to be understood as implying cause and effect.[2] Despite this confusion, many resources for statistical analysis choose to employ the terms independent and dependent variables. In an effort to avoid confusion, we will not use these terms.

Regression analyses are bidirectional. For instance, if there is a correlation between stubbornness and empathy, regression can be used to predict a person's level of empathy based on their known level of stubbornness. Conversely, a person's stubbornness can be predicted based on their known empathy score. The bidirectionality of the analyses should remind us that causal relationships should not be inferred.

The word "prediction" implies a temporal sequence between the variables; this is potentially misleading. We might be tempted to conclude that a *current* predictor variable is used to predict a *future* predicted variable. However, no futurity is implied with regression. To say that empathy can be predicted from stubbornness is not to suggest that stubbornness occurs first and then empathy later. In fact, the temporal relationship can even be known to run the opposite direction. For example, we could use SAT scores to predict IQ scores even though it is obvious that a person's intelligence is needed to generate the SAT score. To say "*X* predicts *Y*" is merely to say that if a *Y* value is unknown, *X* can be used to make a prediction of it.

In simple regression, only one predictor variable is used. **Multiple regression** employs more than one predictor variable. The inclusion of additional predictor variables usually increases the accuracy of the prediction. This is attractive to researchers. For a recent example, multiple regression was used to determine that a student's level of sexual identity-related distress could be predicted by combining information about a person's public sexual identity, private sexual

2 Regression techniques can be used with experimental designs. Interested readers are referred to advanced statistics texts (e.g. Ryan, 2008) for a discussion of regression analyses in the context of true experiments.

identity, organizational religiosity, intrinsic religiosity, and level of same-sex attraction (Yarhouse, Dean, Stratton, & Lastoria, 2018). Although multiple regression is beyond the scope of this text, it is based of the concepts discussed in this chapter. (See Box 16.1 to learn more about how the regression concept underlies other more complex statistical analyses.)

Box 16.1 Next Steps with Regression Analyses

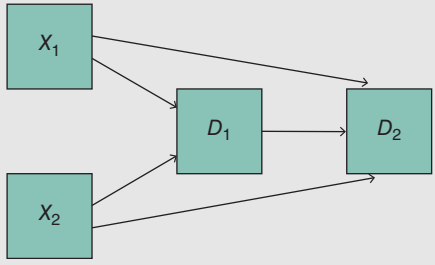
This is an introductory statistics textbook; as a result, we should not be surprised to learn that there is much more to know about each one of the statistical procedures that is being presented. We have already learned there are many versions of follow-up *t* tests and ANOVAs that can be applied to complex research designs. The analytical tool of regression is no different. In fact, there are numerous concepts that build upon the basis of the linear regression idea. This box will briefly identify several of them.

Recall that in linear regression a known relationship between two variables is used to help predict unknown values for one of the variables (Y_p). As was briefly mentioned earlier in the chapter, *multiple regression* uses more than one predictor variable to construct predictor equations for unknown values. If one variable (*X*) can account for a given percentage of the variance in another variable (*Y*), perhaps a third variable (*Z*), also shown to relate to *Y*, can increase the accounted for variance of *Y*. Furthermore, if the use of two predictor variables is better than one, is not three better than two? What about four? And so on. The addition of these other predictor variables introduces new problems and makes the analysis more complicated, but the basic idea remains; use known relationships between variables to help make more accurate predictions of unknown values for a variable of interest.

Factor analysis is a statistical technique designed to take a collection of interrelated variables and shrink them down to the fewest number of actual core concepts. Many times these are inferred variables that are not themselves being actually measured (they are hidden or latent). For instance, we may find that dozens of medically interesting variables seem to be correlated. A factor analysis might show that the dozens of correlated variables may be reducible down to just a handful of core variables, like activity level, social-support network, diet type, and lifestyle. Factor analysis finds hidden patterns in the data by helping to identify a small set of core variables underlying the host of measured variables (these sets of identified hidden variables are oftentimes called "dimensions" or "factors"). These factors can then be listed according to their factor loadings; that is, how much variation in the data they can account for or explain. For instance, we may learn that "type of diet" explains a large percentage of the shared variance among a handful of different correlated medical variables.

It is typically understood that there are two types of factor analysis, exploratory and confirmatory. Exploratory factor analysis takes place when the researcher does not have any preexisting ideas about how many factors may be present behind the nest of correlational data. One of the challenges with exploratory factor analysis is trying to identify and name the underlying concepts that best describe the interrelatability among a set of variables. This is often not an easy determination to make. Confirmatory factory analysis is used to verify a preexisting set of dimensions. These analyses are usually run to test hypotheses derived from exploratory factor analyses about what factors best account for a set of interrelated variables.

Path analysis is often described as a straightforward extension of multiple regression. The purpose of the procedure is to generate size and significance estimates of the hypothesized causal connections between sets of variables. As the name implies, path analysis organizes the predictor variables into a causal sequence of variables where one variable leads to another and so forth. However, rarely are the paths simple relationships where one variable leads to one and only one other variable. Oftentimes the paths determined by the analysis are complex. The path analysis diagram below reflects the idea that two variables (X_1 and X_2) seem to be causing, both directly and indirectly (through D_1), an effect on a given dependent variable (D_2). We can think of path analysis as a form of multiple regression that attempts to develop a causal model between the correlated variables.



Structural equation modeling will be mentioned last because it is a collection of regression-related techniques, including factor analysis and path analysis, which is designed to "model" very complex relationships between numerous variables. Its use is growing in the social and behavioral sciences because of its ability to suggest predictive (and even causal) relationships amid a collection of hidden (latent) variables while relying only on data gathered from observable variables. As computing power increased at the end of the last century, this sophisticated collection of techniques has emerged as a powerful analytical tool for researchers.

Using Correlated Information to Make Predictions

If we were asked to predict the height of a randomly selected American biological male, our best guess would be the mean height for all American males. Assuming the data are normally distributed, most heights will be clustered around this value. It will probably be wrong, but it will most likely be close. It is the best guess we could make. Suppose the mean height is known to be 5 ft 10 in. Now, imagine we were asked to make a prediction for a series of randomly selected American males. Each time we are asked to make a prediction, we should guess 5 ft 10 in. A few times we would be correct, but most of the time we would be incorrect. Moreover, when incorrect, much of the time we would be a little bit incorrect (the selected person might be 5 ft 9 in.), but some of the time we would be very incorrect (the person might be 6 ft 7 in.). Ultimately, using the mean height for all American males would lead to a lot of wrong predictions; but at least our predictions would be unbiased. That is, after we had made a large number of predictions and examined our overall accuracy, we would find that we had overpredicted to the same extent that we had underpredicted.

Now let us change the game. Suppose we are told there is a correlation between the heights of fathers and sons. Furthermore, we are given a table that states the mean height of sons for fathers of a specific height. We might learn that for fathers who are 5 ft 5 in., the mean height of their sons is 5 ft 7 in; for fathers who are 6 ft 6 in., the mean height of their sons is 6 ft 3 in; and so on. Now someone says, "I have selected a biological male. The height of his father is 5 ft 5 in. Guess the height of the son." Now we have correlated information available to aid in prediction. Instead of predicting the mean height of all males, we should predict 5 ft 7 in., the mean height of all males that have fathers who are 5 ft 5 in. Every time we have to make a prediction, we should attempt to obtain the height of the father and use the table to predict the height of the son.

By using correlated information in prediction, we will still make many errors, but the overall error will be smaller than if we had ignored the correlated information. If this strategy seems obvious, then we already have an intuitive understanding of what regression is about. Although regression does not entail using a table to make predictions, an equation accomplishes the same thing. Some score (e.g. height of father) is entered into the regression equation, and the predicted score (e.g. height of son) is computed.

Regression as a Two-Step Process

The regression procedure involves two steps. The first step is to identify two variables that are correlated and to gather this bivariate data. From the bivariate data, we construct a regression equation for later use. The second step involves the application of the regression equation to data from participants *not* included in the original sample. In this second step we only have the value of the X variable available, which is the predictor variable. In this way, the regression equation initially emerges from group data and is then later applied to individuals.

For example, if our ultimate goal is to predict GPAs from performance on the GREs, a large sample of participants is required to obtain both GRE scores and subsequent GPAs. From this data, a regression equation is established. This equation can then be used to predict the GPAs of new applicants. A prediction can be made for *each* applicant based on that person's GRE score.

16.3 Establishing the Regression Line

In linear regression, the regression equation is used to plot a straight prediction line that goes through the middle of the scatter plot. This is called the **regression line.** The term middle, however, has a precise meaning; it will be explained shortly. Formula 16.1 is the formula for a straight line.

Straight line formula

$Y = a + bX$ (Formula 16.1)

In this formula, a is the ***Y* intercept** – that point at which the prediction line crosses the ordinate (i.e. the Y axis) when X is 0. The other constant, b, defines the angle, or slope, of the line.[3] Figure 16.1 shows a scatter plot with several straight lines drawn through the plot. Since each line has a different angle and a different point at which it crosses the Y axis, the constants a and b are different for each line. Yet each line is described by the general equation for a straight line.

The regression equation chooses among an infinite number of straight lines that could be drawn through the scatter plot. A criterion has to be established, which must be met when estimating a and b; then formulas have to be developed

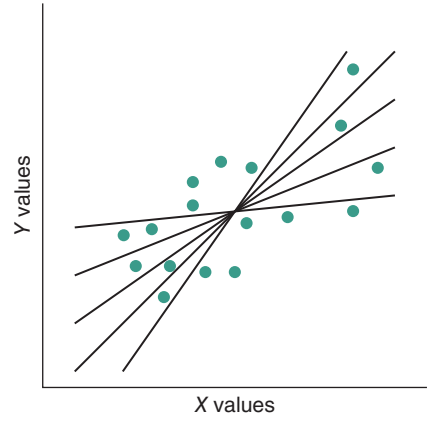


Figure 16.1 Several straight lines drawn through a scatter plot. The regression equation establishes which line, of a potentially infinite number of lines, is the best one to predict a *Y* score given an *X* score.

3 Algebra texts usually use different symbols in the formula for a straight line, e.g. $Y = mX + b$, where m is the slope of the line and b is the intercept.

for a and b that meet the criterion. The least squares method is the most typical method used to determine where the regression line will be drawn through the scatter plot. Before discussing the criterion that is met by the least squares method, it is important that we understand how to interpret a regression line.

Reading the Regression Line: All Predicted *Y*'s Are on the Regression Line

In regression, an X score is used to predict the value of a Y score. Provided there is a correlation between X and Y (which is the only time regression should be used), there is a different Y predicted for each value of X. Every predicted Y (symbolized Y_p) lies on the regression line.[4] To find Y_p for a given X, find the X value on the horizontal axis (abscissa), and draw a line parallel to the vertical axis (ordinate). When that line meets the regression line, another line is drawn at a right angle (parallel to the X axis) until it meets the vertical (Y) axis. The value at which this line intersects the Y axis is Y_p. In Figure 16.2, a Y of 6 would be predicted for anyone who obtained an X of 4.

Reading the Regression Line: The Slope of the Regression Line

The **slope** of the regression line measures the "rise over the run"; how many units the line rises on the Y axis for every one unit moved to the right on the X axis. For instance, if $b = 0.93$, the regression line ascends 0.93 Y units with

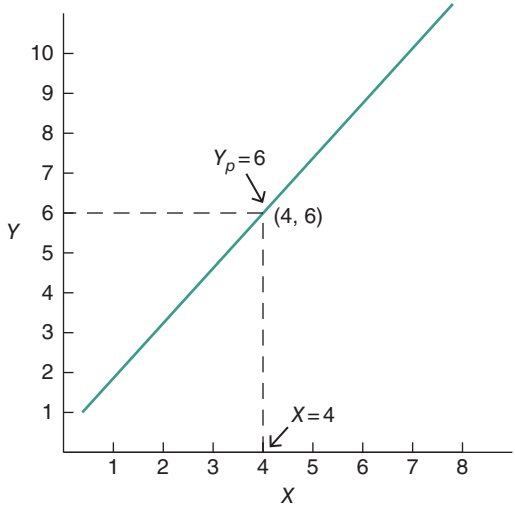


Figure 16.2 Using the regression line to predict Y given $X = 4$.

4 Some textbooks use the symbol Y ("Y prime") instead of Y_p. We will use Y_p to prompt us to think, "predicted Y."

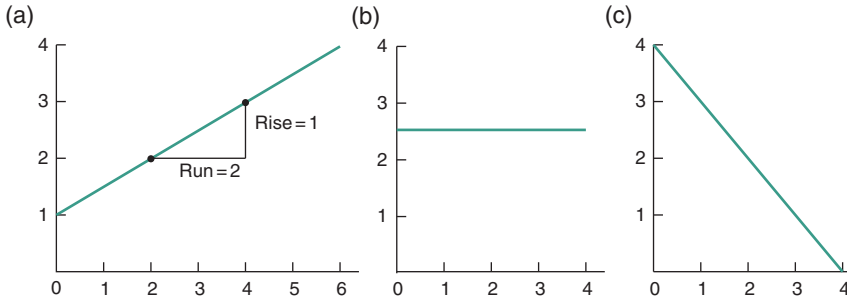


Figure 16.3 The straight lines with different slopes. (a) Slope of 0.5; (b) slope of 0; (c) slope of –1.

each successive unit of X. Suppose for the relationship between height (the Y variable in inches) and weight (the X variable in pounds), a regression line has a slope of +2. This means that an increase of 1 lb corresponds to an increase of 2 in. In addition, a positive slope corresponds to a positive correlation, and a negative slope corresponds to a negative correlation. Larger absolute values of b indicate steeper slopes of the regression line. Figure 16.3 illustrates lines with three different slopes. In Figure 16.3a, the line "rises" one unit on the Y axis for every "run" of two units on the X axis. The slope is 0.5. In Figure 16.3b, the slope of the line neither ascends nor descends. A line that is parallel to the X axis has a slope of 0. In Figure 16.3c, the line descends one unit for every one unit of increase on the X axis. The slope is –1.

The Least Squares Criterion

The criterion that is used to select the best straight line that could be drawn through the scatter plot is called the **least squares criterion.** This criterion assures that the regression line chosen has the least amount of prediction error possible.

If the relationship between X and Y is perfect ($r = \pm 1$), then all data points will lie along a straight line. The regression line will have all the points of the scatter plot on it. Since a correlation of ±1 is extremely rare, there is typically a spread of points surrounding the line; every point not on the line reflects an amount of error. "Error" is the difference between the actual Y score obtained by an individual and Y_p, the score predicted for that person by the regression line.

Figure 16.4 shows a hypothetical regression line for the relationship between GRE scores and subsequent GPA. For purposes of visual clarity, the swarm of data points surrounding the regression line is not depicted. Dan received a GRE score of 650. Next years' predicted GPA for Dan is 3.62. If the GPA turns out to be 3.00, we have missed by 0.62 GPA points. Marcy also scored 650 on the GRE. The same GPA (3.62) was predicted for next year. Marcy later achieved a 4.00 GPA. In this case, we have erred by 0.38 GPA points.

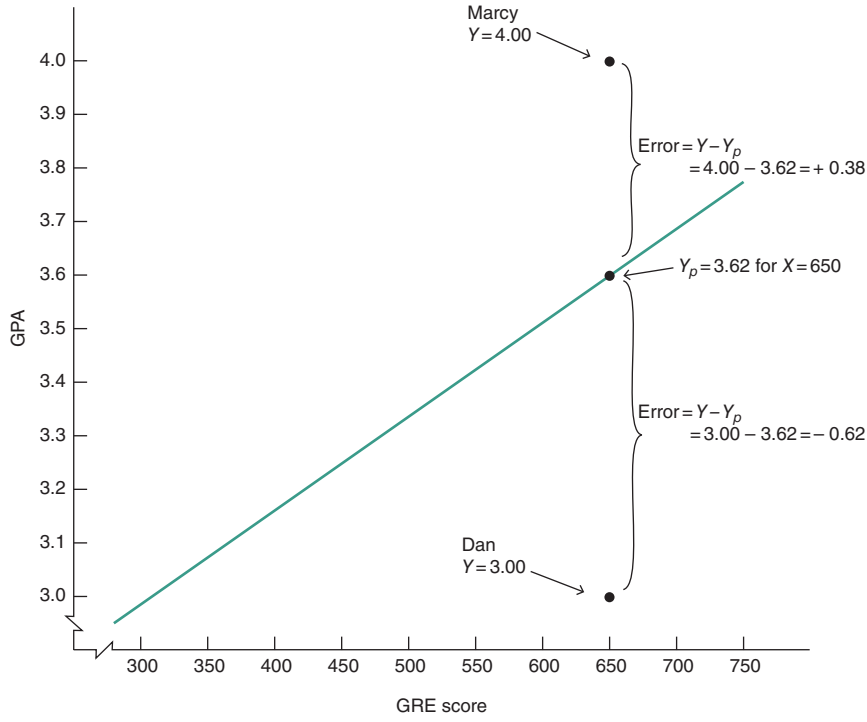


Figure 16.4 An error is the deviation of an actual Y from the predicted Y (Y_p).

Of all the straight lines that could be drawn through the scatter plot, we want the line that creates the least amount of total error possible. Moreover, we want unbiased predictions. This means we want a line that overpredicts to the same extent that it underpredicts. Every data point not falling on the regression line has a corresponding margin of error, defined as $Y - Y_p$ (symbolized as e; $Y - Y_p = e$). Sometimes $Y - Y_p$ will be positive, and sometimes it will be negative, depending on whether Y falls above or below the regression line. An unbiased regression line will perfectly balance the positive and negative values, such that $\Sigma(Y - Y_p) = 0$. However, to use $\Sigma(Y - Y_p)$ as the least squares criterion would be misleading because it suggests an overall summed error of zero. This is similar to the problem addressed in Chapter 4 of using deviation scores to measure variability in univariate distributions. When developing the variance formula, it was noted that $\Sigma(X - M) = 0$. The solution was to square each deviation score in order to remove the negative signs. This same solution is applied to determining the regression line. The least squares criterion requires that the regression line be fitted to the scatter plot in such a way that the sum of the squared errors (Σe^2) is minimized. The way in which the criterion is met mathematically is called the **least squares method.**

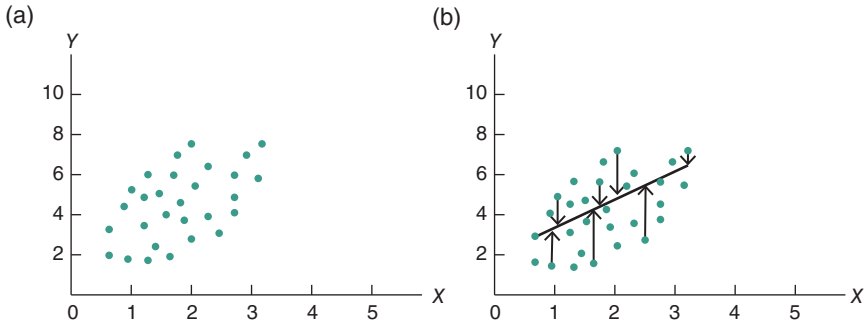


Figure 16.5 (a) The scatter plot has been fitted with a regression line in (b). The line is drawn so that Σe^2 is minimized.

In Figure 16.5b, a regression line has been fitted to the scatter plot depicted in Figure 16.5a. Each data point not on the line possesses some amount of error. No other line drawn through the plot will yield a smaller value for Σe^2.

Establishing the Regression Equation

Linear regression produces a straight line drawn through the middle of a scatter plot. Recall that the general form of the equation for a straight line is $Y = a + bX$. Once the least squares criterion has been specified (i.e. Σe^2 is at a minimum), formulas for a and b can be derived so that the criterion is met (see Draper and Smith, 1966).

Estimate for the intercept

$a = M_Y - bM_X$ (Formula 16.2)

where

a = Y intercept
b = slope of the regression line
M_Y; M_X = means of Y and X values

To arrive at the regression equation, we first substitute the formula for a into the general equation for a straight line:

$$Y = a + bX$$

Regression equation: interim step 1

$Y_p = M_Y - bM_X + bX$ (Formula 16.3)

Rearranging the terms yields Formula 16.4.

Regression equation: interim step 2

$Y_p = M_Y + bX - bM_X$ (Formula 16.4)

Since both X and M_X are multiplied by b, b can be factored out to form Formula 16.5.

Linear regression equation

$Y_p = M_Y + b(X - M_X)$ (Formula 16.5)

Formula 16.5 is the general linear regression equation for predicting Y given X (also called Y on X).[5] This equation, as well as others presented later, is for predicting Y given X. The addendum to the formula section of the end of this chapter presents formulas for predicting X given Y. Only use the formulas in the addendum if Y is determined to be the predictor variable and X is the predicted variable (solving for X_p).

Formula 16.6 is the computational formula for the slope, given the least squares criterion.

Computational formula for the slope

$$b = \frac{n_p(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\left[n_p(\Sigma X)^2\right] - (\Sigma X)^2} \quad \text{(Formula 16.6)}$$

Although it is possible to find Y_p for a given X using a graph with the regression line, this strategy is less precise than plugging the X value into the regression equation and performing the simple arithmetic.

■ **Question** *Use the data below to create a regression equation and find Y_p for $X = 9$.*

X	X^2	Y	Y^2	XY
9	81	11	121	99
6	36	8	64	48
5	25	6	36	30
7	49	9	81	63
4	16	7	49	28

5 There are numerous equivalent formulas for the regression equation. Some textbooks use $Y_p = a + bX$ and provide formulas for b and a. In the author's opinion, however, this is the easiest one with which to work.

Solution

$M_X = 6.20$; $M_Y = 8.20$; $\Sigma X^2 = 207$; $\Sigma X = 31$; $\Sigma Y = 41$; $\Sigma XY = 268$; $n_p = 5$

$$b = \frac{n_p(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\left[n_p(\Sigma X)^2\right] - (\Sigma X)^2}$$

$$b = \frac{5(268) - (31)(41)}{[5(207)] - (31)^2}$$

$$b = \frac{1350 - 1271}{1035 - 961}$$

$$b = \frac{69}{74}$$

$$b = +.93$$

$$Y_p = M_Y + b(X - M_X)$$

$$Y_p = 8.20 + 0.93(X - 6.20)$$

$Y_p = 8.20 + 0.93(X - 6.20)$ is the regression equation for Y given X. Now, any X value can be placed in the equation to yield the Y_p for that given X. The question asks for the predicted Y when $X = 9$.

$$Y_p = 8.20 + 0.93(9 - 6.20)$$

$$Y_p = 8.20 + (0.93)(2.80)$$

$$Y_p = 8.20 + 2.60$$

$$Y_p = \mathbf{10.80}$$ ■

Plotting the Regression Line

To plot the regression line, two points are required. Simply solve the regression equation for any two values of X (ideally, the two points will be a good distance apart). If we make $X = 0$, the Y_p will fall on the Y axis – the Y intercept. There is nothing particularly special about the intercept. However, for purposes of plotting the line, it is oftentimes considered a helpful reference point.

For this example, we will select *X's* of 4 and 8 to draw the regression line.

$$Y_p = 8.20 + 0.93(4 - 6.20)$$

$$Y_p = 8.20 + (0.93)(-2.20)$$

$$Y_p = 8.20 - 2.05$$

$$Y_p = \mathbf{6.15}$$

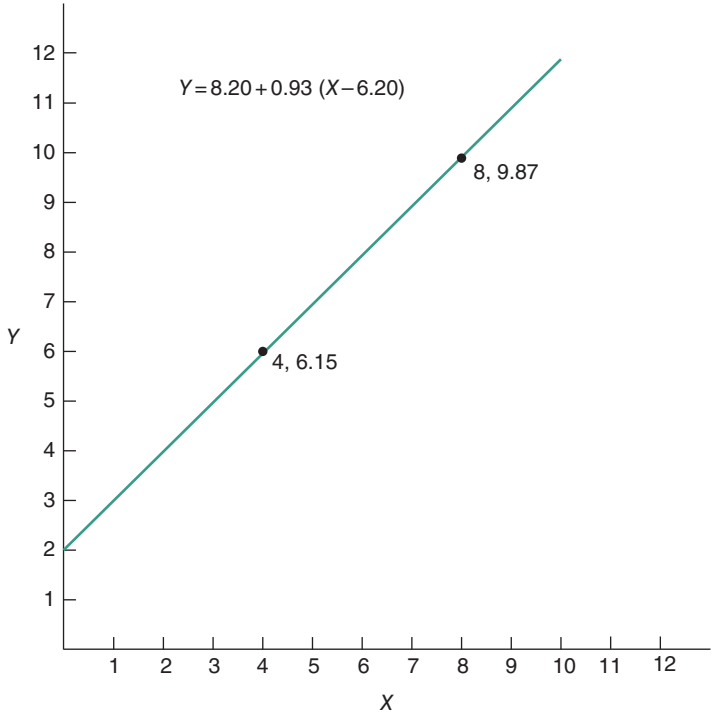


Figure 16.6 Drawing the regression line using two points: (4, 6.15) and (8, 9.87).

A Y of 6.15 is predicted for every person who has an X score of 4. What is the Y_p when $X = 8$?

$$Y_p = 8.20 + 0.93(8 - 6.20)$$
$$Y_p = 8.20 + (0.93)(1.80)$$
$$Y_p = 8.201.67$$
$$Y_p = \mathbf{9.87}$$

The two coordinates are $X = 4$, $Y_p = 6.15$ and $X = 8$, $Y_p = 9.87$. Figure 16.6 shows the regression line for Y given X.

More About the Slope

Formula 16.6 is used to compute the slope from raw data. Formula 16.7 can be used if r (the correlation), s_Y (the standard deviation of Y), and s_X (the standard deviation of X) are provided. Since the correlation is used in Formula 16.7, it is referred to as the correlation formula for the slope.

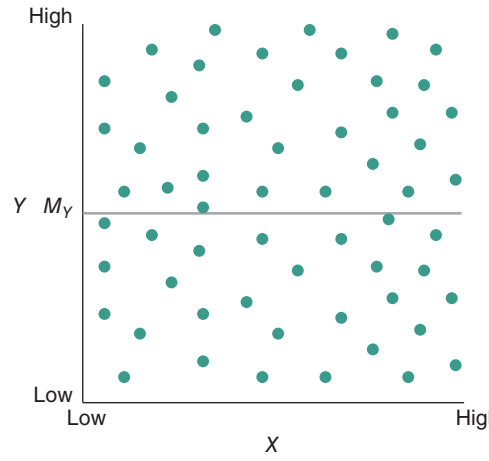


Figure 16.7 When $r = 0$, the slope of the regression line is zero ($b = 0$). The regression line will intersect the Y axis at M_Y.

Correlation formula for the slope

$$b = r\left(\frac{s_Y}{s_X}\right) \quad \text{(Formula 16.7)}$$

■ **Question** *What is the slope and* Y_p when $r = 0$*?*

Solution

$$b = 0\left(\frac{s_Y}{s_X}\right) = 0$$

It does not matter what values s_Y and s_X take, when $r = 0$, $b = 0$. A slope of zero means that the regression line goes neither up nor down as the value of X changes. In Figure 16.7, the scatter plot shows a regression line when $b = 0$. When $r = 0$, b will always = 0, and Y_p is always M_Y:

$$Y_p = M_Y + 0(X - M_X)$$
$$Y_p = M_Y.$$

In Figure 16.7, note that the regression line is parallel to the X axis and intersects the Y axis at M_Y. ■

Analysis of Regression

That X and Y are correlated at the population level is assumed when performing a regression analysis. We do not have a separate section of the text dealing with the assumptions behind the regression analysis, but they are very similar to the

assumptions for other inferential statistical procedures, namely, population representativeness in the samples, independent observations, the use of interval or ratio data, and data that is normally distributed. Additionally, a regression analysis assumes X and Y are correlated at the population level. (There are two additional assumptions discussed later in the chapter). Unlike some of the others, the assumption that X and Y are correlated can be investigated. In fact, an ANOVA can and should be run to test the null hypothesis that $b = 0$. The F ratio is a ratio of two variances; the numerator is the variance of the Y scores that is predicted by the regression equation. This variance measures the systematic changes in Y that occur as the X value changes. This is the portion of variable Y variance that is shared with variable X. The denominator is the unpredicted variance in the Y scores. This variance measures the changes in Y scores that are unrelated to changes in the value of X. As with other ANOVAs, these variances are weighted according to the degrees of freedom associated with them. If the amount of Y variance accounted for by X is similar, per *df*, to the amount of Y variance not explained by X, the F ratio will be unimpressive (close to 1), and the null hypothesis cannot be rejected.

In deference to brevity, we will not walk through the associated ANOVA formulas or summary table. However, virtually all statistical software programs, including Microsoft® Excel and SPSS®, generate an ANOVA to test the null hypothesis that $b = 0$ when a regression analysis is performed. (More information regarding the use of Excel and SPSS to test this null hypothesis can be found at the end of this chapter.) Although the specifics of the analysis of regression process are not covered here, oftentimes the F associated with this analysis is a necessary part of reporting the findings. (See Section 16.7.)

How Accurate Is the Regression Equation?

Regression equations look very scientific. There is a tendency to equate "looking scientific" with "accurate." Regression equations, however, are merely statistical tools, and just as some carpentry tools are more refined than others, some regression equations are more accurate than others. For regression equations, the measure of accuracy is embodied in the concept of *prediction error*. In regression, the measure of prediction error is called the **standard error of the estimate**, which is symbolized as s_e.[6] An understanding of the conceptual basis of the standard error of the estimate first requires a discussion of conditional distributions.

6 There are many different symbols used to represent the standard error of the estimate. We have chosen to use the symbol s_e, so that the subscript reminds us that it is a measure of error.

Conditional Distributions

A bivariate distribution is based on two related univariate distributions. One univariate distribution is comprised of all of the observations of *X*, and the other is comprised of all of the observations of *Y*. They have a correlated relationship; together, the distribution of the pairs of scores comprises one bivariate distribution.

Imagine we have conducted a study and recorded each participant's *X* and *Y* scores on a chip and then deposited all the chips in a box. Then someone asks to see all of the chips that have an *X* score of 5. Do we think every *Y* score written on a chip with a 5 will be the same? Unless the correlation between *X* and *Y* is +1 or −1, there will be an array of *Y* values associated with an *X* value of 5. In other words, for that given value of *X*, there exists a corresponding distribution of *Y* scores. The spread of *Y* scores for a given *X* is called a **conditional distribution**; this means that every *X* value has a *conditional* distribution of *Y* scores. Each conditional distribution has all the characteristics of any distribution of scores; it has a mean, median, standard deviation, range, and so forth.

Figure 16.8 represents a few conditional distributions for different values of *X*. The figure is idealized in that each conditional distribution assumes a normal shape. (This figure is also a bit misleading since a flat surface forces us to illustrate the concept using only two dimensions. More visually appropriate would be to imagine these conditional distributions protruding off the surface into a third dimension.) With very large samples, normally distributed conditional

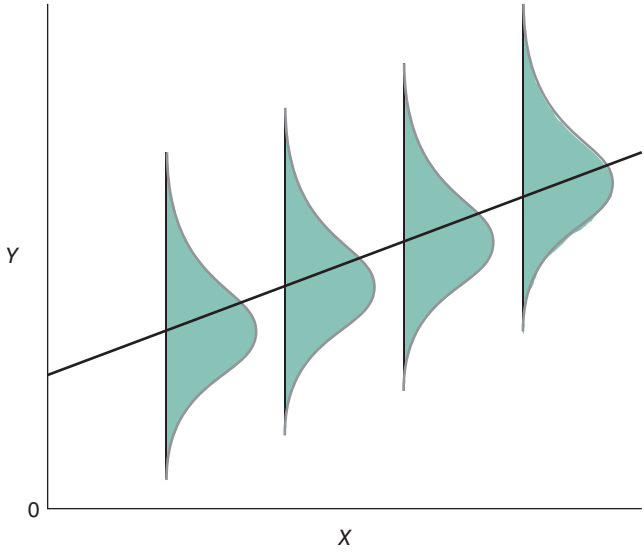


Figure 16.8 Normally distributed conditional distributions of four values of *X*.

distributions are usually the case. Presented later will be a discussion of the implications for prediction when this condition is not met.

In the absence of correlated information, repeatedly predicting the mean of a univariate distribution will minimize prediction errors. (Recall the task of predicting the height of an American biological male without any additional information.) In the presence of correlated information, if conditional distributions are normal, the mean is still the best prediction; but now it is *the mean of each conditional distribution*. Each X score has an associated Y distribution of scores. When asked the question "What would we predict for this X?," the regression equation answers "The mean of the conditional distribution of Y scores associated with that particular X." The regression line connects all the means of the conditional distributions (see Figure 16.8). The regression line connects the means of each conditional distribution; for this reason, it can be referred to as the "line of moving means."

Francis Galton has the distinction of originating the concept of regression. Spotlight 16.1 presents the man in the context of his discoveries.

Spotlight 16.1 Sir Francis Galton

Francis Galton (1822–1911) is credited with developing the concepts of correlation and regression, although his understudy Karl Pearson was responsible for many of the mathematical underpinnings of correlation. Galton led a full and varied life. Born into a wealthy English family, he was afforded the luxury of indulging his scientific curiosities. In the mid-1800s, European explorers were mapping the interior of Africa. Perhaps inspired by the travels of his prodigious cousin, Charles Darwin, Galton departed for Africa at the age of 28. His maps of unknown regions of Africa were published, for which he received a gold medal from the Geographical Society. For the next few years, Galton dabbled in geography and meteorology. He coined the term *anticyclone*, invented three-dimensional weather maps, and even invented spectacles that could be used to read underwater (they did not sell well).

Galton was intellectually gifted and preoccupied with measurement. All throughout life, he zealously counted things. Once, when attending a lecture, he counted the number of fidgets per minute of members of the audience and looked for variation as a function of audience attentiveness versus boredom. His observations were published in *Nature* (Galton, 1885). While his portrait was being painted, he counted the number of brush strokes in an hour. Multiplying by the number of hours it took to paint his portrait, he estimated the total number of strokes at 20 000. (Contemporary clinicians might wonder if Galton suffered from obsessive–compulsive disorder.)

Without question, the greatest influence on Galton's work was Darwin's *Origin of Species*, published in 1859. The mature years of Galton's career were devoted almost entirely to the quantification of heredity. This work led to the development of statistical tools such as correlation and regression. He established the Anthropometric Laboratory and collected thousands of observations on physical and mental attributes, many from parents and their offspring. Indeed, Galton is viewed as the father of *biometrics*, the quantitative aspect of biology, as well as the father of mental testing. Galton's book, *Hereditary Genius* (1869), put forth the belief that intelligence is inherited. His evidence was based primarily on studies that counted the number of eminent people who also had eminent relatives, an admittedly weak methodology by today's standards. In trying to find a quantitative method for showing a link between cross-generational abilities, he came upon the ideas of correlation and regression. Galton had collected data on the heights of parents and their children and had drawn a graph with the average of the parents' heights on one axis and the height of their child on the other axis. However, he was unable to arrive at a suitable method for statistically relating the two distributions. Galton wrote in his autobiography of the moment in which he was struck by the solution to his problem.

> But I could not see my way to express the results of the complete table in a single formula. At length, one morning, while waiting at a roadside station near Ramsgate for a train, and poring over the diagram in my notebook, it struck me that the lines of equal frequency ran in concentric ellipses (Galton, 1908, p. 302).

The insight that "struck" Galton was the *bivariate normal surface*, that is, the infinite series of conditional distributions depicted in the text of this chapter. Galton's discovery of correlation and regression has revolutionized the fields of biology and the social sciences.

Galton's abiding belief in the inheritance of physical, mental, and, in fact, moral attributes led him to begin a movement to better the human race through selective breeding, which he called *eugenics*, meaning "good birth." To be fair to Galton, there is little evidence he could foresee where his ideas might lead. At the time, it held great appeal for many of the brightest scientists, writers, and politicians of the era. Karl Pearson, Sir Ronald Fisher, George Bernard Shaw, George Orwell, Teddy Roosevelt, and Winston Churchill were just some of the prominent and ardent supporters of the eugenics movement. However, the sorrowful story of eugenics in America includes political actions endorsing forced sterilizations and marriage restrictions for the poor, ethnic minorities, and the handicapped, immigration laws restricting entry to peoples from certain parts of the world, and even more well-documented horrific and drastic measures undertaken by German scientists (e.g. Lifton, 2000) and politicians (e.g. Kuhl, 1994) in Nazi Germany. There is simply no positive spin that can throw into better light this darkest chapter of modern scientific reasoning. However, we can say that despite the evils that emerged from the resonance of Galton's utopian ideal, the techniques he invented for exploring the world around us have revolutionized the quantitative aspects of many disciplines in the social and behavioral sciences.

Formulas for the Standard Error of the Estimate[7]

To recap, the regression line connects the means of all of the conditional distributions, and Y_p is the mean of the conditional distribution associated with a given X value. If a conditional distribution has little variability, most of the Y scores bunch around the mean of the distribution. As a result, less error in prediction is made relative to predictions made when a conditional distribution is highly variable. It would make sense that an overall measure of prediction error would be based on the amount of variability found across the set of conditional distributions. Therefore, it should come as no surprise that the definitional formula for the estimated standard error of the estimate looks very much like a standard deviation formula. Formula 16.8 is based on the average of the squared errors for all possible predicted Y scores.

Definitional formula for s_e

$$s_e = \sqrt{\frac{\Sigma(Y - Y_p)^2}{(n-2)}} \quad \text{(Formula 16.8)}$$

The measure of error for a given X is the standard deviation of the Y scores around the Y_p value. If the amount of variability of Y scores differed for each value of X, then a researcher would be in the awkward position of having a different s_e for each X. With two additional assumptions, s_e can be used as a measure of prediction error for any value of X. The first assumption is that each conditional distribution is normally distributed. If Y_p is the mean of each conditional distribution, then it is important that the mean falls at the center of each conditional distribution. (Recall the discussion in Chapter 3 about the problems with using the mean as a measure of centrality for skewed distributions.) The second assumption is that each conditional distribution has the same standard deviation. This assumption is termed **homoscedasticity** (*homo* meaning same, *scedastic* meaning scatter). Formula 16.8 is the definitional formula for the overall amount of prediction error, irrespective of the value of X and the predicted Y score.

The definitional formula for s_e emphasizes the fact that the standard error of the estimate is a standard deviation. The definitional formula is computationally prohibitive since every X score would have to be entered into the regression

7 In this text, regression is discussed as a set of inferential techniques. It is assumed that sample data are used to make inferences about population parameters, in this instance, the standard error of the estimate. Therefore, when we read "the standard error of the estimate," bear in mind that we are estimating a population parameter. The formulas for s_e are presented with the assumption that sample data is being used.

equation and then $\Sigma(Y - Y_p)^2$ would have to be computed for all obtained Y values. When working from raw data, however, the following computational formula can be used.

Computational formula for s_e

$$s_e = \sqrt{\left[\frac{1}{n_p(n_p-2)}\right]\left[(n_p\Sigma Y^2 - (\Sigma Y)^2) - \left(\frac{[n_p\Sigma XY - (\Sigma X)(\Sigma Y)]^2}{n_p\Sigma X^2 - (\Sigma X)^2}\right)\right]} \quad \text{(Formula 16.9)}$$

How Does the Size of the Correlation Affect Prediction Error?

An equivalent formula for the standard error of the estimate is presented, which allows a researcher to examine the relationship between r and s_e. Formula 16.10 is the correlation formula for the standard error of the estimate. It can be used for computational purposes if s_Y, n_p, and r are provided.

Correlation formula for s_e

$$s_e = s_Y\sqrt{(1-r^2)\left[\frac{n_p}{(n_p-2)}\right]} \quad \text{(Formula 16.10)}$$

Suppose the correlation between X and Y is perfect. Substituting r in the formula with either +1 or −1 yields $s_e = 0$. A perfect correlation means no prediction error.

Interpreting the Standard Error of the Estimate

With a normal distribution, approximately 68% of the scores fall within ±1 standard deviation of the mean. In regression, Y_p is the mean of a conditional distribution; approximately 68% of the actual Y scores fall within $\pm 1 s_e$ of the mean (Y_p) of the conditional distribution. For a given X score, 95% of the actual Y scores fall between $Y_p \pm 1.96 s_e$ (Glass & Stanley, 1970; Shavelson, 1988; Wiggins, 1973).[8] Figure 16.9 shows conditional distributions for three values of X. The dotted lines mark the cutoffs between which 95% of the actual Y scores lie, for each value of X.

Suppose we are interested in predicting the GPAs of first-year graduate students from quantitative GRE (GRE-q) scores. Assume that $s_e = 0.25$ and that the

8 Recall the 68-95-99.7 rule introduced in Chapter 4, roughly reflecting the percent of scores within ±1, ±2, and ±3 standard deviations of the mean. Since s_e is a standard deviation, the same relationship applies here. Of course ±1.96 is a more accurate number than 2 to encapsulate the middle 95% of the scores.

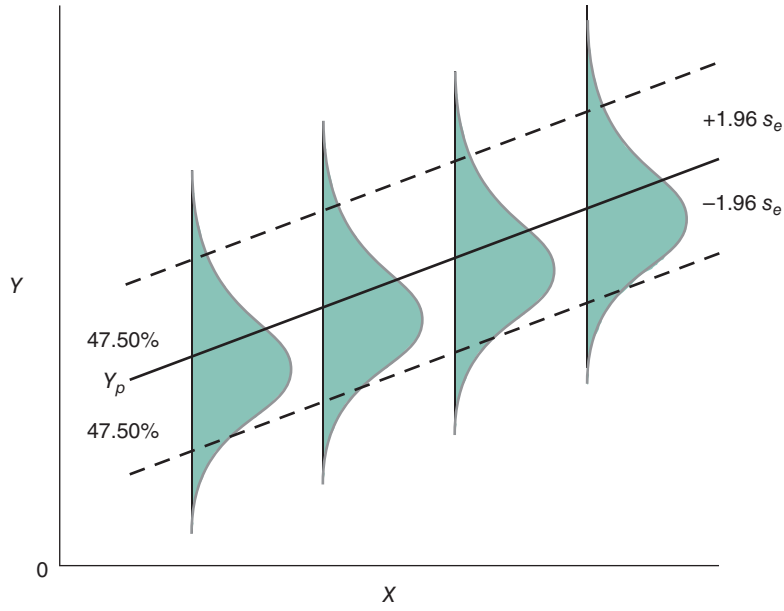


Figure 16.9 Ninety-five percent of the actual *Y* scores fall between $Y_p \pm 1.96\ s_e$.

regression equation predicts a GPA of 3.50 for everyone with a GRE-q of 160. Although we predict a GPA of 3.50 for every student with a GRE-q of 160, there will be error since the correlation between GRE-q and GPA is not perfect. We can estimate that 68% of the students with a GRE-q of 160 will achieve a GPA between 3.75 and 3.25 ($Y_p \pm 1s_e = 3.50 \pm 0.25$). We can also estimate that 95% of the students with a GRE-q of 160 will achieve a GPA between 3.99 and 3.01 [$Y_p \pm 1.96s_e = 3.50 \pm 1.96(0.25)$].

16.4 Putting It All Together: A Worked Problem

A team of researchers is interested in the relationship between marital satisfaction among couples and the subsequent marital satisfaction reported by their children. A large sample of couples had been studied 15 years earlier; their scores from a marital satisfaction questionnaire (MSQ) are still available. They obtain the data from all couples who had a 15-year-old child at the time of the original study. A random sample of seven couples is drawn, their MSQ scores are recorded, and their children, now 30 years of age, are located. All seven children are now married and the MSQ is administered to each of them. The researchers are interested in the correlation between parents' and children's level of marital satisfaction. They would also like a regression equation that

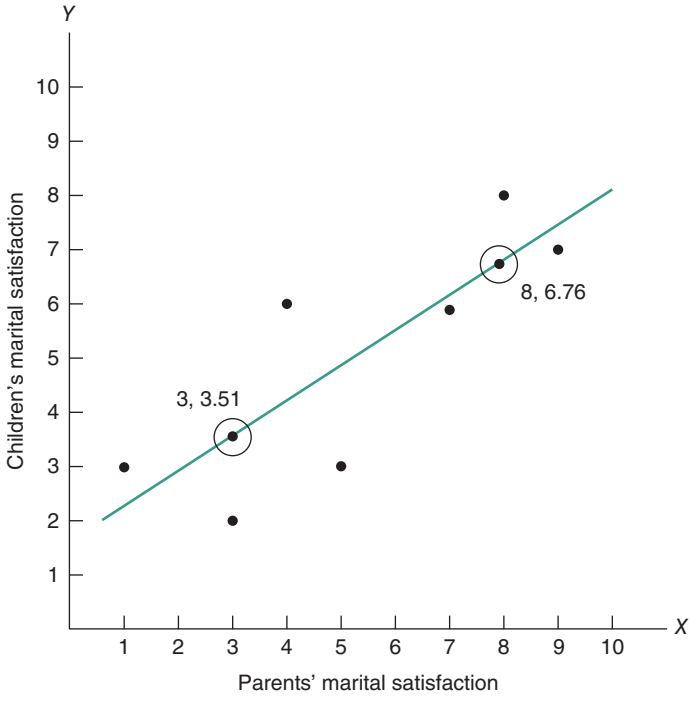


Figure 16.10 The scatter plot and regression line for the data on marital satisfaction.

can be used to predict the future marital satisfaction of children, knowing only the level of satisfaction reported by their parents. The following table and list of steps show the calculation of the correlation, regression equation, and the standard error of the estimate. Figure 16.10 shows the scatter plot of the pairs of scores. Marital Satisfaction scores can range from 0 (very dissatisfied) to 10 (very satisfied). An interpretation of the findings is also provided.

Step 1. Construct a scatter diagram to see if there is a linear relationship between X and Y. Figure 16.10 indicates that the relationship is linear; there is no curve to the swarm of data points. (The regression line is drawn later.) Moreover, the scatter plot reveals a positive correlation between X and Y because the swarm of points ascends from the lower left to the upper right of the graph.

Step 2. Compute the correlation between X and Y. Mathematically speaking, it is not necessary to compute the correlation in order to establish the regression equation. However, the correlation is always of substantive interest; in addition, regression is useless if there is no correlation between the variables. The correlation coefficient is computed using the Pearson raw score formula and is found to be large, +.80.

Marital Satisfaction	
Parents: X	***Children: Y***
1	3
3	2
7	6
9	7
8	8
4	6
5	3
$M_X = 5.29$	$M_Y = 5.00$
$n_p = 7$	
$\Sigma X = 37$	$\Sigma Y = 35$
$\Sigma X^2 = 245$	$\Sigma Y^2 = 207$
$s_X = 2.87$	$s_Y = 2.31$
$\Sigma XY = 217$	

$$r = \frac{n_p(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{\left[n_p(\Sigma X^2) - (\Sigma X)^2\right]\left[n_p(\Sigma Y^2) - (\Sigma Y)^2\right]}}$$

$$r = \frac{7(217) - (37)(35)}{\sqrt{\left[7(245) - (37)^2\right]\left[7(207) - (35)^2\right]}}$$

$$r = \frac{1519 - 1295}{\sqrt{(1715 - 1369)(1449 - 1225)}}$$

$$r = \frac{224}{\sqrt{(346)(224)}}$$

$$r = \frac{224}{278.40}$$

$$r = +.\mathbf{80}$$

Step 3. Compute the slope. Formula 16.6 is used when working from raw data. Since r has been computed and s_X and s_Y are provided, Formula 16.7 would be the formula of choice. Nonetheless, the computational formula is used for illustrative purposes.

$$b = \frac{n_p(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\left[n_p(\Sigma X)^2\right] - (\Sigma X)^2}$$

$$b = \frac{7(217) - (37)(35)}{[7(245)] - (37)^2}$$

$$b = \frac{1519 - 1295}{1715 - 1369}$$

$$b = \frac{224}{346}$$

$$b = +\mathbf{0.65}$$

Step 4. The necessary values for establishing the regression equation are now available. Use the regression equation to plot the regression line. Pick any two values of X (preferably at some distance from each other), compute Y_p for each value, plot each Y_p, and draw a straight line. X values of 3 and 8 have been chosen. Figure 16.10 shows the regression line.

$$Y_p = M_Y + b(X - M_X)$$

$$Y_p = 5.00 + 0.65(X - 5.29)$$

$$Y_{p\ for\ X=3} = 5.00 + 0.65(3 - 5.29) = \mathbf{3.51}$$

$$Y_{p\ for\ X=8} = 5.00 + 0.65(8 - 5.29) = \mathbf{6.76}$$

Plot points at coordinates 3, 3.51 and 8, 6.76.

Step 5. Compute the standard error of the estimate. Formula 16.10 would be the easiest to use because r and s_Y are known. However, Formula 16.9, the computational formula, will be used to illustrate its computational steps.

$$s_e = \sqrt{\left[\frac{1}{n_p(n_p - 2)}\right]\left[\left(n_p\Sigma Y^2 - (\Sigma Y)^2\right) - \left(\frac{\left[n_p\Sigma XY - (\Sigma X)(\Sigma Y)\right]^2}{n_p\Sigma X^2 - (\Sigma X)^2}\right)\right]}$$

$$s_e = \sqrt{\left[\frac{1}{7(7-2)}\right]\left[\left(7(207) - (35)^2\right) - \left(\frac{[7(217) - (37)(35)]^2}{7(245) - (37)^2}\right)\right]}$$

$$s_e = \sqrt{[0.029]\left[(1449 - 1225) - \left(\frac{[1519 - 1295]^2}{1715 - 1369}\right)\right]}$$

$$s_e = \sqrt{[0.029]\left[(224) - \left(\frac{(224)^2}{346}\right)\right]}$$

$$s_e = \sqrt{(0.029)(224 - 145.02)}$$
$$s_e = \sqrt{(0.029)(78.98)}$$
$$s_e = \sqrt{2.29}$$
$$s_e = \mathbf{1.51}$$

The measure of prediction error is s_e = 1.51. Suppose, for a specific 15-year-old adolescent biological male, we want to predict his level of marital satisfaction when he is 30 years old. We administer the MSQ to his parents, find that their score is 7, and, using the regression equation, find Y_p = 5.00 + 0.65 (7 – 5.29) = 6.11. Therefore, we predict that he will report a level of marital satisfaction of 6.11 when he is 30 years old. However, what about the accuracy of our prediction? We can state that approximately 68% of all 15-year-old children who have parents who scored a 7 on the MSQ will report, 15 years later, marital satisfaction scores between 4.60 and 7.62 (6.11 ± 1.51).

By now, we might have wondered why a prediction equation goes by the term *regression*. Galton coined the term for good reason. Box 16.2 details the origin of the term and explains why the term reflects a fundamental aspect of making predictions.

Box 16.2 Why Is a Prediction Equation Called a Regression Equation?

While in grade school, one of the authors of this text, Laurence, had a teacher who wanted to reward effort. She gave a test at the beginning of the term (pretest) and another test at the end of the term (posttest). She subtracted each student's pretest score from their posttest score to arrive at a difference score ($Y - X = D$). A positive D score indicated improvement, a 0 showed no improvement, and a negative D score meant that the student had done worse on the second test. She then used the D score as a measure of effort and assigned her grades accordingly; those students with positive D scores received the highest grades, and those students with negative D scores received the lower grades. Her intent was commendable. She wanted to make her grading fair by not disadvantaging the students who performed poorly on the pretest. She wanted to impress on us the importance of always "trying really hard," irrespective of our previous scores. As she reminded us, "No matter how good you are; there is always room for improvement." Although Laurence did quite well on the pretest, he recalled that he did not do as well on the posttest and received a rather mediocre grade. Years later, after learning something about regression analysis, he exclaimed, "I was robbed!" (Not that he harbored any deep-seated resentment.) The teacher had failed to understand a fundamental concept of regression, called *regression toward the mean*. Without intending, she had doomed most of the students who did well on the pretest. Similarly, she had practically

assured that the students who scored low on the pretest would eventually receive high grades.[9]

Regression toward the mean is a built-in characteristic of using one variable to predict a second variable, *when the correlation between the variables is less than perfect*. Galton was the first to discover this phenomenon and referred to it as "regression to mediocrity." In his work on heredity, he gathered measures of fathers' and sons' heights (Galton, 1869). He found that their heights were positively correlated (although not perfectly). Tall fathers tended to have tall sons and short fathers tended to have short sons. However, very tall fathers tended to have sons who were not as tall as the father, and very short fathers tended to have sons who were not as short as their father. In general, the heights of sons tended to "regress" toward the mean height of all sons. This is the meaning of "regression toward the mean." A statistical way of stating the regression effect is as follows. Suppose we examine just those fathers who have heights that are two standard deviations above the mean height of all fathers. We will find that their sons will not be, on the average, two standard deviations above the mean of the height of all sons. The sons of these tall fathers will tend to be above the mean of all sons, but not two standard deviations above the mean. The same logic holds for short fathers and their sons. The regression effect is also found in situations that have nothing to do with heredity.

Whenever two variables are imperfectly correlated (i.e. less than ± 1), regression toward the mean is likely to occur. This fact has a very important implication for designing studies. Suppose we want to show the advantages of a study program for poor students. We administer some test of ability, take the lowest 10% of the group, and put them into the study. After six months in the program, we administer the same ability test and conduct a dependent-samples *t* test, comparing the pretest and posttest means. We will most likely find a significant improvement in ability. Should we conclude the study program was effective? Not according to the foregoing discussion of the regression effect. Those students who scored low on the pretest will tend to score to some degree higher on the posttest independent of any effect from the educational program. (Including a control group of low-ability students who did not receive the program would not eliminate the regression effect, but it would allow the investigator to examine how much improvement was due to regression toward the mean.)

So, whenever a teacher that hands out grades based on improvement is found, kindly explain to them the concept of regression toward the mean.

9 The other author of this text, Paul, feels he may have had a different feeling about this grading system, given his typical performance on grade-school pretests!

16.5 The Coefficient of Determination in the Context of Prediction

The concept of r^2 was discussed in Chapter 15. The coefficient of determination reflects the amount of shared variance between X and Y, that is, the amount of variance in the Y scores accounted for by the variance in X scores. Now that we have learned some of the intricacies of regression, r^2 can be presented from the perspective of prediction. Indeed, the coefficient of determination is easier to understand when viewed with respect to prediction.

From the discussion of the standard error of the estimate, we know that as the correlation increases, the amount of prediction error decreases. The correlation formula for the standard error of the estimate (Formula 16.10) was stated as

$$s_e = s_Y \sqrt{(1-r^2)\left[\frac{n_p}{(n_p-2)}\right]}$$

A simpler version of Formula 16.10 leaves off the component incorporating sample sizes (this feature of the formula becomes increasingly irrelevant as n_p increases). This formula merely *estimates* the standard error of the estimate.

Correlation formula for estimating s_e

$s_e = s_Y \sqrt{(1-r^2)}$ (Formula 16.11)

The percentage of prediction error can range from 100 to 0%. Using Formula 16.11, when $r = 0$, there is no improvement (reduction in error) over the value of s_Y $\left(s_e = s_Y\sqrt{1-0^2} = s_Y\right)$. In this case, prediction error is at a maximum – 100%. What happens when r = .50 and the coefficient of determination is r^2 = .25?

$$s_e = s_Y \sqrt{1-(.50)^2}$$

$$s_e = s_Y \sqrt{1-.25}$$

$$s_e = s_Y \sqrt{.75}$$

$$s_e = s_Y (.87)$$

Prediction error is now 87% of the value of s_Y. If prediction error is 100% when $s_e = s_Y$, then when r = .50, prediction error decreases by 13% (100 – 87%). We might ask, "decreases prediction error relative to what?" Well, relative merely to

predicting the mean of the Y distribution. Some correlations and their corresponding reductions in prediction error follow.

r	Reduction in Prediction Error (%)
1.00	100
.75	34
.50	13
.25	3
.00	0

Whenever correlated information is available, we should use it to improve prediction. However, as is shown in the preceding table, substantial reductions in prediction error are only achieved with strong correlations. Also, note that the relative reduction in prediction error "gains speed" as the size of the correlation increases. While there is only a 10% improvement in error between correlations of .25 and .50, there is a 21% improvement in error rate between the correlations of .50 and .75.

16.6 The Pitfalls of Linear Regression

Chapter 15 discussed several factors that lead to the misinterpretation of a correlation. Since regression is intimately connected with correlation, it should be no surprise to learn that the same conditions that create misleading correlations will adversely affect the usefulness of regression analyses. This section discusses some of the factors that can undermine the accuracy of the prediction equation. (For a more detailed presentation of these factors, refer back to Section 15.5.)

Restricted Range

Figure 16.11 reveals a linear relationship between X and Y. What if the sampled values of X are confined to the area between the vertical lines? The scatter plot between the lines fails to capture the true relationship between X and Y. As was previously discussed in Chapter 15, the problem of range restriction can arise when using GRE scores to predict GPA in graduate school. The available GREs are the GREs of the students who have been accepted to a graduate program. This sample does not accurately reflect the full range of GRE scores; scores above the population mean are overrepresented. Not only will the correlation

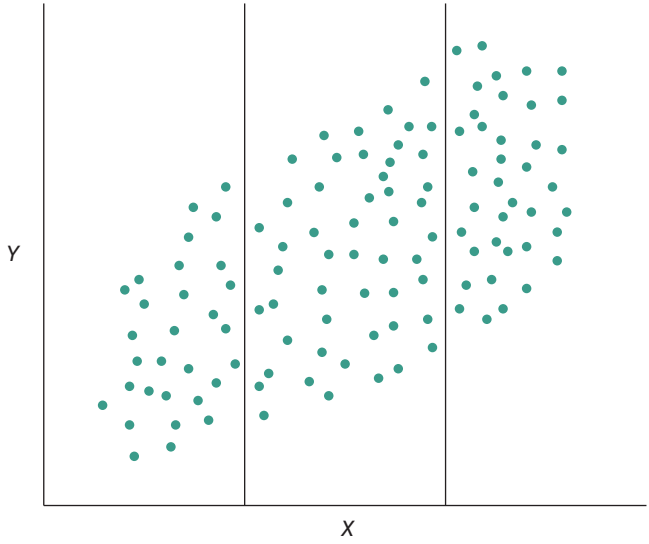


Figure 16.11 A linear relationship may be impossible to detect if either *X* or *Y* has a restricted range of scores.

between GRE and GPA underestimate the strength of the relationship between the variables, but also the accuracy of the regression equation will diminish.

Examining the scatter plot will not help to decide if there is a range restriction for *X* and/or *Y*. We will need to know the range of *possible* values for each variable and then separately examine the *X* and *Y* sample distributions to see if these scores are an accurate representation of the population from which they are drawn.

Extreme Scores

An extreme score can create the false impression that there is a linear relationship between *X* and *Y*. Figure 16.12 shows a scatter plot in which there is one outlier in the upper right corner of the plot. Were it not for this extreme point, the regression line would be parallel to the *X axis*, indicating no relationship between *X* and *Y*. However, the extreme point could lift the regression line so that its slope is nonzero. With the outlier included, the regression equation might be relied upon for predictions when, in fact, there is no relationship between *X* and *Y*. Thankfully, this pitfall is mitigated as the sample size increases. This is another reason to create a scatter plot; it will reveal the presence of an extreme score.

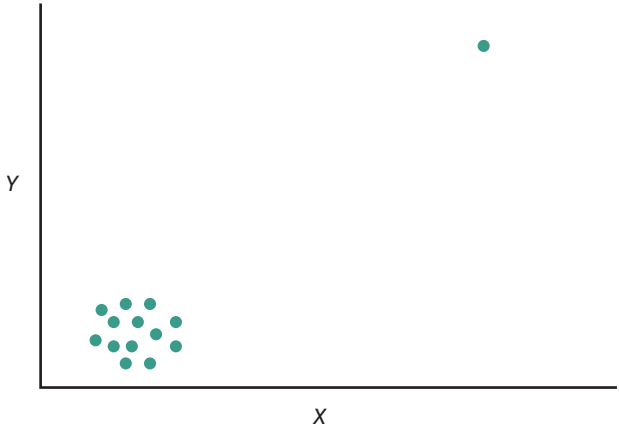


Figure 16.12 Even though the majority of scores suggest no relationship exists between *X* and *Y*, the extreme score in the upper right can create the impression of a linear relationship between *X* and *Y*.

Overgeneralization

It is a mistake to apply a regression equation to different populations from the one used to establish the regression equation. Recall the hypothetical study in which parents' marital satisfaction was used to predict the level of marital satisfaction among their children, 15 years later. All the children in the sample were 15 years old when values of the predictor variable (parents' MSQ score) were obtained. It is most likely still appropriate to apply the regression equation to predict the future marital satisfaction of any 15-year-old adolescent. However, it is dubious to think we could use the equation to predict the future marital satisfaction of 2-year-old children or 20-year-old children. The regression equation was not constructed using data from parents of children with these ages. To apply the regression equation to a different population is to overgeneralize the results obtained from one sample.

Violating Homoscedasticity

What happens to prediction error when the standard deviations of the conditional distributions are dissimilar (called **heteroscedasticity**)? Figure 16.13 shows a heteroscedastic plot for *Y* given *X*. There is a positive linear relationship between the variables, but the amount of prediction error increases as the value of *X* increases. The standard error of the estimate would be different for every value of *X*. The magnitude of the prediction error for each predicted *Y* would fluctuate accordingly. Prediction in this case would be worse for higher values of *X* as compared with lower values of *X*. As a result, we would not be justified

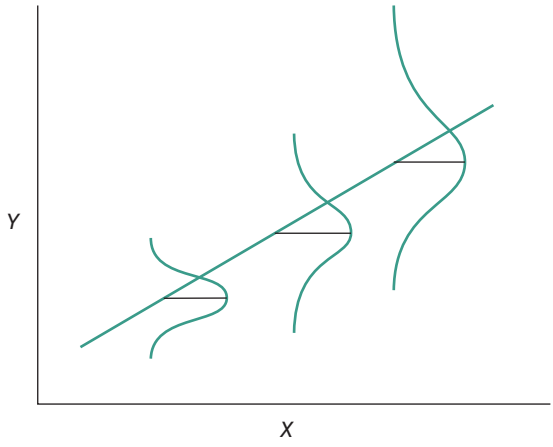


Figure 16.13 Conditional distributions with progressively larger standard deviations as the value of *X* increases. This is a violation of homoscedasticity.

in assuming that 68% of the actual *Y* scores for a given *X* score lie between $Y_p \pm 1s_e$.

Be careful not to get lost in the mechanics of calculating regression equations. Regression is a powerful statistical tool that can give the researcher the ability to predict unknown values, but there are also many pitfalls to be carefully avoided.

16.7 How to Present Formally the Conclusions of a Linear Regression Analysis

The proper reporting of regression equations is a bit different from the reporting of *t*, *F*, and *r* tests. When reporting a linear regression, we must identify the predictor and predicted variables, register the rejected *F* testing the null that $b = 0$, and include a quantitative measure of the relationship between them (slope) and, typically, the r^2 value as well. For instance, "A simple linear regression was calculated to predict height based on shoe size. Statistical evidence for a regression equation was found, $F(1, 6) = 41.63$, $p < .05$, with an r^2 of .87. The resulting equation suggests height increases 0.46 in. for each unit increase in shoe size." Notice that a "0" preceding the decimal is typically not reported in professional writings. A failure to reject the null might read, "A simple linear regression was calculated to predict height based on shoe size. No statistical evidence was found for a regression equation, $F(1, 6) = 1.63$, *n.s.*"

Summary

Linear regression is a statistical method utilizing bivariate data to predict the value of one variable when information about another variable, correlated with the first one, is available. Regression proceeds in two steps. First, a prediction equation is established from a random sample of participants. Second, the prediction equation is applied to individual cases where the value of only one variable is known.

The regression equation defines a line along which each predicted Y for any given X lies. The slope is the angle of the regression line. It reflects the "rise over the run." The slope states the number of units Y changes as X changes by one unit. A positive slope indicates a positive correlation, whereas a negative slope indicates a negative correlation. A slope of zero is a straight line parallel with the X axis, intersecting the Y axis at the mean of the Y distribution.

A correlation of zero will yield a slope of zero, neither ascending nor descending with increasing values of X. Since a slope of zero will intersect the Y axis at M_Y, a correlation of zero means that M_Y should always be predicted in the absence of correlated information. An analysis of regression can be run testing the null hypothesis that $b = 0$ for the population of bivariate scores. If this null hypothesis cannot be rejected, the findings of a regression analysis are meaningless.

An error in predication is the difference between the actual score obtained by the participant and the score predicted for the person. The prediction line must be drawn such that the sum all of the errors will equal zero, indicating an unbiased prediction equation. Additionally, the regression line is to be drawn so that the sum of the squared errors, Σe^2, is at a minimum. This is called the least squares method.

Each X score is associated with an array or distribution of Y scores. These sets of Y scores for given values of X are called conditional distributions. The regression analysis assumes them to be normal and to have similar standard deviations across the range of possible X values (the assumption of homoscedasticity). For any given X score, it is the mean of the corresponding conditional distribution of Y's that is the predicted value (Y_p).

The standard error of the estimate is a measure of prediction accuracy. Sixty-eight percent of the actual Y scores will fall within $\pm 1 s_e$ of any Y_p. When the correlation is ±1, prediction error is zero. This reflects the fact that when $r = \pm 1$, *all* points lie on the regression line; there is no array of Y scores associated with any X score.

The viability of a regression analysis is based on several factors. A regression equation should only be used with individuals who are represented in the sample from which the regression equation was established. Range restrictions of X and/or Y as well as the presence of extreme scores or heteroscedasticity will also invalidate regression analyses.

Using Microsoft® Excel and SPSS® to Create a Linear Regression Line

Excel

General instructions for data entry into Excel can be found in Appendix C.

Data Entry

Enter the bivariate data into two adjacent columns, being sure to keep the data from each participant together in the same row. Label the columns appropriately. (See Figure 16.14 for an example.)

Data Analysis

1) Excel has built-in programs for many inferential tests, including linear regression. To access it, click on the Data tab on the top menu and then click **Data Analysis**. (Some versions of Excel have a "Tools" tab. The Data Analysis function may be under this tab.) If this option is not found, the Data Analysis ToolPak needs to be installed. See Excel instruction materials for how to install this feature.
2) With the Data Analysis box open, select **Regression**.
3) Input the data range for each variable by dragging over the data set for each variable into the corresponding **Input Range** box. Make sure that the predicted variable (*Y*) is placed in the **Input Y Range** box and the predictor variable (*X*) is placed in the **Input X Range** box. Accidentally switching these

Shoe size	Height
72	10
66	9
74	13
68	10
63	7
70	10
73	12
67	9

Summary output

Regression statistics	
Multiple *R*	0.9348898
R Square	0.874019
Adjusted *R* square	0.8530222
Standard Error	0.709876
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	20.97645601	20.97646	41.62623	0.000656801
Residual	6	3.023543 99	0.503924		
Total	7	24			

	Coefficients	*Standard error*	*t Stat*	*P-value*
Intercept	–21.521 685	4.892 131 999	–4.39924	.004572
X variable 1	0.4 560 099	0.070 679 057	6.451 839	.000657

Figure 16.14 A worked example of a regression analysis using Microsoft Excel.

around will give us a faulty slope value. In the Figure 16.14 example shown, we are predicting height (Y) using shoe size (X). (If we include the labels in the data range, make sure to click the **Labels** box to exclude those cells.) Leave the **Constant to Zero** box unchecked (we do not want to force the regression line through origin of the graph), and leave the **Confidence Level** box at the 95% default value.

4) Decide on an **Output** option. The default is to place it on a separate worksheet. Leave the other boxes below also unchecked. These are for analyses that are more sophisticated.
5) Click **OK**. (Increase column width as necessary so the longer output labels can be read.)
6) Three tables are produced. Below are the outputs of particular importance to a regression analysis. **Multiple R** is the correlation between the two variables (Pearson's r). **R Square** is the coefficient of determination (r^2). **Standard Error** is the standard error of the estimate. [The **ANOVA** analysis in the middle box tests the null hypothesis that $b = 0$. If the regression ***F*** is not large and the probability associated with it (***Significance F***) is larger than .05, then we cannot reject this null, and the regression analysis is meaningless. This is akin to testing the null hypothesis that $\rho = 0$ in Chapter 15 and failing to reject the null.] The coefficient associated with the X variable (**X Variable 1**) in the last box is the slope. Excel's regression analysis does not produce either M_X or M_Y. We will need to use the **Descriptive Statistics** function found in the **Data Analysis** toolbox to generate those values. (See Figure 16.14 for a worked example.)
7) The slope, M_X, and M_Y can be used to construct a regression line equation and solve for the corresponding Y_p for any given value of X.

SPSS

General instructions for inputting data into SPSS can be found in Appendix C.

Data Entry

In SPSS, each row of the data file represents a participant. Since bivariate data is used in calculating a regression equation, create a series of variables within **Variable View** corresponding to the variables measured. Then, go to **Data View,** and input the data, being careful to keep data from each participant within a given row. See Figure 16.15 for an example.

Data Analysis

1) Click **Analyze** on the tool bar, select **Regression**, and then click **Linear**.
2) Use the arrow key to move the variable to be predicted into the **Dependent** box and the predictor variable into the **Independent(s)** box. Be careful not to get these two confused. Doing so would generate a wrong slope value.

	Shoe size	Height
1	72	10
2	66	9
3	74	13
4	68	10
5	63	7
6	70	10
7	73	12
8	67	9

Figure 16.15 An example of entered data for a regression analysis in SPSS.

Regression

Model summary

Model	*R*	*R* square	Adjusted R square	Std. error of the estimate
1	.935[a]	.874	.853	.710

[a] Predictors: (constant), shoe size

ANOVA[a]

Model		Sum of squares	*df*	Mean square	*F*	Sig.
1	Regression	20.976	1	20.976	41.626	.001[b]
	Residual	3.024	6	.504		
	Total	24.000	7			

[a] Dependent variable: height

[b] Predictors: (constant), shoe size

Coefficients[a]

Model		Unstandardized coefficients		Standardized coefficients	*t*	Sig.
		B	Std. error	Beta		
1	(Constant)	–21.522	4.892		–4.399	.005
	Shoe size	.456	.071	.935	6.452	.001

[a] Dependent variable: height

Figure 16.16 Output tables from a worked example using SPSS to run a regression analysis.

3) Since we will need M_X and M_Y to create the regression equation, click the **Statistics** box in the upper right-hand corner, and then select **Descriptives**. Click **Continue**.
4) Click **Ok** on the main menu and run the analysis.
5) The first box will give us both M_X and M_Y.
6) The slope can be found in the **Coefficients** box at the bottom. The slope is the **unstandardized coefficients B** value associated with the predictor variable.
7) Other helpful information presented includes an ANOVA test of the null hypothesis $b = 0$. In the **ANOVA** box, if the probability (**Sig.**) associated with the **F** value is larger than .05, then the null cannot be rejected, and the analysis is meaningless. (This is akin to testing the null hypothesis that $\rho = 0$ in Chapter 15 and failing to reject the null.) In the **Model Summary** box, **R** is the Pearson *r*, **R Square** is the coefficient of determination (r^2), and **Std. Error of the Estimate** is the standard error of the estimate. (See Figure 16.16 for a worked example. Note: Some unreferenced tables have been removed from the figure.)

Key Formulas

Straight line formula

$Y = a + bX$ (Formula 16.1)

Estimate for the intercept

$a = M_Y - bM_X$ (Formula 16.2)

Regression equation: interim step 1

$Y_p = M_Y - bM_X + bX$ (Formula 16.3)

Regression equation: interim step 2

$Y_p = M_Y + bX - bM_X$ (Formula 16.4)

Linear regression equation

$Y_p = M_Y + b(X - M_X)$ (Formula 16.5)

Computational formula for the slope

$$b = \frac{n_p(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\left[n_p(\Sigma X)^2\right] - (\Sigma X)^2} \quad \text{(Formula 16.6)}$$

Correlation formula for the slope

$$b = r\left(\frac{s_Y}{s_X}\right) \quad \text{(Formula 16.7)}$$

Definitional formula for s_e

$$s_e = \sqrt{\frac{\Sigma\left(Y - Y_p\right)^2}{(n-2)}} \quad \text{(Formula 16.8)}$$

Computational formula for s_e

$$s_e = \sqrt{\left[\frac{1}{n_p(n_p-2)}\right]\left[\left(n_p\Sigma Y^2 - (\Sigma Y)^2\right) - \left(\frac{\left[n_p\Sigma XY - (\Sigma X)(\Sigma Y)\right]^2}{n_p\Sigma X^2 - (\Sigma X)^2}\right)\right]} \quad \text{(Formula 16.9)}$$

Correlation formula for s_e

$$s_e = s_Y\sqrt{(1-r^2)\left[\frac{n_p}{(n_p-2)}\right]} \quad \text{(Formula 16.10)}$$

Correlation formula for estimating s_e

$$s_e = s_Y\sqrt{(1-r^2)} \quad \text{(Formula 16.11)}$$

Addendum

The following formulas predict X given Y.

Regression equation

$$X_p = M_X + b(Y - M_Y)$$

Formulas for the slope

$$b = \frac{n_p(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\left[n_p(\Sigma Y)^2\right] - (\Sigma Y)^2}$$

$$b = r\left(\frac{s_X}{s_Y}\right)$$

Formulas for the estimated standard error of the estimate

Correlation formula for estimating s_e

$$s_e = s_X\sqrt{(1-r^2)}$$

Computational formula for s_e

$$s_e = \sqrt{\left[\frac{1}{n_p(n_p-2)}\right]\left[\left(n_p\Sigma X^2 - (\Sigma X)^2\right) - \left(\frac{\left[n_p\Sigma XY - (\Sigma X)(\Sigma Y)\right]^2}{n_p\Sigma Y^2 - (\Sigma Y)^2}\right)\right]}$$

Key Terms

Regression
Regression equation
Multiple regression
Regression line
***Y* Intercept**
Slope
Least squares criterion
Least squares method
Standard error of the estimate
Conditional distribution
Homoscedasticity
Heteroscedasticity

Questions and Exercises

1 If blindly guessing a value from a known distribution of normally distributed scores, what value should be chosen? Why?

2 What are the two important uses of regression analysis?

3 Why must we use care when interpreting the findings of regression analyses in the same way we use care when interpreting correlations?

4 What does it mean to say the regression line is based on the least squares method?

5 What is the name of the place on the Y axis where the regression line passes through?

6 Why cannot $\Sigma(Y - Y_p)$ be used as the basis for measuring error?

7 What concept does s_e represent?

8 What is the difference between a small standard error of the estimate and a large one?

9 If $r = 0$, what must b equal?

10 What does the slope tell us? What does it not tell us?

11 Even though a data set may have many different Y values associated with a given X value, why is the generated Y_p always the same?

12 When is it not reasonable to perform a regression analysis on bivariate data? Why?

13 Why is the assumption of normality for conditional distributions important?

14 When using the standard error of the estimate as a measure of prediction error, why is the assumption of homoscedasticity important?

15 Why does a multiple regression analysis typically explain a larger percentage of variability in a predicted variable than a linear regression analysis?

16 Why must a researcher be careful when using a prediction equation with an individual who has characteristics unlike the people who were included in the original sample from which the regression equation was derived?

17 Use the correlation formula (Formula 16.7) to determine the slope for predicting Y from X of the following bivariate distributions.
a $r = -.51$, $s_Y = 3.75$; $s_X = 8.54$.
b $r = .33$, $s_Y = 1.71$; $s_X = 2.17$.
c $r = .82$, $s_Y = 1.03$; $s_X = 1.54$.

18 Suppose a tax official learns that the slope for the regression line between the price of gas and the number of nonwork-related miles driven by drivers to be –0.40. Interpret the likely relationship between gas prices and transportation activities like traveling for vacation.

19 Suppose a regression analysis of bivariate data gathered on the variables "hours studied per week" and "GPA" results in a regression equation as follows: $Y_p = 2.92 + 0.08(X - 18)$. What would be our predicted GPA if we adopted each of the following study plans?
a 10 hours a week.
b 22 hours a week.
c 30 hours a week.

20 Provide the requested statistics based on the following data set.

X	Y
2	4
7	12
3	7
7	14
5	11
3	5

a Calculate the slope of the regression line for predicting Y from X.
b Establish the regression equation for predicting Y from X.
c Calculate the corresponding s_e using the computational formula.

21 Provide the requested statistics based on the following data set.

X	Y
12	4
4	11
13	7
6	14
9	3
8	9

a Calculate the slope of the regression line for predicting Y from X.
b Establish the regression equation for predicting Y from X.
c Calculate the corresponding s_e using the computational formula.
d Determine Y_p when $X = 1$.

22 In a study on pain tolerance, a researcher is interested in predicting the amount of time that participants are able to keep their hands in ice-cold water. Based on previous research, the researcher knows that vitamin E intake over the past 12 hours is correlated with tolerating a painful stimulus, at least when the stimulus is freezing water. The following table lists the pairs of scores for the study sample.

Vitamin E: (*X*)	Tolerance Times (in seconds): (*Y*)
5	23
9	32
22	65
12	40
16	42

a What is the *b?*
b What is the Y intercept?
c What is the s_e*?*
d What tolerance time should we predict for someone who has taken 16 units of Vitamin E the morning of the study?
e Within what time interval should we be 68% confident lies the right answer?

23 A sociologist is interested in predicting yearly income (Y) based on prior education level (X), with education level defined as the number of years of formal schooling. The following data were collected from six individuals.

Education: (X)	Income ×1000: (Y)
10	15
14	29
9	14
14	37
12	20
13	23

a What is the b?
b What is the Y intercept?
c What is the s_e?
d What income should we predict for someone with ten years of education?
e Within what pay range should we be 95% confident lies the right answer?

24 An admissions committee needs to predict whether a particular student will be able to make passing grades during the first year of graduate school. In order to make it past the first year, the student will have to achieve a 3.00 GPA in the first semester of graduate school. The admissions committee has data from past years on the relationship between undergraduate GPA and the subsequent first-year graduate school GPA. Those data are as follows.

Undergraduate GPA: (X)	Graduate School GPA: (Y)
3.50	3.33
3.98	3.63
3.10	3.40
2.90	3.41
3.40	3.40

a What is the b when predicting graduate school GPA?
b Should they admit a student with a 3.00 GPA?
c Of the incoming students with a GPA of 3.67, what is the graduate school GPA predicted range for the middle 68%?

25 (This problem uses the research scenario and data from Chapter 15, question 31.) A researcher is interested in the relationship between smoking and illness. A sample of 13 smokers in a large office is randomly selected and asked to report the average number of cigarettes they smoke per day. The researcher then obtains the company records that monitor the number of sick days each employee has taken over the past six months of employment. Run a regression analysis predicting the number of sick days, and answer the following questions.

Number of Cigarettes	Number of Sick Days
11	1
10	1
26	5
15	3
9	2
16	2
20	2
8	1
3	0
24	4
21	6
5	0
14	3

a What is the slope?
b What is the regression equation?
c What is the Y intercept?
d What is the s_e?
e What number of sick days should be predicted for an employee who smoked 15 cigarettes a day?
f What is the 68% confidence range of missed days for someone who smokes 10 cigarettes per day?

26 (This problem uses the research scenario and data from Chapter 15, question 32.) A psychologist is interested in the relationship between

Intelligence and Word Processing Speed on a keyboard. The gathered data from 12 participants are below. Either use a computer program to test the null hypothesis that $b = 0$, or find the correlation. Should a regression analysis be run on this data? Why or why not?

Intelligence	Word Processing Speed
108	28
96	46
90	55
111	40
119	34
105	38
98	57
93	47
117	48
127	73
101	56
103	48

27 Suppose another researcher, accessing a different population (older individuals who originally learned how to type on a typewriter), found the following bivariate data regarding intelligence and typing speed (in words per minute). It was found that $r(10) = .86$. Run a regression analysis predicting Word Processing Speed, and answer the following questions.

Intelligence	Word Processing Speed
109	48
94	36
100	45
117	59
104	39
90	32
116	57
87	37
122	53
88	46
126	59
101	45

a What is the slope?
b What is the regression equation?
c What is the Y intercept?
d What is the s_e?
e What words per minute should be predicted for a person with an IQ of 100?
f What is the 95% confidence range of words per minute for this person?

28 Gerson, Plagnol, and Corr (2016) examined the relationship between social media use and happiness (known as "subjective well-being in the professional literature"). Examine the manufactured data below which has been designed to reflect merely one aspect of the study (use of social media for social comparison purposes). Higher subjective well-being numbers mean greater happiness. Run a regression analysis predicting subjective well-being, and answer the following questions.

Using social media for social comparison purposes

Use (hours/day)	Subjective well-being
3.1	3
2.8	4
1.8	6
3.0	4
3.3	2
1.5	7
0.8	6
2.2	4
1.8	6
0.6	8

a What is the slope?
b What is the regression equation?
c What is the s_e?
d What subjective well-being should be predicted for a person who spends three hours a day on social media for social comparison purposes?
e What is the 95% confidence range of subjective well-being for this person?
f Any methodological concerns?

29 (Based on Chapter 15, Problem 33.) Compute the regression equation and s_e for both conditions. If, during drilling, the dentist rates the patient's discomfort as 7, what should we predict is the patient's discomfort rating? Answer the same question for the rubber dam condition, and report the standard error of the estimate.

30 (Based on Chapter 15, Problem 34.) For a woman who reports 54 menstrual symptoms, how many symptoms will she report during pregnancy? Identify two values of pregnancy symptoms, between which we should find the correct number of pregnancy symptoms for 68% of the women who report 54 menstrual symptoms.

Part 6 Review

Linear Correlation and Linear Regression

Review of Concepts Presented in Part 6

As with previous review sections, the purpose here is to revisit both the similar concepts that hold Chapters 15 and 16 together and the concepts that distinguish them one from another. First let us look at the numerous similarities. Although previous chapters have dealt with the measurement of more than one data point per participant (Chapter 10, "Dependent-Samples *t* Test," and Chapter 14, "Repeated-Measures ANOVA"), there was only one variable being measured. These previous chapters involved the use of repeated-measures designs; the repeated measurement of the same variable under different conditions. In Chapters 15 and 16, the concept of bivariate data is introduced, where two different variables are measured for each participant. The purpose of measuring bivariate data is to explore the nature and strength of the relationship between them. Both the correlation and regression concept help achieve that goal.

Unfortunately, when describing relationships between variables, it is tempting at least to think of, if not state, a specific causal structure for the relationship. However, the same cautions introduced in Chapter 1, as well as at other places in the text, are in effect here as well. The language of causality is restricted to data gathered experimentally, where a hypothesized causal variable is controlled and manipulated by the experimenter and a dependent variable is carefully measured. Typically, bivariate data is gathered in nonexperimental situations and so causal language is not warranted; however, that is not necessarily the case.

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

Another common concept is the notion of shared variance, the degree to which the variance associated with one variable corresponds to variance found in a second variable. It is captured in both chapters, as the coefficient of determination (r^2) in Chapter 15, and extrapolated into a measure of prediction error (s_e) in Chapter 16.

There are also many common pitfalls associated with analyzing bivariate data. Although there are advanced techniques associated with both the correlation and regression procedures to allow for the analysis of nonlinear relationships between variables, the introductory concepts presented in these chapters can only be used on bivariate data reflecting a linear relationship. Furthermore, bivariate data sets that do not capture the entire range of scores for each variable can lead to a mischaracterization of their relationships. Finally, the appropriateness of including or excluding outlier data can become problematic, either underrepresenting or overrepresenting the strength of the relationship between the variables.

A final area of commonality concerns the visual representation of the data, typically in the form of a scatter plot, on a two-dimensional graph. Although not meant to replace the necessary role for a mathematical analysis, especially when making decisions regarding null hypotheses, a visual representation of the relationship can help make sense of the data in a number of nonquantifiable ways. Most notably, a visual representation is necessary to avoid many of the pitfalls previously mentioned.

There are also several areas of distinction between the two chapters. First, we should recognize that the regression concept is built upon the correlation concept. It takes the idea of quantifying the shared variance between two variables, extends it, and applies it to a specific purpose – the prediction of unknown values.

Another distinction concerns variable designation. Correlations are bidirectional. The assignment of X and Y to the variables is inconsequential to the outcome; the r will be the same either way, as will a decision regarding the null hypothesis. Although a regression analysis allows for both the prediction of Y from X and the prediction of X from Y, the mathematics used to make those predictions require the proper specification of a predictor and predicted variable; otherwise the conclusions will be in error.

Each technique also has a distinct approach to testing a null hypothesis, an important objective if the data is to be used either as statistical evidence for a relationship or as a predictor variable. With correlational analyses, the null hypothesis of no relationship ($\rho = 0$) can be tested. With a regression analysis, the null hypothesis of no slope ($b = 0$) can be tested. This is being identified as a distinction between the two concepts, since the symbols that are used differ. However, underlying both tests is the same mathematical question. As a result, one test is not more powerful than the other.

Finally, although many of the assumptions behind both inferential tests are shared, a regression analysis includes a couple additional ones, namely, the normality of conditional distributions and the concept of homoscedasticity. A regression analysis assumes the distribution of Y data corresponding to a given value of X to be normally distributed across the range of the X variable, and it assumes those conditional distributions will have similar measures of dispersion.

Since real-world research problems do not come with a label informing the researcher of which test to use for analysis, it is important for us to work on our diagnostic skills. Understandably, the exercises at the end of each particular chapter only require the use of the tests found and studied within that chapter for solution. The work exercises at the end of chapters are designed to get us familiar with using the tools just described to solve a statistical problem. They are not designed to challenge our diagnostic skills (i.e. knowing which test to use for a given situation). The following review section is designed to help us develop these abilities.

The questions and exercises below will help us review the statistical differences between the correlation and regression concepts and will also continue the work of helping us recognize the appropriate diagnostic cues for the proper application of *all* of the tests presented up to this point in the text. In keeping with this goal, the hypothesis testing exercises will not identify which test is appropriate for the described scenario. We will need to use the available information presented in the exercise to make that determination. (Note: Most of the exercises below can be solved either with or without the use of statistical software.)

Questions and Exercises

1 Select the proper statement.
- **a** Correlation is to regression as relationship is to prediction.
- **b** Correlation is to regression as prediction is to relationship.
- **c** Correlation is to regression as independent variable is to dependent variable.
- **d** Correlation is to regression as dependent variable is to independent variable.

2 What does a regression analysis add above and beyond what is learned from a correlational analysis?

3 Sarah is an above-average tennis player but decides to change sports and take up ping-pong.

a What information is needed to see if Sarah can expect to be an above-average ping-pong player?
b What additional information is needed to predict how Sarah will eventually perform in this new sport?

4 Cheryl is an above-average runner but decides to change sports and take up wrestling. If there is no known relationship between running and wrestling, predict how Cheryl will rate as a wrestler.

5 Describe in what way r is related to prediction error.

6 If a researcher wanted to investigate the strength of the relationship between a measure of political conservatism and a person's degree of philanthropic giving, what statistical tool seems most appropriate? Why?

7 If a researcher wanted to investigate the relationship between political attitudes and philanthropy by classifying people as either liberals, conservatives, or independents and then investigate their charitable giving as recorded on their tax returns, what statistical tool seems most appropriate? Why?

8 If a researcher wanted to predict charitable giving based on the level of political conservatism espoused by an individual, what statistical tool seems most appropriate? Why?

9 If a researcher wanted to investigate the relationship between biological sex, political attitudes, and philanthropy by classifying people as either male or female and as either liberals, conservatives, or independents and then investigate their charitable giving as recorded on their tax returns, what statistical tool seems most appropriate? Why?

10 For the following bivariate data set, answer the following questions:

X	Y
3	9
5	4
7	0
4	2
9	0
2	10
5	5

(Continued)

X	Y
4	7
2	8
1	8

a What is the correlation between X and Y?
b What is Y_p if $X = 10$?
c Can the null hypothesis that $\mu_X = \mu_Y$ be rejected?

11 Imagine the following data set is gathered regarding biological male–female sibling pairs and their average daily time spent doing leisurely activities.

Male	Female
3.9	4.2
1.6	1.9
2.9	1.6
3.5	2.3
3.3	2.1
4.9	4.0
5.5	3.5
4.4	5.3
3.7	2.1
3.5	2.0

a Test the null that there is no difference between biological males and biological females regarding daily leisure time.
b Test the null that there is no relationship for time spent in leisure between siblings.
c What if we know our friend, Jimmy, spends two hours playing video games every day in an otherwise highly structured life, how many hours might Jenny, his sibling, have to engage in a leisurely activity?

12 Suppose we are interested in the relationship between the number of siblings a child has and the age at which they learn to walk. We ask mothers of seven two-year-olds how many older siblings the child has and the age the

child began to walk. Data are listed below. Answer the questions that follow.

Number of siblings	Age of walking (months)
1	16.0
2	8.5
0	16.5
3	10.0
1	11.0
2	10.0
4	9.0

a Is there evidence of a relationship?
b If so, how strong is the relationship?
c Suppose a new child enters into the study with five older siblings. When might they begin to walk?
d How confident are we in our prediction?

13 Some recent research suggests early retirement may lead to memory decline (Rohwedder and Willis, 2010). The researchers gathered data based on a memory test given to numerous individuals aged 60–64 from a variety of countries throughout the world. An average memory score for each country cohort was generated as well as national statistics on the percentage of retired people in the same age window. The data below are similar to what the researchers found.
a Is there a relationship between memory performance for those aged 60–64 and the percent of retirees in that same age window?
b What analysis could be used to forecast what is going to happen to memory scores if the number of early retirees increases in a given country?
c As a country goes from 20 to 40% early retirees, what sort of memory decline could be expected?
d Is the causal relationship between the variables clear?

Country	Percent retired	Memory score
Sweden	39	9.8
United States	47	10.6
Switzerland	47	9.6

(Continued)

Country	Percent retired	Memory score
Denmark	58	10.4
England	59	10.5
Greece	69	8.4
Germany	69	9.3
Spain	73	6.3
Netherlands	77	9.3
Italy	81	7.7
France	86	8.0
Belgium	87	8.3
Austria	90	9.1

14 A soft drink company believes their new energy drink actually helps users lose weight responsibly. To test this hypothesis, 15 adult consumers are gathered, their current weight is measured, and they are asked to use the soft drink company's product at least twice a day. After six months, the participants are contacted and weighed. The pre-measure is subtracted from the post-measure to create a variable called "weight change." If we assume the body weight of adults does not change over time, what test should be used to evaluate the null hypothesis?

Part 7

Inferential Statistics

Nonparametric Tests

17

The Chi-Square Test

17.1 The Research Context

Starting with Chapter 8, we have been discussing a variety of different inferential tests. However, all of the tests up to this chapter require an interval or ratio scale of measurement and involve finding means as well as other mean-based statistics (e.g. the variance). These scales of measurement are used for continuous quantities and capture a sense of *how much* more or less one entity is compared with another (see Chapter 2). However, not all research questions can be answered by using continuous measures; instead of asking "how much," many research questions ask "how many." The following are example of research situations are examples in which the data come in the form of a **frequency count**:

► **Example 17.1** A developmental psychologist hypothesizes that fear of strangers occurs more often at a certain age. A random sample of children, ages 2 through 6, is taken. For each age category, the number of children who are afraid of strangers is counted. The dependent variable is *not* the amount of fear the children have but rather how many children are afraid. Each child is categorized based on the presence or absence of fear. ◄

► **Example 17.2** A political scientist hypothesizes that students grow more politically liberal over their four years of undergraduate education. A random sample of freshmen and seniors is obtained, and the number of students reporting liberal and conservative views is tabulated for each class. Once again, the dependent variable is *not* how liberal or conservative are the students. Each student is classified as *either* liberal *or* conservative. ◄

► **Example 17.3** A health psychologist hypothesizes that the percentage of urban dwellers who struggle with anxiety issues is higher than those who live in rural areas. A random sample of urban and rural residents is gathered and assessed. A frequency count of the number (percentage) of those who suffer with anxiety-related issues in each environmental setting is tabulated. ◄

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

► **Example 17.4** A social psychologist is interested in the relationship between obedience and service in the armed forces. A random sample is obtained of former service personnel as well as individuals who never served in the military. An experimental task is administered that allows the researcher to assess *whether* the participant will follow an unpleasant order. The data are collected in the form of a frequency count of the number of participants who obey or disobey the order. ◄

► **Example 17.5** A clinical psychologist hypothesizes that schizophrenics prescribed with medication after being discharged from the hospital are less likely to be rehospitalized compared with released schizophrenic patients not maintained on medication. A year after discharge, the number of medicated and nonmedicated patients who were and were not rehospitalized is counted. ◄

In each of the preceding research contexts, the data come in the form of a frequency count that is tabulated for each category. In addition, note that the frequency count data have a discontinuous either-or quality. For example, the child either is afraid of strangers or is not afraid of strangers, the student is either liberal or conservative, the participant either obeys or does not obey, and so on. Recall from Chapter 2 that variables having an either-or quality are measured on a nominal scale. Because nominal data are arranged by categories, nominal data are oftentimes referred to as **categorical data**.

A statistical test that analyzes categorical data is the **chi-square test** (pronounced *kigh* square, symbolized as χ^2). One important difference between the chi-square test and the previously discussed tests is that the chi-square test makes no assumptions about population parameters or population characteristics for its use. For this reason, the chi-square test is one example of a **nonparametric test.** Tests that *do* make assumptions about population parameters are known as **parametric tests.** For example, the F test assumes that the population distributions are normally distributed and have equal variances. If these assumptions are grossly violated, interpretations of the test results can be misleading. Nonparametric tests do not make assumptions about the shape of population distributions; for this reason, they are sometimes referred to as **distribution-free tests**.

There are times when an investigator uses a scale of measurement that would normally lead to the use of a parametric inferential test. However, if the assumptions of the test are not met, the data can be transformed and analyzed using a nonparametric test. The decision to switch from a parametric to a nonparametric test should not be taken lightly. Nonparametric tests are generally not as powerful as parametric tests; this means it is more difficult to reject a false null hypothesis when using a nonparametric test. If given the choice to conduct either a parametric or a nonparametric test, the parametric alternative should be chosen; the Type II error rate will be smaller. Of course, if the data are based

on a nominal or ordinal scale, there is no alternative; a nonparametric test must be used. Thankfully, the chi-square tests, although nonparametric, are still rather powerful. A more detailed discussion of parametric versus nonparametric tests is provided in Chapter 18. The following sections present two versions of the chi-square test, one for single-factor research situations and another for two-factor research situations.

17.2 The Chi-Square Test for One-Way Designs: The Goodness-of-Fit Test

The goodness-of-fit test is the categorical counterpart to the one-way ANOVA. It is designed to detect differences among a set of categories falling along a single factor. In these methodological situations, an ANOVA is typically used if the measuring scale is interval or ratio; the chi-square goodness-of-fit test is used if the measuring scale is nominal.

Chapter 2 defined a frequency distribution as the number of observations for each score in a distribution. When using nominal data, the frequency distribution is the number of observations per category. The chi-square test uses the frequency distribution of a sample to make an inference about the frequency distribution of a population. The **goodness-of-fit test** uses the chi-square statistic to analyze how well the sample data "fit" (correspond) with the hypothesized frequency distribution. In doing this, we must first state a null hypothesis that indicates what the population data would look like if there were *no* effect. The frequencies of the distribution specified by the null hypothesis are called **expected frequencies**, symbolized as f_e. The frequencies of the distribution obtained from the sample are called **observed frequencies**, symbolized as f_o.

The Null and Alternative Hypotheses for the Chi-Square Test for Goodness of Fit

The goodness-of-fit test requires us to specify the population frequency distribution that will be used as the null hypothesis. Frequencies are typically presented as percentages. How do we arrive at the null hypothesis; that is, what frequencies should be specified for each category? Hypothesized population frequencies can be determined either *rationally* or *empirically*. First, consider the rational approach. Suppose we would like to find out if people prefer cola *A* or *B*. We conduct a blind taste test with 100 participants. If there is no preference among the participants between the colas, what percentage of people would we expect to choose cola *A*, and what percentage would we expect to choose cola *B*? On the average, we would expect 50% of the participants to select cola *A* and 50% to choose cola *B*. This null hypothesis can be represented as follows:

	A	*B*
H_0	50%	50%

Now suppose we add a third drink, cola *C*. What would be the expected frequency distribution for a null hypothesis of no preference? The answer is

	A	*B*	*C*
H_0	33.33%	33.33%	33.33%

If we add a fourth beverage, the expected frequencies would be 25% for each category. A null hypothesis of no preference states that the expected frequencies are equally distributed across the categories. The alternative hypothesis states that the population frequencies are not distributed equally across the categories.

Another example of the rational approach to specifying the null hypothesis would be the distribution of expected frequencies as predicted by a theory. A genetic theory of ulcer susceptibility might predict the percentage of rats that will develop ulcers under stress. The theory might hypothesize that after four generations of inbreeding, 40% of the offspring of rats will show stomach ulcers under stress. This null hypothesis can be represented as

	Ulcers	
	Yes	**No**
H_0	40%	60%

The alternative hypothesis states that the population distribution of frequencies is not distributed in the expected manner.

The *empirical* approach to specifying expected frequencies requires existing data, although not necessarily the data that we have collected. For example, suppose the percentage of people who voted Democratic in the last gubernatorial election was 60%, whereas 30% voted Republican and 10% of the population voted Other. Since the last election, however, a popular Republican president campaigned for the present Republican gubernatorial candidate. Before the election, several thousand voters are polled to determine if the percentage of people in the Democrat, Republican, and Other categories has changed. Therefore, using the data from the last election, the null hypothesis would be stated as:

	Democrat	**Republican**	**Other**
H_0	60%	30%	10%

The alternative hypothesis states that the relative frequency of voters across the categories differs from the last election.

Computing the Chi-Square Statistic for the Goodness-of-Fit Test

The purpose of the goodness-of-fit test is to determine if the observed frequencies, obtained from a sample of participants, meaningfully differ from the expected frequencies derived from an understanding of the null hypothesis. Recall from previous discussions of null hypotheses that it is *highly* unlikely that sample means will be identical *even when* the H_0 is true. In much the same way, simply finding that the observed frequencies differ from the expected frequencies is itself not a sufficient reason to reject the null hypothesis. The chi-square goodness-of-fit test allows an investigator to determine the likelihood that the difference between the expected and observed frequencies is due to chance. The following scenario will show us how to compute a goodness-of-fit chi-square test.

Two students are discussing the reasons why their peers choose a particular undergraduate major. The psychology student makes the somewhat provocative statement that business majors select their program of study because they are most interested in making money. The other student, a business major, insists that money is one motive, but not the most important motive for entering the business world. They decide to conduct a study in which business majors are asked to respond to one statement: "Making money is the most important reason for majoring in business." Participants are told to either Agree, Disagree, or be Undecided about the statement. One version of a rationally derived null hypothesis states that the percentage of responses in each category is the same.

Agree	Disagree	Undecided
33%	33%	33%

Although the null hypothesis is typically stated in terms of percentages, the chi-square test is not performed using percentages. Specifying the exact expected frequencies in each category requires knowledge of the total number of participants in the study. Assume 90 participants were used. To determine the exact (expected) frequency for a given cell (category), f_e, each percentage (or proportion) is multiplied by n, the number of participants.

$$33\% \text{ of } 90 = 0.33(90) = 30 \text{ Agree responses}$$

$$33\% \text{ of } 90 = 0.33(90) = 30 \text{ Disagree responses}$$

$$33\% \text{ of } 90 = 0.33(90) = 30 \text{ Undecided responses}$$

Instead of converting percentages into frequencies, the f_e for each category can also be found by dividing the number of participants (observations) by the

number of categories: $N/C = 90/3 = 30$. This method only works, however, when the expected frequencies for each cell are hypothesized to be the same. There will be research situations when the null hypothesis will state different percentages for each cell. In these situations, the expected frequencies will differ from cell to cell. Familiarity with the null hypothesis is critical for the goodness-of-fit chi-square test.

The next step in computing the χ^2 statistic is to find the observed frequencies, f_o. This is accomplished by simply counting the number of participants in the sample who agreed, disagreed, and were undecided. With expected and observed frequencies calculated, Formula 17.1 is used to compute χ^2.

Formula for χ^2

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \quad \text{(Formula 17.1)}$$

Assume the data show observed frequencies in each category as indicated in the following.

Agree	Disagree	Undecided
$f_e = 30$	$f_e = 30$	$f_e = 30$
$f_o = 60$	$f_o = 20$	$f_o = 10$

In using Formula 17.1, arithmetic operations are performed for each cell:

$$\frac{(f_o - f_e)^2}{f_e}$$

After we obtain this value for each cell, the Σ symbol in Formula 17.1 directs us to sum all the values in each category.

Computational Steps

Step 1. For the first category, subtract the expected (hypothesized) frequency from the observed frequency (the data), $f_o - f_e$.

Step 2. Square the difference, $(f_o - f_e)^2$. This removes any negative signs.

Step 3. Divide the number found in step 2 by the f_e *specified for that cell,* $(f_o - f_e)^2/f_e$.

Step 4. Repeat the first three steps for each cell.

Step 5. Sum all the quantities from all the categories.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\begin{array}{lll} \textit{Agree} & \textit{Disagree} & \textit{Undecided} \end{array}$$

$$\chi^2 = \frac{(60-30)^2}{30} + \frac{(20-30)^2}{30} + \frac{(10-30)^2}{30}$$

$$\chi^2 = \frac{900}{30} + \frac{100}{30} + \frac{400}{30}$$

$$\chi^2 = 30 + 3.33 + 13.33$$

$$\chi^2 = \mathbf{46.66}$$

Deciding Whether to Reject the Null Hypothesis

This section shows us the steps involved in deciding whether to reject the null hypothesis. The next section addresses the characteristics of the sampling distribution of the χ^2 statistic. This section will give us a deeper understanding of how hypotheses are tested in chi-square analyses.

In viewing the formula for χ^2, it should be clear that χ^2 will become larger as the difference between the expected and observed frequencies increases.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

A close fit between the expected frequency distribution and the observed frequency distribution will lead to a relatively small χ^2.

Similar to other statistical tests, the larger the χ^2 statistic, the less likely it will be generated from a null distribution. To see if the χ^2 is large enough to reject the null hypothesis, it is compared with a critical value, χ^2_{crit}. If χ^2_{obt} is equal to or greater than χ^2_{crit}, the null hypothesis is rejected; statistical evidence has been found suggesting a population consistent with the null hypothesis did not generate our observed frequencies. The difference between the obtained and expected values is probably *not* due to chance.

The critical value for χ^2 is found by using Table A.8, a portion of which is shown in Table 17.1.

The degrees of freedom for the goodness-of-fit test are the number of columns in the design minus one. In the "reason for majoring in business" example, the number of columns (categories) is equal to 3. Therefore, $df = C - 1 = 3 - 1 = 2$. What would be the critical value for χ^2 if alpha were set at .05? Referring to Table 17.1, the answer is 5.99. Since the obtained $\chi^2 = 46.66$ and $46.66 > 5.99$, we would reject the null hypothesis.

We will need to wait until later in the chapter to learn about follow-up analyses, but for now we can state that statistical evidence suggests the opinions of business majors were not equally distributed across the three options, $\chi^2(2, n = 90) = 46.66$, $p < .05$.

Table 17.1 A portion of the table of critical values for the chi-square distribution.

Proportion of critical region					
	Alpha Level				
df	***.10***	***.05***	***.02***	***.01***	***.001***
1	2.71	3.84	5.41	6.64	10.83
2	4.60	5.99	7.82	9.21	13.82
3	6.25	7.82	9.84	11.34	16.27
4	7.78	9.49	11.67	13.28	18.46
5	9.24	11.07	13.39	15.09	20.52
6	10.64	12.59	15.03	16.81	22.46

Consider another worked problem using the goodness-of-fit test, this time where the expected cell frequencies are not all the same.

■ **Question** *A professor states papers are graded using the following categories: excellent, above average, average, and below average. Furthermore, the professor maintains that turned-in papers are graded graciously and offers the following distribution of percentages as an estimate of the manner in which the grades are distributed.*

Excellent	Above Average	Average	Below Average
25%	35%	25%	15%

A group of students suspects the professor may indeed be generous but only in the perception of being an easy grader. All of the students who had previously taken this professor's course in the last three years are available, and amazingly, they still have their papers. A random sample of 100 former students is taken, and the actual distribution of evaluations is recorded.

Excellent	Above Average	Average	Below Average
25%	35%	25%	15%
$f_o = 20$	$f_o = 25$	$f_o = 30$	$f_o = 25$

Note that since the sample size is 100, the observed frequencies sum to 100. Is there a reason to conclude that the professor's claim is mistaken?

Solution Before the formula for χ^2 can be used, the f_e's need to be calculated.

25% = (0.25) 100 = 25 expected in the *Excellent* category
35% = (0.35) 100 = 35 expected in the *Above Average* category
25% = (0.25) 100 = 25 expected in the *Average* category
15% = (0.15) 100 = 15 expected in the *Below Average* category

Placing the expected and observed frequencies in a table shows

Excellent	Above Average	Average	Below Average
$f_e = 25$ $f_o = 20$	$f_e = 35$ $f_o = 25$	$f_e = 25$ $f_o = 30$	$f_e = 15$ $f_o = 25$

Computing χ^2,

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2 = \frac{(20-25)^2}{25} + \frac{(25-35)^2}{35} + \frac{(30-25)^2}{25} + \frac{(25-15)^2}{15}$$

$$\chi^2 = \frac{25}{25} + \frac{100}{35} + \frac{25}{25} + \frac{100}{15}$$

$$\chi^2 = 1 + 2.86 + 1 + 6.67$$

$$\chi^2 = \mathbf{11.53}$$

$$df = C - 1 = 4 - 1 = 3$$

$\alpha = .05$; therefore $\chi^2_{crit} = 7.82$

Since $11.53 > 7.82$, the H_0 is rejected.

As a consequence of rejecting the null hypothesis, we can conclude that statistical evidence has been found suggesting the observed frequency distribution has been generated by a population that is different than the one claimed by the professor. By examining the specific observed frequencies in the cells, we see that the "excellent" category has the fewest number of students. Indeed, the sample shows that most of the students received an "average" on their papers. This is contrary to the professor's claim to be an easy grader. However, the mere fact that the null hypothesis has been rejected does not mean we know *which cells* are causing this rejection. We have yet to introduce a follow-up analytical tool needed for further investigation. Additionally, a rejected null hypothesis does not automatically mean that the professor is mistaken. For instance, suppose that the observed frequency of the "excellent" category was found to be 70, with the remaining 30 frequency counts spread across the other categories. The null hypothesis would be rejected, but the fact that the "excellent" category had such a high count would actually *strengthen* the professor's claim.

After rejecting a null hypothesis, we have to examine the pattern of cell frequencies and perform some additional analyses to interpret properly the meaning of the finding. ■

17.3 The Chi-Square Distribution and Degrees of Freedom

The chi-square statistic indicates how well the hypothesized expected frequencies correspond to the observed frequencies. The closer the fit, the smaller the value of χ^2. The test statistic for the chi-square distribution is χ^2, which is the basis for the sampling distribution, just as the t statistic and F ratio are the basis for the t and F distributions.

As with previous descriptions of sampling distributions, the **chi-square distribution** is theoretical, formed by taking an infinite number of samples from a null population and computing the chi-square statistic for each sample. The relative frequency of each value of chi-square is plotted to show a chi-square distribution. A separate chi-square distribution is created for each *df*, thereby establishing a family of chi-square distributions. The shape of each distribution is defined by the number of categories used to compute χ^2.

Note that the degrees of freedom for a χ^2 test are *not* based on the number of participants in the study. Instead, the number of categories determines the degrees of freedom. Suppose we conducted a study with four categories and 150 participants. If the first three categories contained a total of 100 participants, the number of participants in the fourth category would be automatically determined (50). In other words, the frequency count of three of the four categories is free to vary. Once the count of three categories is specified, the count in the fourth category is strictly determined. As a result, in the simple one-way design, *df* equals the number of categories minus 1, ($df = C - 1$).

Characteristics of Chi-Square Distributions and Rejecting the Null Hypothesis

Rejecting the Null Hypothesis

The null hypothesis for the χ^2 test is specified by the expected frequencies in each cell of the design. The χ^2 test measures the degree to which the observed frequencies correspond to the expected frequencies. If the null hypothesis is true, then the observed frequencies for each cell will be very close to the expected frequencies, and the value of χ^2 will be small. If the null hypothesis is false, the expected and observed frequencies will be discrepant from one another. Even if the null hypothesis were true, because of sampling error, we would not expect the observed and expected frequencies to match perfectly.

As the expected and observed frequencies become increasingly discrepant, χ^2 becomes larger. How large does χ^2 have to be in order to reject the null hypothesis? If the χ^2 computed on the sample data is unlikely to occur when the null hypothesis is true, then we conclude that the null hypothesis is false. As with previous inferential tests, the meaning of "unlikely" is defined by the level of alpha. If $\alpha = .05$, the null hypothesis is rejected if the probability of obtaining a χ^2 statistic of a given size is equal to or less than .05. As a result, each chi-square distribution can be marked with a critical value to identify the percentage of χ^2 values that lie, for instance, in the upper 10, 5, or 1% of the sampling distribution.

Figure 17.1 shows chi-square distributions for 1, 5, and 8 degrees of freedom. The rejection region for each of the distributions when $\alpha = .05$ is also shown in Figure 17.1. An obtained value of χ^2 that falls in the rejection region would allow us to reject the null hypothesis. Note, however, that as the degrees of freedom increase, a *larger* χ^2_{obt} is required to reject the null hypothesis. Take another look at the formula for χ^2.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}.$$

Assume that the null hypothesis is true and, therefore, any discrepancies between f_o and f_e are due to sampling error. Remember that the formula for χ^2 requires that we sum all the categories in the design. What happens as the number of categories increases? Even if the value of $(f_o - f_e)$ for each cell is quite small, summing a large number of cells will lead to a large χ^2, even when the null hypothesis is true. Therefore, for a given level of alpha, as the number of categories increases, a larger χ^2_{obt} is necessary to reject the null hypothesis.

Finally, the χ^2 test is nondirectional. The observed frequencies can fit poorly with the expected frequencies by being either too large or too small. Although the total number of observed and expected frequency counts will be equal, which cells underpredict and which cells overpredict the frequency counts is not something stated by the alternative hypothesis.

Characteristics of the Chi-Square Distribution

The characteristics of the chi-square distribution are as follows:

1) Since the numerator of the χ^2 statistic is squared, all values of χ^2 are positive.
2) Chi-square distributions are unimodal and typically positively skewed. However, as the *df* increases, the chi-square distribution approximates the shape of a normal distribution.
3) As the *df* increases, the critical value of χ^2, beyond which the rejection region lies, becomes relatively larger.

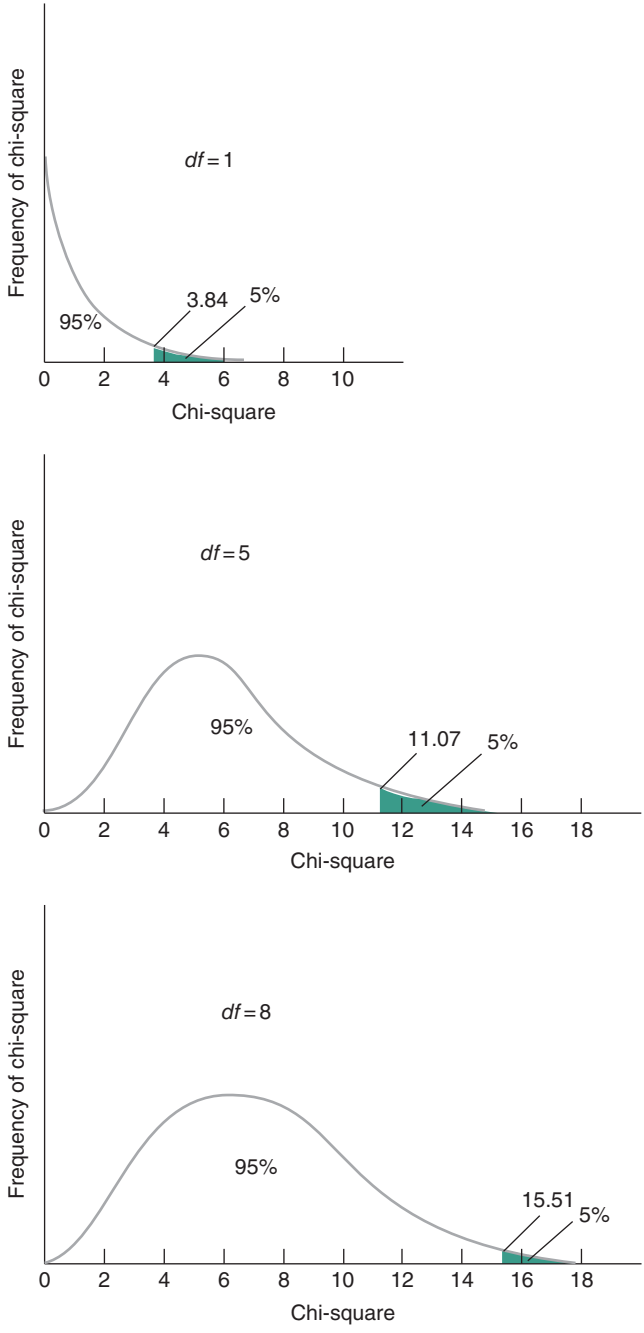


Figure 17.1 Chi-square distributions for degrees of freedom of 1, 5, and 8.

17.4 Two-Way Designs: The Chi-Square Test for Independence

The chi-square test can also be used to determine if there is a relationship between two factors. We may recall that a cell structure featuring two factors was also used by the two-way ANOVA presented in Chapter 13. Similar to the goodness-of-fit test, the chi-square **test for independence** also uses the frequency counts from a set of observations across both factors. Some examples of research questions in which the chi-square test is applied with two variables are as follows:

1) A clinical psychologist hypothesizes that birth complications are associated with a subsequent diagnosis of schizophrenia. Three comparison groups are included in the study: a group of schizophrenic patients, a group of depressed patients, and a group of normal participants. Each participant in the study is categorized on *two* variables: diagnosis and history of birth complications. This design is represented in the following table, called a **contingency table** (also called a *frequency* or *cross-tabulation table*). Since there are two rows and three columns, the table is called a 2×3 contingency table.

	Schizophrenic	Depressed	Normal
Birth Complications	20	6	8
No Birth Complications	8	20	22

The numbers in the cells refer to the number of participants that meet the classification criteria for both variables. For example, 20 schizophrenics were found to have had birth complications, 20 depressed patients did not have birth complications, 8 normal participants had birth complications, and so on.

2) A social psychologist hypothesizes that biological males are more likely than biological females to help someone in an emergency and that helping will be affected by the presence or absence of bystanders. This two-way design, without obtained frequencies, is represented in the following 2×2 contingency table.

	Bystanders	
	Present	Absent
Males		
Females		

3) A health psychologist hypothesizes that individuals who have high cardiovascular fitness are tougher negotiators than those who have low cardiovascular fitness. A sample of college students is classified as fit, not fit, and somewhat fit (based on resting heart rate and blood pressure readings). An experimental task that requires negotiating a conflict is presented to each participant. The outcome is classified as either a win or a loss. The design is represented as a 2×3 contingency table.

	Fit	Not fit	Somewhat fit
Win			
Loss			

The Null Hypothesis and the Concept of Independence

The null hypothesis, H_0, applied to a two-way design states that the two variables are independent. The alternative hypothesis, H_1, states that the two variables are not independent; that is, they are related. This is conceptually similar to testing for evidence of an interaction in a two-way ANOVA analysis (see Chapter 13). Figure 17.2 shows bar graphs of the hypothetical data from the schizophrenia and birth complications example. The illustration on the left shows the relative number of schizophrenics, depressed, and normal participants who had complications associated with their births. Compare this graph with the one adjacent, which depicts the relative number of participants among the diagnostic categories who do *not* have a history of birth complications. Just by "eyeballing" the two graphs (birth complications versus no birth complications), we can see that the pattern of data is different in the birth complications category compared with the pattern of data in the no birth complications category. This indicates that the manner in which the data are distributed for one variable *depends* on the level of the second variable.[1] In this example, the two variables do not appear to be independent.

Now consider Figure 17.3. Hypothetical data have been used to reflect independence between the diagnostic category and birth complications category. The graph on the left is now quite similar to the graph on the right in the pattern of the bars representing relative frequency. In other words, the likelihood of someone being classified as schizophrenic, depressed, or normal does not appear to be related to the presence or absence of birth complications. Stated differently, the frequency distribution for one variable has the same pattern for

1 The word "depends" in this context does not imply a causal relationship between the variables.

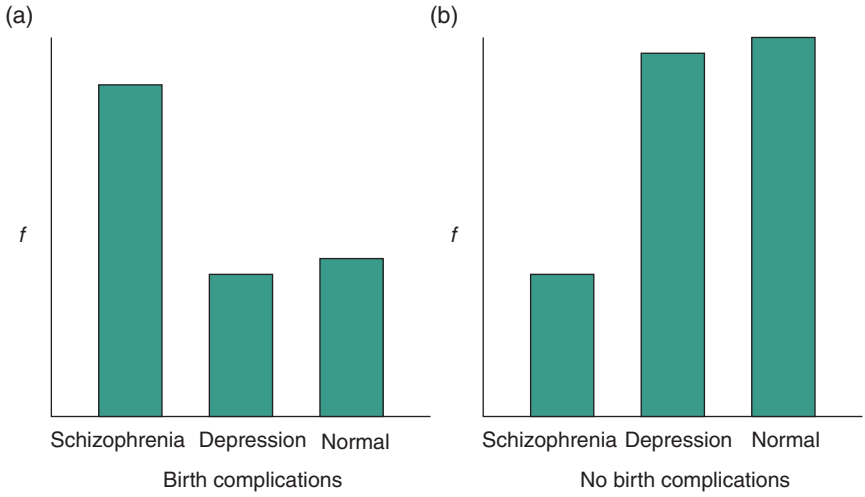


Figure 17.2 Each bar reflects the number of people given one of the three diagnoses. The pattern of bars in (a) is different from the pattern of bars in (b). This indicates that the distribution of various diagnoses *depends* on the presence or absence of a history of birth complications. Viewed together, these graphs reflect a relationship between diagnostic category and birth complications, with schizophrenics showing a greater frequency of birth complications.

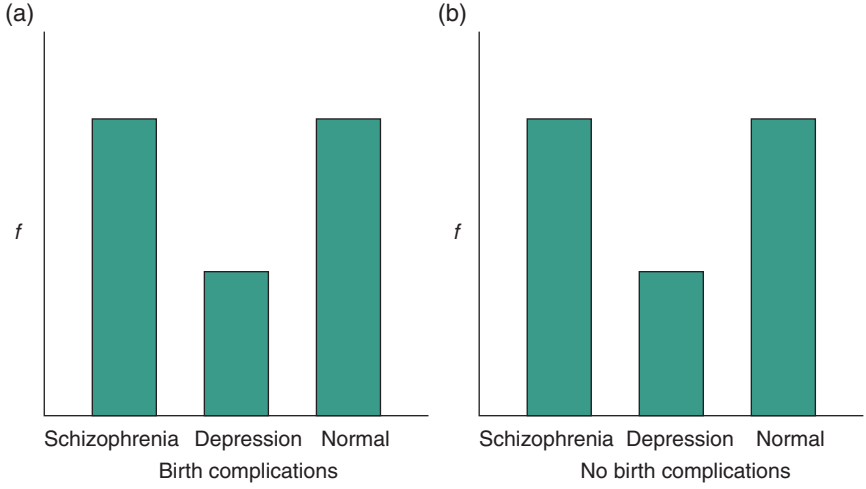


Figure 17.3 Unlike Figure 17.2, the relative occurrence of different diagnoses is similar in (a) and (b). This reflects the fact that birth complications and diagnostic category are independent, that is, unrelated.

each level of the second variable. The chi-square test for independence checks to see if the frequency distribution for one classification variable is different, depending on the level of the second classification variable. If the distributions are different, then the null hypothesis that the variables are independent is rejected.

The failure of the test for independence to reject the null hypothesis for values represented in Figure 17.3 informs us that this analysis is restricted to *between* categories and not *within* the conditions of a category. In this example, we find those who are depressed to be less frequent to those who are schizophrenic and normal. Although the chi-square test for independence corresponds with a two-way ANOVA in terms of investigating an interaction, it does not investigate main effects.

Computing χ^2 for a Two-Way Design

Whether we are conducting a goodness-of-fit test or testing the independence of two variables, the formula for χ^2 is the same:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Moreover, observed frequencies are still obtained from sample data. Specifying the expected frequencies under the null hypothesis, however, is not as straightforward as it was with the goodness-of-fit test.

Computing Expected Frequencies in the Two-Way Design

Computing the proper f_e value for each cell requires a bit of arithmetic. Let us use new frequency count data from the schizophrenia and birth complications study as a working example to illustrate the calculation of expected frequencies. The observed data should be organized, counted, and presented in a two-factor grid such as Table 17.2. From this observed data, we can generate the expected data to test the null of independence.

Using the observed data, marginal frequency counts need to be determined. This is done by simply adding up the frequency counts in all conditions associated with each row and each column (see Table 17.3). Additionally, the total number of participants can be placed in the lower right-hand corner. In our

Table 17.2 Observed frequencies for a worked problem.

	Schizophrenic	Depressed	Normal
Birth Complications	$f_o = 20$	$f_o = 6$	$f_o = 8$
No Birth Complications	$f_o = 8$	$f_o = 20$	$f_o = 22$

Table 17.3 Observed, expected, and marginal frequencies for a worked problem.

	Schizophrenic	Depressed	Normal	
Birth Complications	f_o = 20 f_e = 11.33	f_o = 6 f_e = 10.52	f_o = 8 f_e = 12.14	**34**
No Birth Complications	f_o = 8 f_e = 16.67	f_o = 20 f_e = 15.48	f_o = 22 f_e = 17.86	**50**
	28	**26**	**30**	**84**

example, the total number of participants is 84. From the marginal means, we can determine various values such as the total number of participants having birth complications is 34, the total number of participants having a diagnosis of schizophrenia is 28, and the total number of participants having no psychiatric diagnosis is 30.

Table 17.3 also shows the expected frequencies for each cell of the matrix. The following explains how the expected frequency is calculated for the uppermost left cell, which is the cell that corresponds to schizophrenic *and* birth complications. In computing this expected frequency, we only need to consider the Birth Complication row and the Schizophrenic column. If we were to take a participant at random from the total number of participants in the study, what is the probability that that participant would have had a complicated birth? Since there is a total of 84 participants and 34 of them had birth complications, the answer is **34/84 = 0.4048.** Now consider just the Schizophrenic column. What is the probability that a participant selected at random from the entire study sample would have a diagnosis of schizophrenia? Since there are 28 schizophrenics in the total sample of 84 participants, the answer is **28/84 = 0.3333.** To find the probability of someone being diagnosed as schizophrenic *and* having had birth complications, multiply (0.4048)(0.3333) to arrive at **0.1349.** The expected frequency for this cell is the number of people we would expect to find in this cell if the null hypothesis is true: **0.1349(84) = 11.33.** The formula below reflects a simplified version of the process described above.

Formula for computing f_e

$$f_e = \frac{f_c f_r}{N} \quad \text{(Formula 17.2)}$$

where

f_c = the frequency total for the relevant *column*
f_r = the frequency total for the relevant *row*
N = the total number of participants

Formula 17.2 can be used to compute the expected frequencies for each of the cells represented in Table 17.3.

Schizophrenic and Birth Complications

$$f_e = \frac{(28)(34)}{84} = \mathbf{11.33}$$

Depressed and Birth Complications

$$f_e = \frac{(26)(34)}{84} = \mathbf{10.52}$$

Normal and Birth Complications

$$f_e = \frac{(30)(34)}{84} = \mathbf{12.14}$$

Schizophrenic and No Birth Complications

$$f_e = \frac{(28)(50)}{84} = \mathbf{16.67}$$

Depressed and No Birth Complications

$$f_e = \frac{(26)(50)}{84} = \mathbf{15.48}$$

Normal and No Birth Complications

$$f_e = \frac{(30)(50)}{84} = \mathbf{17.86}$$

Computing χ^2 and Testing the Null Hypothesis

The task of computing χ^2 is easier if we work from a frequency summary table. In Table 17.4, the observed and expected frequencies are ordered in the first two columns. The next two columns are interim calculations necessary for computing χ^2. The value of χ^2 computes to 16.77.

The degrees of freedom associated with a two-way design are $df = (R - 1)(C - 1)$, which is the number of rows (R) minus one multiplied by the number of columns (C) minus one. For the foregoing example, $df = (2 - 1)(3 - 1) = 2$. (Notice that the df for the goodness-of-fit test used C to symbolize the number of categories. Here C symbolizes the number of columns.)

The critical value for the chi-square test is found in Table A.8. If alpha is set at .05, the critical value associated with $df = 2$ is 5.99. Since 16.77 is larger than 5.99, we have statistical evidence to reject the null hypothesis. Therefore, we have reason to believe that a dependency exists between psychiatric diagnosis and birth complications.

Table 17.4 Computing χ^2.

f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
20	11.33	8.67	75.17	6.63
6	10.52	–4.52	20.43	1.94
8	12.14	–4.14	17.14	1.41
8	16.67	–8.67	75.17	4.51
20	15.48	4.52	20.43	1.32
22	17.86	4.14	17.14	0.96
				$\chi^2 = 16.77$

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \mathbf{16.77}$$

The steps involved in conducting a chi-square analysis for a two-way design are:

Step 1. Specify the null and alternative hypotheses. H_0 is a statement that the two variables are independent. H_1 states that the two variables are not independent.
Step 2. Specify alpha; it is typically .05 or .01.
Step 3. Use Formula 17.2, $f_e = f_c f_r / N$, to compute all expected cell frequencies.
Step 4. Place the observed and expected frequencies in a frequency summary table and perform the computational steps for χ^2.
Step 5. Compute *df by* $(R - 1)(C - 1)$.
Step 6. Find χ^2_{crit} in Table A.8 of the appendix and decide whether to reject the null hypothesis.
Step 7. Interpret the findings.

17.5 The Chi-Square Test for a 2 × 2 Contingency Table

We have just completed an example of how to compute χ^2 for a 2 × 3 contingency table. If the two-way design is in the form of a 2 × 2 table, a shortcut method can be used that does not involve computing expected frequencies for each cell. Formula 17.3 requires us to place the letters *A, B, C,* and *D* in the cells as indicated in the following diagram. In this case, only the observed cell frequencies are needed to compute χ^2.

A	B	A + B
C	D	C + D
A + C	B + D	N

Chi-square formula for a 2 × 2 contingency table

$$\chi^2 = \frac{N(AD - BC)}{(A + B)(C + D)(A + C)(B + D)} \quad \text{(Formula 17.3)}$$

where

A, B, C, D = the observed frequencies, f_o, in each cell
$AD = f_o$ for cell $A \times f_o$ for cell D
$BC = f_o$ for cell $B \times f_o$ for cell C
N = total number of participants

Box 17.1 looks at one of the studies that explored the "What is beautiful is good" stereotype, a very robust finding of social psychology. Formula 17.3 is used for the analysis.

Box 17.1 What Is Beautiful Is Good

Social psychologists have discovered that attractive people are the recipients of a positive social stereotype that can be summarized as "What is beautiful is good." Compared with physically unattractive people, those who are good looking are assumed to be more successful, mentally healthier, smarter, and happier (e.g. Dion et al. 1972; Eagly, Ashmore, Makhijani, & Longo, 1991), as well as a host of other positive qualities (e.g. Segal-Caspi, Roccas, & Sagiv, 2012). Many social psychologists believe this occurs because people are more eager to bond with attractive people; this motive causes us to project desirable attributes onto them (e.g. Lemay, Clark, & Greenberg, 2010). Interestingly, the initial findings in this area were based almost exclusively on paper-and-pencil ratings of pictures of attractive and unattractive "target" persons. Benson, Karabenick, and Lerner (1976) wondered if people's *overt behavior* would be influenced by the attractiveness of another person. More specifically, they hypothesized that males would be more likely to help a physically attractive female in comparison to a physically unattractive female.

Method

The study took place at an airport several decades ago, a time before cell phones. The experimenters placed a completed graduate school application on a shelf in a telephone booth. Attached to the application was a picture of the applicant. In the Attractive condition, the picture was of a female that had been pre-rated by judges to be extremely attractive. In the Unattractive condition, the picture affixed to the application was of a female pre-rated as extremely unattractive. An addressed, stamped envelope accompanied the

application, with a clearly displayed note from the applicant: "Dear Dad, Have a nice trip. Please remember to mail this application before you leave Detroit on your (*time of departure*) flight to New York. Love, Linda." The time of departure was constantly altered to indicate that the flight had already left.

The experimenter surreptitiously observed the male participants enter the telephone booth and categorized each one as either Helpful or Nonhelpful based on the following criteria. First, only participants who looked at the application were included in the study. A helpful response consisted of either mailing the application in a nearby mailbox or turning the application over to an employee at the airport. If the experimenter lost sight of the participant, a helpful response could be determined by whether the application arrived at the experimenter's psychology department, the address that was placed on each envelope. A response was considered nonhelpful if the participant, after looking at the application, either left it in the telephone booth, destroyed it, or left with the application but never mailed it.

The design of this study conforms to a 2×2 contingency table with one factor being Attractive/Unattractive and the other factor being Helpful/Nonhelpful. The dependent variable is in the form of a frequency count; therefore, a chi-square test is appropriate. Observed frequencies for each cell are shown in the following table, and Formula 17.3 is used to analyze the data.

	Attractive	Unattractive	
Helpful	A 55	B 35	A + B 55 + 35 = **90**
Non helpful	C 55	D 71	C + D 55 + 71 = **126**
	A + C 55 + 55 = **110**	B + D 35 + 71 = **106**	N = **216**

$$\chi^2 = \frac{N(AD - BC)^2}{(A+B)(C+D)(A+C)(B+D)}$$

$$\chi^2 = \frac{216[(55)(71) - (35)(55)]^2}{(55+35)(55+71)(55+55)(35+71)}$$

$$\chi^2 = \frac{216(3905 - 1925)^2}{(90)(126)(110)(106)}$$

$$\chi^2 = \frac{216(3\,920\,400)}{132\,224\,400}$$

$$\chi^2 = \frac{846\,806\,400}{132\,224\,400}$$

$$\chi^2_{obt} = \mathbf{6.40}$$

With $\alpha = .05$ and $df = 1$, the critical value of χ^2 is 3.84. The obtained value of χ^2 exceeds the critical value (6.40 > 3.84); therefore, the null hypothesis that states there is no association between helping and the attractiveness of the target person is rejected.

How are the findings interpreted? With the aid of additional analyses, it was reported that statistical evidence was found suggesting that men are more likely to help an unknown female when she is attractive compared with unattractive, $\chi^2(1, N = 216) = 6.40$, $p < .05$.

17.6 A Measure of Effect Size for Chi-Square Tests

Recall that hypothesis tests only indicate the degree of certainty associated with rejecting the null hypothesis; that is, how certain are we that there is an effect? However, the *certainty* of an effect and the *size* of an effect are not the same thing. For instance, with large sample sizes comes great statistical power – power to detect even very small effects. As with other inferential tests, a separate measure of effect size can be helpful when interpreting the meaning of a rejected null hypothesis.

For the chi-square tests, the standard effect size measure is **Cramér's *V*** (or Cramér's *phi* – pronounced "fie," rhymes with *fly*) and is oftentimes symbolized as ϕ (although it may also be represented as V). Once the chi-square has been calculated, this measure of effect size can be a fairly straightforward hand calculation. The formula is

Formula for Cramér's *V*

$$\phi = \sqrt{\frac{\chi^2}{(N)\left(df_{row/column}\right)}} \quad \text{(Formula 17.4)}$$

where

χ^2 = the obtained chi-square value
N = the total number of participants in the study
$df_{row/column}$ = the degrees of freedom for either the rows or the columns, whichever is smaller (For a goodness-of-fit test, simply use the df value.)

This procedure, in effect, generates a number that proportions the obtained chi-square value to the size of the sample. Larger numbers reflect greater effect

sizes; however, what counts as a small, medium, or large effect changes as the degrees of freedom change. In general, values that are .10 or lower are usually considered small effects, and values that are .30 or higher are usually considered large effects. Of course, Cramér's V should only be found if the obtained chi-square has resulted in a rejection of the null hypothesis.

For the psychiatric disorder and birth complications example, Cramér's V would be calculated as follows:

$$\phi = \sqrt{\frac{\chi^2}{(N)\left(df_{row/column}\right)}}$$

$$\phi = \sqrt{\frac{16.77}{84(1)}}$$

$$\phi = \sqrt{0.2}$$

$$\phi = \mathbf{0.45}$$

This would be considered a large effect size.

17.7 Which Cells are Major Contributors to a Significant Chi-Square Test?

In the chapters covering the analysis of variance, we learned that a significant F ratio could be further analyzed with follow-up (sometimes called "post hoc") comparisons. The purpose of these analyses is to locate the source(s) of a significant F ratio. Within the context of the chi-square analysis, a post hoc analysis answers the question, "Which cells are *major contributors* to a significant chi-square value?" Unlike follow-up tests that are used with parametric data, this analysis does *not* contrast two cells to see if they differ from one another. Rather, each cell is analyzed separately to determine which cells make a *major* contribution to the χ^2 value. To do this we must first calculate the residual for each cell. This is determined by subtracting the expected frequency from the observed frequency, $f_o - f_e$. These residuals can be standardized, in much the same way z scores standardize raw scores, by dividing this difference by the square root of the expected frequency (Haberman, 1973). Formula 17.5 mathematically represents this statistic.

Formula for the standardized residual

$$R = \frac{f_o - f_e}{\sqrt{f_e}} \quad \text{(Formula 17.5)}$$

An adjustment can be made to the measure that better takes into account the overall size of the sample. The adjusted standardized residual divides the difference between the observed and expected frequencies by the standard error. Many statistical software packages like SPSS generate this more refined measure. For simplicity purposes, only the standardized residual formula will be presented here.

How large must a cell residual be before it is considered a major contributor to the significant χ^2? The conventional standard suggests an absolute value that *equals or exceeds* 2.00 (i.e. $R \geq |2.00|$) is to be considered a major contributor to the significant chi-square test. Formula 17.5 should only be used after the null hypothesis is rejected. Table 17.5 presents the standardized residuals for each category of the schizophrenia and birth complications study. Although the valence of R is irrelevant when determining the cell significance, it is important for interpreting the meaning of any cell designated to be a major contributor. If R is equal to or greater than +2.00, it means the number of observations in that cell is more than would be expected by chance. If R is equal to or greater than −2.00, it means the number of observations in the cell is lower than would be expected by chance. In Table 17.5, note that two cells have R values greater than |2.00|. Cell 1, Schizophrenic/Birth Complications, has a large positive R (2.57). This means that there are *more* schizophrenics *with* a history of birth complications than would be expected by chance. Cell 4, No Birth Complications/Schizophrenic, has a large negative R (−2.13). This means that there are *fewer* schizophrenics *without* a history of birth complications than would be expected by chance. Since none of the other cells yield an R that equals or exceeds |2.00|, these are the only two cells considered to have made a major contribution to the significant chi-square finding.

Table 17.5 The standardized residuals for each category of the hypothetical study on schizophrenia and birth complications (BC).[a]

Cell	f_o	f_e	$\sqrt{f_e}$	R
Cell 1	20	11.33	3.37	2.57
Cell 2	6	10.52	3.24	−1.40
Cell 3	8	12.14	3.48	−1.19
Cell 4	8	16.67	4.08	−2.13
Cell 5	20	15.48	3.93	1.15
Cell 6	22	17.86	4.23	0.98

[a] Cell 1: Schizophrenic/BC; Cell 2: depressed/BC; Cell 3: normal/BC; Cell 4: schizophrenic/No BC; Cell 5: depressed/No BC; Cell 6: normal/No BC. Cells with $R \geq |2.00|$ make a major contribution to the significant chi-square test.

The analysis of the standardized residuals is a useful technique; it allows us to make a more specific interpretation of a significant χ^2. Standardized residuals can also be used to analyze significant χ^2 goodness-of-fit tests, like the two examples presented previously in this chapter. The cell residuals are calculated the same way, and the conventional standard for any cell to be considered a major contributor is the same.

17.8 Using the Chi-Square Test with Quantitative Variables

Although the chi-square test is usually performed using discrete variables (e.g. citizen or foreigner; home owner or renter; single, married, divorced, or widowed), it is possible to use the test with a quantitative variable that is treated as a categorical variable. For example, participants may be administered a scale measuring dominance. The scores may range from 0 to 30, but the participants could be classified as either High Dominance or Low Dominance. In fact, the chi-square test for independence can be used when both variables are continuous. A researcher, for instance, may hypothesize a relationship between "need for achievement" and "annual income." In Table 17.6, participants are assigned to one of three categories of "need for achievement" and one of four categories of "annual income."

If there is some reason to suspect that the variable underlying a measure is *not* continuous, but rather a set of discrete categories, using a chi-square test is a good idea. Unlike parametric tests, a chi-square test using scaled numbers converted into categories does not assume the underlying scale to be interval or ratio.

Table 17.6 A 3 × 4 contingency table.[a]

	Annual Income ($1000)			
Need for Achievement	**<40**	**41–60**	**61–80**	**>80**
>40	39	44	57	85
21–40	45	40	53	62
<20	76	52	40	30

[a] Two quantitative variables are presented as discrete variables. Scores on a measure of need for achievement are collapsed to form low, medium, and high categories. Income is represented as four discrete categories. Cell values are observed frequency counts.

17.9 Assumptions of the Chi-Square Test

One of the advantages of the chi-square test is that there are very few assumptions that need to be met to conduct the test:

1) As with any inferential test, the sample should be representative of the population to which we want to generalize our findings.
2) The data should be in the form of a frequency count. The chi-square analysis does not analyze differences between means.
3) Each observation must be *independent* of every other observation. "Independence" not only means no influence between the participants but also that each participant is only represented once in the frequency counts.
4) Expected cell frequencies need to be of sufficient size; usually "5" is used as a rule of thumb. Chi-square analyses with small observed frequencies and/or small expected frequencies can underrepresent the Type I error rate when rejecting null hypotheses.

17.10 How to Present Formally the Conclusions for a Chi-Square Test

The basic information and formatting needed when presenting nonparametric findings in professional writings are not much different than those needed for the more typical analytical techniques of *t*'s, *F*'s, and *r*'s. However, there are a few differences. A typical sentence structure for a rejected null hypothesis for a chi-square goodness-of-fit test might take this general form: "A chi-square goodness-of-fit test was performed to determine if the types of majors were equally preferred by undergraduate students. Statistical evidence was found suggesting major preference was not equally preferred, χ^2 (2, N = 108) = 6.22, p < .05." A typical sentence structure for a rejected null hypothesis for a chi-square test of independence might take this general form: "A chi-square test of independence was calculated comparing biological sex and type of undergraduate major. Statistical evidence suggesting a relationship was found, χ^2 (2, N = 22) = 6.05, p< .05. Follow-up analyses found evidence that biological males were disproportionately more likely than biological females to major in the sciences (R = 2.75 and R = –3.25, respectively)." Notice how the sample size has been included within the parentheses that describe the basic features of the design. With other tests, the degrees of freedom are sufficient to communicate the nature of the sample size. The degrees of freedom for a chi-square, however, only reflect the number of cells in the design and not the frequency count. For this reason, a count of the sample size is typically included.

A typical sentence for a null that is not rejected might take the general form: "A chi-square test of independence was calculated comparing biological sex and type of undergraduate major. No statistical evidence of an interaction was found, χ^2 (2, N = 22) = 3.05, *n.s.*"

Summary

The chi-square test is used to analyze the frequency counts of nominal data. Tests that make assumptions and inferences about population parameters are called parametric tests. The chi-square test does not make assumptions about the shape of a population distribution and does not use means or standard deviations to infer population parameters; therefore, it is called a nonparametric test. A chi-square analysis tests the correspondence between a hypothesized distribution of frequency counts and an observed distribution of frequency counts. The null hypothesis states that there is no difference between expected and observed frequency distributions. The alternative hypothesis is a statement that the expected and observed distributions are sufficiently different such that the difference is unlikely to be due to sampling error.

The chi-square goodness-of-fit test is analogous to a one-way ANOVA with two or more groups (categories) distributed along a single factor. The chi-square test for independence is analogous to a two-way ANOVA with two or more groups (categories) distributed along two different factors. A two-way design that uses categorical data is called a contingency table. In a two-way design, the chi-square analysis tests whether two variables are independent. The null hypothesis states that there is no relationship between two variables, but the alternative hypothesis states that the variables are related; that is, they are not independent. This is analogous to the two-way ANOVA test for an interaction. Both chi-square tests generate an observed statistic that can be compared with a critical chi-square value derived from a sampling distribution reflecting that particular research design.

In situations where the null hypothesis can be rejected, a measure of effect size, Cramér's V, can be found. This value reflects the proportion of the obtained chi-square value to the size of the sample. Furthermore, a residual analysis can be used as a follow-up test to explore which cell or cells of the research design are major contributors to the large chi-square value.

The assumptions of the chi-square test include the sample to be representative of the population of interest, the data to be in the form of a frequency count, each observation to be independent of every other observation, and expected cell frequencies to be of sufficient size.

Using Microsoft® Excel and SPSS® to Calculate a Chi-Square

Excel

There is no specific Excel function for the chi-square goodness-of-fit test. However, Excel, just like a calculator, can be used to generate the necessary values to perform the test, for example, $f_o - f_e$, $(f_o - f_e)^2$, and so on.

For ease of calculation for two-factor chi-squares (chi-square test for independence), creating a Pivot Table can be of help.

General instructions for data entry into Excel can be found in Appendix C.

Data Entry

Enter the bivariate data into two adjacent columns, being sure to keep the data from each participant together in the same row. Categorical data need not be numerical. (See Figure 17.4 for an example.)

Biological sex	Major
male	arts
male	sciences
female	humanities
male	sciences
female	arts
male	sciences
female	humanities
female	humanities
female	arts
male	humanities
female	sciences
female	arts
male	sciences
female	humanities
female	humanities
male	sciences
female	arts
female	sciences
female	humanities
male	sciences
male	humanities
female	humanities

Count of biological sex	Column labels		
Row labels	**Female**	**Male**	**Grand total**
Arts	4	1	5
Humanities	7	2	9
Sciences	2	6	8
Grand total	**13**	**9**	**22**

Figure 17.4 A worked example of a chi-square test for independence analysis using Microsoft Excel.

Data Analysis

1) For Excel to run a chi-square test for independence, pivot tables must be created.
2) Click in any cell that has data and then click the **Insert** tab and select **Pivot Table**. The entire data range should become activated. Select a location for the Pivot Table and click **OK**.
3) On the right-hand side of the monitor, we will see a Pivot Table Field List displaying our two variables (Biological Sex, Major in our example). We will place one variable (Biological Sex) in the **Column Labels** box and the other variable (Major) in the **Row Labels** box (either one in either box is fine). Either variable can be placed in the $\sum$ **Values** box. All variables can be moved by dragging. The resulting values in the table are the observed frequencies that can be used to calculate the chi-square statistic. (See Figure 17.4 for an example.)
4) From this point the arithmetic features of Excel can be used to calculate f_e's as well as the χ^2_{obt}.

SPSS: Chi-Square Goodness-of-fit Test

General instructions for inputting data into SPSS can be found in Appendix C.

Data Entry

SPSS will run a goodness-of-fit chi-square, but the frequency counts for each category will need to be calculated ahead of time. Once this is done, create two variables using **Variable View**, one labeled "Category" and the other labeled "Frequency." Use a nominal scale to identify the various categories (1, 2, 3,…) and label them. Label the category values under **Values** in the **Variable View**. (See Figure 17.5 for an example.)

Go to **Data** and drop down to **Weight Cases**. Move the "Frequency" variable into the **Frequency Variable** box. This will properly weight each category based on the number of scores associated with it.

Data Analysis

1) Click **Analyze** on the tool bar, select **Nonparametric Tests**, then **Legacy Dialogues**, and then **Chi-square**.

Figure 17.5 An example of entered data for a chi-square goodness-of-fit test is SPSS.

	Category	Frequency
1	1	25
2	2	39
3	3	44

Chi-square test

Frequencies

Category

	Observed *N*	Expected *N*	Residual
Arts	25	36.0	−11.0
Sciences	39	36.0	3.0
Humanities	44	36.0	8.0
Total	108		

Test statistics

	Category
Chi-square	5.389[a]
df	2
Asymp. sig.	.068

[a]0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 36.0.

Figure 17.6 Output tables from a worked example using SPSS to run a chi-square goodness-of-fit test.

2) Move "Category" into the **Test Variable List** box.
3) Leave the default value under **Expected Values** if the null hypothesis predicts each category to be equal. (If the null hypothesis is more complex, click **Values** and put in, in order of the category number, the percent expected for each category numerical label.)
4) Click **OK**.
5) The first output box will present the observed frequency (**Observed N**), expected frequency (**Expected N**), and the difference between them (**Residual**) for each category. The second box will present the overall chi-square statistic (**Chi-Square**), the degrees of freedom (**df**), and the probability of getting a chi-square of that value if the null hypothesis is true (**Asymp. Sig.**). There is statistical evidence to reject the null hypothesis if this number is ≤ .05. (See Figure 17.6 for an example.)

SPSS: Chi-Square Test for Independence

General instructions for inputting data into SPSS can be found in Appendix C.

Data Entry

SPSS will run a chi-square test for independence. First, we need to create the two categorical variables in **Variable View**. We may want to label each numerical value used under the **Values** tab while in **Variable View**. For example,

	Biological_Sex	Major
1	1	1
2	1	2
3	2	3
4	1	2
5	2	1
6	1	2
7	2	3
8	2	3
9	2	1
10	1	3
11	2	2
12	2	1
13	1	2
14	2	3
15	2	3
16	1	2
17	2	1
18	2	2
19	2	3
20	1	2
21	1	3
22	2	3

Figure 17.7 An example of entered data for a chi-square test for independence is SPSS.

"Biological_Sex" may be labeled 1 = male and 2 = female. See Figure 17.7 for an example of properly inputted data.

Input the data properly being sure to remember that each row represents a case; typically, that is a participant.

Data Analysis

1) Click **Analyze** on the tool bar, select **Descriptives Statistics**, and then select **Crosstabs**.
2) Select one of the two categorical variables and use the arrow key to locate it in the **Row(s)** box. Select the other categorical variable, and use the arrow key to locate it in the **Column(s)** box. Then click **OK**.

Chi-square test

Crosstabs

Chi-square tests

	Value	*df*	Asymp. sig. (2-sided)
Pearson chi-square	6.051[a]	2	.049
Likelihood ratio	6.231	2	.044
Linear-by-linear Association	.120	1	.729
N of valid cases	22		

[a] 5 cells (83.3%) have expected count less than 5. The minimum expected count is 2.05.

Figure 17.8 Output tables from a worked example using SPSS to run a chi-square test for independence.

3) The second table (**Crosstabulation**) generated will present the observed frequencies for each combination of categories between the two variables.
4) Go back to **Analyze**, then **Descriptive Statistics**, and then **Crosstabs**. Then click on **Statistics**. Click **Chi-square**. (If we want to generate simultaneously an effect size measure like Cramér's *V*, we can do that here as well. For instance, we could click **Phi and Cramer's V**.)
5) Click **Continue** to leave the **Cells** menu and then click **OK** to run the analysis.
6) The third table of the output is entitled **Chi-Square Tests**. The first line labeled "Pearson Chi-Square" is the test statistic. Along with that value, we can find the *df* value (**df**) and the probability of getting a chi-square value of that size if the null is true (**Asym. Sig. (2-sided)**). Just as with previous inferential tests, we are looking for a significance value of .05 or less as statistical evidence to reject the null (assuming our alpha value is 5%). See Figure 17.8 for a visual example of the SPSS chi-square table.
7) If a residual analysis is needed, the various residual options can be found under the **Cells** tab on the **Crosstabs** menu.

Key Formulas

Formula for χ^2

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \quad \text{(Formula 17.1)}$$

Formula for computing f_e

$$f_e = \frac{f_c f_r}{N} \quad \text{(Formula 17.2)}$$

Chi-square formula for a 2 × 2 contingency table

$$\chi^2 = \frac{N(AD - BC)}{(A + B)(C + D)(A + C)(B + D)} \quad \text{(Formula 17.3)}$$

Formula for Cramér's *V*

$$\phi = \sqrt{\frac{\chi^2}{(N)\left(df_{row/column}\right)}} \quad \text{(Formula 17.4)}$$

Formula for the standardized residual

$$R = \frac{f_o - f_e}{\sqrt{f_e}} \quad \text{(Formula 17.5)}$$

Key Terms

Frequency count
Categorical data
Chi-Square test
Nonparametric test
Parametric test
Distribution-free tests
Goodness-of-fit test
Expected frequencies
Observed frequencies
Chi-square distribution
Test for independence
Contingency table
Cramér's *V*

Questions and Exercises

1 What assumptions are not needed for a chi-square analysis?

2 Why do expected frequencies and observed frequencies always equal each other?

3 What is the difference between a chi-square goodness-of-fit test and a chi-square test for independence?

4 How are the degrees of freedom for chi-square tests understood differently from degrees of freedom for most other inferential tests?

5 Why is there just one critical value for a chi-square test when the test is bidirectional? (That is, the null can be wrong in more than one way – the expected cell means can be too larger or too small.)

6 A one-way ANOVA is to a two-way ANOVA as:
- **a** A nonparametric test is to a parametric test.
- **b** A parametric test is to a nonparametric test.
- **c** A chi-square is to correlation.
- **d** A goodness-of-fit chi-square is to a chi-square test for independence.

7 Which pairing is most similar? Why?
- **a** Standardized residuals and omega-squared
- **b** Cramér's *V* and Fisher's *LSD*
- **c** Omega-squared and Cramér's *V*
- **d** Fisher's *LSD* and r^2

8 Which pairing is most similar? Why?
- **a** Standardized residuals and Fisher's *LSD*
- **b** Cramer's *V* and Tukey's *HSD*
- **c** Tukey's *HSD* and r^2
- **d** r^2 and standardized residuals

9 Three different drug treatments are used to control hypertension. At the end of treatment, the investigator classifies patients as having either a favorable or an unfavorable response to the medication. Set alpha at .05, and conduct a chi-square test regarding the null hypothesis of no relationship. If necessary, use the *R* statistic to determine which cells make a major contribution to the χ^2. Interpret the findings. If there is evidence of an effect, what is the effect size?

	Treatment			
Response	I	II	III	
Favorable	70	160	168	398
Unfavorable	30	40	32	102

10 A psychologist hypothesized that biological males are more likely than biological females to accumulate objects of trivial significance because of a biological basis to acquire and possess. Both male and female students were loaned No. 2 pencils with which to take a multiple-choice exam. A box labeled "pencils" was positioned next to a table upon which students were to place their answer sheets. The investigator counted the number of males and females who returned the pencils. The hypothesis was that males would be more likely to keep the object. Conduct a chi-square test to analyze the data. If there is evidence of an effect, what is the effect size?

	Kept pencil	Returned pencil
Males	15	40
Females	38	17

11 Specify the correct *df* for each of these designs.
a 2×2
b 3×4
c 4×5
d 1×3

12 Supply the requested information for each of the following designs. Assume $\alpha = .05$.

	Design	χ^2_{obt}	*df*	χ^2_{crit}	Reject H_0?
a	2×2	4.5			
b	3×3	9.0			
c	1×5	17.22			
d	2×4	5.55			

13 Two students find themselves in a discussion about the ways in which police decide to pull people over for traffic violations. They maintain that police are more likely to pull someone over if there is some evidence that the driver has beliefs that are offensive to the officer. They enlist the aid of 50 drivers. Twenty-five of them are asked to place the following sticker on their car bumper: Stop Police Brutality! The other 25 drivers are given a sticker that reads, Smile! Assume there is no difference in the way in which the participants of the two groups drive. Over the next 6 months, the number of times the police stop the drivers of each group is recorded. Drivers displaying the brutality sticker are stopped 18 times; drivers displaying the smile sticker are stopped 5 times. No driver is stopped more than once.
a State the null and alternative hypotheses.
b Specify f_o for each cell.
c Compute χ^2 and test the null hypothesis. Set $\alpha = .05$.
d Interpret the findings.
e If an effect is found, what is the effect size?

14 For each matrix, fill in the missing observed and marginal frequencies. Next, compute the f_e for each cell.

a

30	?	?	20	120
?	?	40	?	100
?	80	60	40	N =?

b

7	?	14
?	?	?
?	18	30

15 Here are two more. For each matrix, fill in the missing observed and marginal frequencies. Next, compute the f_e for each cell.

a

27	?	13	57
?	?	?	?
52	30	?	N = 140

b

?	?	34
?	36	?
72	?	132

16 Assume that all marginal frequencies are given for a 2 × 3 design. What are the fewest number of cells that must have frequencies specified in order to determine the rest of the cell frequencies?

17 If a research design employing a chi-square analysis has 3 rows and 4 degrees of freedom, how many columns must it have? Suppose it has 8 degrees of freedom. How many columns must it have?

18 A marketing psychologist is hired as a consultant to an association of recreational vehicle dealers. The dealers would like to know if they should seasonally alter their advertising focus. The psychologist collects data on the number of RVs sold in each season of the year. Conduct a chi-square test with $\alpha = .05$ on the following observed frequencies. Perform follow-up tests if appropriate. Interpret the findings for the dealers.

Spring	Summer	Fall	Winter
160	190	170	130

19 Frank and Lester (1988) have found that young adults, ages 15–24 years old, more often commit suicide on a Sunday. The following hypothetical data are consistent with their findings. Conduct a chi-square test on these data, and make a decision regarding the null hypothesis.

Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
56	29	17	22	25	15	33

20 Below are some data regarding a potential relationship between driving conditions and automobile accidents. Run the appropriate chi-square test to see if a relationship exists.

	Accident	No accident
Rain	29	35
No rain	31	48

21 Mothers frequently report that they had more difficulty delivering their first child in comparison with subsequent children. Kaitz, Roken, and Eidelman (1988) tested this common belief by obtaining data from primiparous (first-time mothers) and multiparous (more than one past delivery) mothers. Primiparous (n = 49) and multiparous (n = 75) mothers were asked to rate their labor as Easy, Medium, or Difficult. The following data are adapted from their study. Conduct a chi-square analysis, and test the hypothesis that there is a difference between primiparous and multiparous mothers for discomfort experienced during delivery. Set alpha at .05. In addition, if needed, use a residual analysis to identify cells that make a major contribution to a significant χ^2.

	Easy	Medium	Difficult
Primiparous Mothers	2	20	27
Multiparous Mothers	19	46	10

22 Kaitz et al. (1988) tested the hypothesis that primiparous mothers are less successful at recognizing their newborn babies in comparison with multiparous mothers. After less than 5 hr of exposure to their newborns, both primiparous and multiparous mothers were presented with seven photographs of babies, one of which was their own child. The investigators found that 30% (8/27) of the primiparous mothers and 79% (34/43) of multiparous mothers accurately identified their babies. The authors attribute this difference to a "short-lived impairment of perceptual/cognitive skills associated with their more stressful childbearing experience." Conduct a chi-square analysis on the following data. Can the null hypothesis of no relationship be rejected for this set of data?

	Correct ID	Incorrect ID	
Primiparous	8	19	27
Multiparous	34	9	43
	42	28	N = 70

23 We have learned that the *df* for a goodness-of-fit test is $C - 1$ and $(R - 1)(C - 1)$ for a two-way design. As the number of categories or cells of a design increases, χ^2_{crit} increases. Look at the formula for χ^2 and explain why it makes sense for χ^2_{crit} to increase as the number of categories increases.

24 Both the number of people getting tattoos and the reasons they get them have changed dramatically in the last few decades. Suppose a researcher employs the help of the employees of a tattoo parlor and has them record the various reasons for getting a tattoo as stated by the recipients. A content analysis organized the reasons into the following four categories:

a To memorialize a meaningful family event (e.g. marriage, birth of child) **1724**

b To memorialize a meaningful nonfamily event (e.g. first-time skydiving, visiting a foreign country) **1031**

c Conformity reasons (friends or family are doing it together) **635**

d Personal expression of an important idea (freedom, peace, etc.) **879**

Does this data set lend itself to a chi-square test? (Ignore the fact that the participants are not a random sample of all those who received a tattoo.) Run an analysis testing the null that each category is equally likely.

25 Is a younger mother more likely to give birth to a physically immature baby? In this study, younger mothers (under 20 years of age) are compared with older mothers (30–35 years of age). An immature baby is defined as having a birth weight equal to or less than 2500 g. Every baby is assigned to a category based on the age of the mother and whether the baby is below or above the weight cutoff defining physical immaturity. Set alpha at .05, and perform a chi-square analysis on the following hypothetical data. Interpret the findings. If there is evidence of an effect, measure the size.

	Birth weight		
Age	**≤ 2500 grams**	**> 2500 grams**	
Under 20	45	20	65
30–35	10	39	49
	55	59	114

26 A researcher is interested in the association between diabetes and prolonged healing of wounds. The research question is, "Do diabetics show prolonged healing?" Conduct a chi-square analysis on the following data. Set alpha at .05.

	Healing		
Patient	**Normal**	**Prolonged**	
Diabetic	125	329	454
Nondiabetic	245	111	356
	370	440	810

27 A dermatologist is interested in comparing four different treatments for dandruff. After six weeks of treatment, a colleague judges each patient as either improved or not improved. Is there any reason to conclude that the treatments have a differential effect on dandruff? Set alpha at .05 and conduct a chi-square analysis. If there is reason, what is a measure of the effect size?

Preparation	No Improvement	Satisfactory Improvement	
A	22	24	46
B	19	17	36
C	23	28	51
D	17	22	39
	81	91	172

28 Suppose we are interested in the relationship between the part of the country people live in and which of two sports they enjoy. We collect data on sports participation from people around the country and obtain the following results (see table below). Is there any evidence for a relationship between part of the country and sports enjoyment? Set alpha at .05 and conduct the appropriate chi-square analysis. If warranted, perform follow-up tests and measure the effect size.

Part of the country	Tennis	Golf
Northeast	7	9
Southeast	6	18
Southwest	20	25
Midwest	15	20

29 A political scientist was interested in seeing if there was a relationship between education level and position on gun rights. To simplify matters participants were asked to identify themselves as either "pro-" gun ownership or opposed to gun ownership ("con"). Several participants were surveyed, and the following results were found (see table below). Please run the appropriate test to see if the null hypothesis of no relationship between education level and gun rights position can be rejected. Set alpha at .05. If warranted, perform follow-up tests and measure the effect size.

Ed. level	Gun rights	Ed. level	Gun rights	Ed. level	Gun rights
H.S.	Pro	Graduate	Con	H.S.	Con
Bachelor's	Pro	H.S.	Pro	H.S.	Pro
H.S.	Con	H.S.	Con	Bachelor's	Pro
Bachelor's	Pro	Bachelor's	Pro	H.S.	Con
Bachelor's	Pro	Graduate	Con	Bachelor's	Pro

(Continued)

Ed. level	Gun rights	Ed. level	Gun rights	Ed. level	Gun rights
H.S.	Con	H.S.	Con	H.S.	Con
H.S.	Pro	Bachelor's	Pro	H.S.	Pro
H.S.	Pro	H.S.	Con	Graduate	Con
Bachelor's	Con	Graduate	Con	Graduate	Con
H.S.	Pro	H.S.	Con	H.S.	Con
H.S.	Pro	H.S.	Pro	H.S.	Pro
Graduate	Con	H.S.	Con	Bachelor's	Pro
Bachelor's	Con	Bachelor's	Pro	Bachelor's	Pro
Bachelor's	Pro	Graduate	Con	Bachelor's	Con
Graduate	Pro	Bachelor's	Con	Graduate	Pro
H.S.	Con	H.S.	Pro	Graduate	Con
Graduate	Pro	H.S.	Pro	Bachelor's	Pro
Graduate	Con	H.S.	Pro	H.S.	Pro
H.S.	Pro	Graduate	Con	Graduate	Con
Bachelor's	Pro	Graduate	Con	Bachelor's	Pro
H.S.	Con	Bachelor's	Pro	H.S.	Pro
Bachelor's	Con	Bachelor's	Pro	H.S.	Pro

18

Other Nonparametric Tests

18.1 The Research Context

Popular inferential tests, such as the *t* test or the ANOVA, are known as **parametric tests** because they test hypotheses about population parameters – usually means. In addition, these tests rest on certain assumptions: scores in the populations are normally distributed, population distributions have equal variances, and the data is measured on either an interval or a ratio scale. Although the *t* test and the ANOVA are robust tests (i.e. they can be used even when, for example, the populations are not normally distributed), gross violations of the population assumptions can invalidate parametric tests. In addition, some research questions do not lend themselves to the use of interval or ratio scales; therefore, a parametric test may not be applicable. In Chapter 17, for example, the chi-square was presented as a test performed on frequency count data.

Statisticians have developed numerous hypothesis tests that do not make assumptions about population parameters; these are called *nonparametric tests*. They can be used when assumptions about population characteristics are violated and/or when the scale of measurement used to gather the raw data is nominal or ordinal.

Many of the parametric tests previously discussed in this text have a nonparametric alternative. However, only four nonparametric tests will be covered in this chapter.[1] The *Spearman* rank correlation coefficient is used to measure the strength of association between two variables when at least one variable is measured on an ordinal scale. The *point-biserial* correlation coefficient is used to measure the strength of association between a variable measured with an interval or ratio scale (a continuous measure) and a dichotomous variable (an "either-or" variable). The *Mann–Whitney U test* is the nonparametric

1 For a detailed treatment of nonparametric tests, see Corder and Foreman (2014) or Siegel and Castellan (1988).

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

alternative to an independent-samples *t* test. It is performed using ordinal data. The *Wilcoxon signed-ranks test* is the nonparametric counterpart to the dependent-samples *t* test. It too is performed using ordinal data. As each test is discussed, appropriate research examples for the test are given. In addition, the methods of calculation and procedures for testing the null hypothesis are presented.

18.2 The Use of Ranked Data in Research

There are two different reasons why a researcher may end up using ordinal data and running a nonparametric test. In one situation, the researcher starts with collecting ordinal data. For example, an investigator might ask if there is a relationship between popularity and intelligence of children. The data are collected by asking a teacher to *rank* students from most to least popular *and* from most to least intelligent. Here is another example, suppose a researcher wants to know if there is a relationship between tennis players' national rankings and their heights. All of the players would be ranked according to height, and then the two ranked variables would be correlated. In many situations, the use of an ordinal scale has advantages. Recall from Chapter 2 that one of the assumptions of an interval scale is that numerically equal distances on the scale represent equal distances on the dimension underlying the scale. Imagine how we would go about rating the talent of several football teams. If we used an interval scale, it would be difficult to convince someone that the rating distance between any two adjacent teams is the same. For example, the difference between the best team and the second best team may be closer than the difference between the tenth- and eleventh-place teams. Using an ordinal scale circumvents this problem; rankings only make claims about the *relative position* of each event compared with the others.

There is a second reason why a researcher may use ordinal data. Collected data may be from an interval or ratio scale, but one or more of the population assumptions needed to run a parametric test may not be met. In these situations, the scores can be converted into ranks, and a nonparametric test can be used instead. To accomplish the conversion into ranks, we simply organize the original scores in ascending or descending order and assigns ranks accordingly. The original scores are then discarded, and an appropriate nonparametric analysis is performed on the newly created ranked data.

Incidentally, ranking interval- or ratio-scaled data can help a researcher address the problem of an outlier score. When numbers are converted into ranks, the value assigned to an outlier score is just one unit higher or lower, as the case may be, to the rest of the values in the data set. This technique brings the outlying data point into close proximity with the rest of the data set.

18.3 The Spearman Rank Correlation Coefficient

Chapter 15 explained that the Pearson formula, a statistic that measures the degree of association between two variables, could be used to compute a correlation. However, the data gathered in some research situations may not meet the assumptions for the Pearson correlation, or the data may reflect a certain type of nonlinearity described as *monotonic* (in these nonlinear relationships, the nonlinearity does not reverse directions as it does in shapes such as ∪ or ∩). In these situations, a Spearman rank correlation coefficient may be an appropriate option.

To perform a Spearman correlational analysis, however, the data will first need to be converted into ranks. Suppose a researcher hypothesizes a relationship between "Need for Approval" and "Ingratiating behaviors." In Table 18.1, the *Score* columns are the continuous measures for the two variables. The *Rank* columns show each score's rank in the distribution. The analysis is unaffected by whether a rank of 1 is assigned to the highest or lowest score. In this example, a rank of 1 is assigned to the highest score, a rank of 2 to the next highest score, and so on. Participant 1 scored an 8 on "Need for Approval," which was the highest score, and so a rank of 1 has been assigned to that value. Participant 7 scored a 6 on the measure of "Ingratiation," and since a 6 is seven scores from the top of the distribution, a rank of 7 has been assigned.

Figure 18.1a is the scatter plot for the continuous measures of "Need for Approval" and "Ingratiation." Since the correlation is extremely high, the points of the scatter plot line up fairly well. However, observe how the line is curved (nonlinear). In this situation, a conversion of the raw data into ranked data may create greater linearity. After converting to ranks, the line becomes straighter,

Table 18.1 Converting continuous measures to ranks.

	Need for Approval		Ingratiation	
Participant	**Score**	**Rank**	**Score**	**Rank**
P_1	8	1	11	1.5
P_2	7	2.5	11	1.5
P_3	7	2.5	10	3
P_4	6	4	9	4
P_5	5	5	8	5
P_6	4	6	7	6
P_7	3	7	6	7
P_8	2	8	2	8

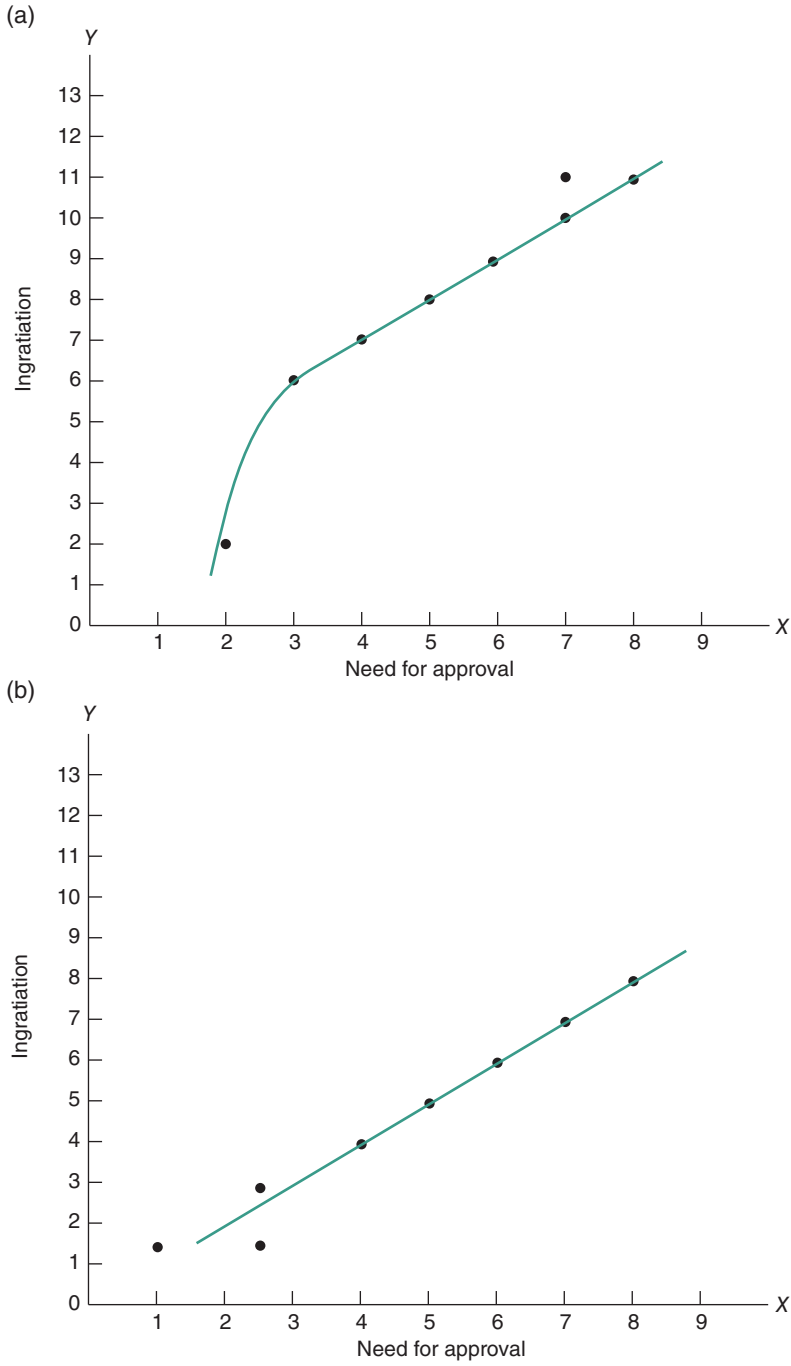


Figure 18.1 **Observe** how the nonlinear scatter plot in (a) becomes linear when the continuous measures have been changed to ranks in (b).

showing a more linear relationship between X and Y (Figure 18.1b). If the relationship between X and Y displays a form of curvilinearity that reverses direction (*nonmonotonic*), for instance, looking like ∪ or ∩, conversion into ranks will not work. However, mildly curved plots often straighten out when ranked. If ranking the scores is successful in making the data linear, we have a choice of correlation formulas. We can use the Pearson formula presented in Chapter 15 (but only if all other population assumptions are met), or we can use a nonparametric formula: the **Spearman rank correlation**, symbolized as r_s. (Charles Spearman actually symbolized the coefficient as ρ_s. However, since ρ has come to symbolize a population correlation, modern usage favors r_s for the Spearman rank correlation of a sample.) The Spearman formula is a computationally simplified Pearson formula applied to rankings. As a result, the essential features of a correlation coefficient still apply:

1) The correlation can assume any value between −1 and +1.
2) The sign of the correlation reflects the nature of the relationship.
3) r_s^2 reflects the amount of shared variance between X and Y.
4) When testing the statistical significance of r_s, the null hypothesis is *usually* $\rho = 0$.

Formula for Spearman rank correlation, r_s

$$r_s = 1 - \frac{6\Sigma D^2}{n_p\left(n_p^2 - 1\right)} \quad \text{(Formula 18.1)}$$

where

D^2 = the squared difference between a pair of ranks
n_p = the number of *paired* scores

■ **Question** *Applying Formula 18.1 to the ranked data in Table 18.1, what is the correlation between Need for Approval and Ingratiation?*

Solution

Participant	Need for Approval Rank	Ingratiation Rank	D	D^2
P_1	1	1.5	−0.5	0.25
P_2	2.5	1.5	1	1
P_3	2.5	3	−0.5	0.25
P_4	4	4	0	0

(*Continued*)

Participant	Need for Approval Rank	Ingratiation Rank	D	D^2
P_5	5	5	0	0
P_6	6	6	0	0
P_7	7	7	0	0
P_8	8	8	0	0
			$\Sigma D = 0$	$\Sigma D^2 = 1.5$

$$\Sigma D^2 = 1.5$$

$$n_p = 8$$

$$r_s = 1 - \frac{6D^2}{n_p\left(n_p^2 - 1\right)}$$

$$r_s = 1 - \frac{6(1.5)}{8(8^2 - 1)}$$

$$r_s = 1 - \frac{9}{8(63)}$$

$$r_s = 1 - 0.018$$

$$r_s = +\mathbf{0.98} \quad \blacksquare$$

Tied Ranks

When ***converting*** continuous measures into ranks, we will frequently encounter two or more participants that have identical X or Y scores. When the ranks of two scores are tied, take the average of the two contiguous ranks. For example, in Table 18.1, Participants 2 and 3 both scored a 7 on Need for Approval, which happens to be the second highest score, but we cannot simply assign a rank of 2 to both participants. The customary method takes the average of the ranks 2 and 3 and assigns a 2.5 to each one. The next highest score in the distribution will be assigned a rank of 4 (*not* 3). We should try to avoid tied ranks because ties have the effect of inflating the correlation. However, if there are not too many ties and not too many long ties (three-or-more-way ties), then the overestimation will be acceptably small (approximately 0.02) (Welkowitz, Ewen, & Cohen, 1988). There may be no way to avoid tied ranks when we are converting continuous measures; the scores determine the ranks and we have no control over the scores. Of course, in the studies in which the investigator collects the data as ranked data, steps can be taken to avoid ties. The following section presents an example in which data are ranked from the beginning of data collection.

The Planned Use of Ranks

Imagine the following hypothetical theory and study. A social psychologist believes that, in the course of a year, children in a classroom will form a dominance hierarchy. Furthermore, the psychologist believes that the hierarchy is formed based on popularity, with the most popular child ascending to the top of the dominance hierarchy and the least popular child stuck at the bottom.

As a measure of popularity, the classroom teacher is asked to rank all of the children from the most to least popular. To measure dominance, the children are given the opportunity to play a new video game, but they must come to an agreement about which of them will go first, which will go second, and so on. The order in which the children play the game is used as the measure of "Dominance." Table 18.2 presents the rankings for every child on both variables. A rank of 1 is assigned to the most dominant child, and a rank of 10 is assigned

Table 18.2 Using the Spearman rank formula to compute the correlation between dominance and popularity.

Child	Dominance	Popularity	D	D^2
Erin	1	1	0	0
Megan	7	8	−1	1
Caleb	6	9	−3	9
Christopher	8	5	3	9
Karis	3	2	1	1
Justine	4	3	1	1
Austin	5	4	1	1
Ella	10	6	4	16
Jake	9	7	2	4
Oren	2	10	−8	64
			$\Sigma D = 0$	$\Sigma D^2 = 106$

$$r_s = 1 - \frac{6D^2}{n_p\left(n_p^2 - 1\right)}$$

$$r_s = 1 - \frac{6(106)}{10(100-1)}$$

$$r_s = 1 - \frac{636}{990}$$

$$r_s = 1 - 0.642$$

$$r_s = +\mathbf{.36}$$

to the least dominant child. In like manner, a rank of 1 is assigned to the most popular child, and a rank of 10 is assigned to the least popular child.

Follow each step of the calculations of r_s in Table 18.2. The obtained r_s of +.36 is consistent with the hypothesis that the more popular the child, the more likely they will assume a dominant position in the class. Whether this relationship between popularity and dominance would be found using other measures of these variables would need to be tested in future research. However, before making any conclusions about this correlation, a test of the null hypothesis will need to be conducted. This is covered later in this section.

Another Example Using Planned Ranks

Another occasion in which participants are ranked from the beginning of the study occurs when two judges provide rankings on *one* variable. For instance, two psychiatrists might rank hospitalized patients along the dimension of how disturbed the patients appear to be, with a rank of 1 given to the most disturbed person, a rank of 2 assigned to the next most disturbed person, and so on. Two gym teachers could rank students on athletic ability, with a 1 assigned to the student who is viewed as the best athlete, a 2 given to the next best athlete, and so on.

Do not be confused about interpreting the r_s when two judges provide rankings on one variable. The correlation reflects the *strength of association of the rankings of the two judges*. Alternatively stated, the correlation indicates the degree to which the judges agree as to how the participants should be ranked on the variable of interest. If the correlation is high, we can be confident that the judges are consistent in ranking the participants. If the correlation is low, we can infer that the judges are using different criteria when making their rankings; or perhaps, they are using the same criteria, but they do not have access to the same information (e.g. maybe they have observed the participants in different settings). Measuring the correlation between the judgements of different evaluators, a concept referred to as *inter-rater reliability*, is an important component to many social and behavioral science research studies.

Following is an example of a *misinterpretation* of r_s for the rankings of two judges. Suppose two school psychologists rank a group of children from the most to the least friendly. The r_s turns out to be +.45. We should *not* conclude that there is a correlation between children's friendliness and the judges' rankings. These are *not* the two things being measured for association. The correct interpretation is that the judges tend to agree as to how the children should be ranked with respect to friendliness. However, the fact that the correlation is only +.45 suggests they show only a moderate degree of agreement. We should expect r_s to be rather high when two judges rank participants on one variable. As the correlation drops below +.80, our concern about the judging process should increase.

Table 18.3 Using the Spearman rank formula to correlate rankings of two judges.

Bodybuilder	Judge 1	Judge 2	D	D^2
Dickenson	1	1	0	0
Rexford	2	2	0	0
Bricken	10	9	1	1
Bundy	6	7	–1	1
Bower	8	8	0	0
Strobel	4	4	0	0
Couvion	7	6	1	1
Shelton	3	3	0	0
Gray	9	10	–1	1
Hamilton	5	5	0	0
			$\Sigma D = 0$	$\Sigma D^2 = 4$

$$r_s = 1 - \frac{6D^2}{n_p\left(n_p^2 - 1\right)}$$

$$r_s = 1 - \frac{6(4)}{10(100-1)}$$

$$r_s = 1 - \frac{24}{990}$$

$$r_s = 1 - 0.024$$

$$r_s = +\mathbf{.98}$$

Table 18.3 provides a hypothetical example of two judges ranking bodybuilders competing for the title of Mr. All-Too-Wonderful. The obtained r_s of +.98 indicates that the judges are in strong agreement as to how to make their judgments.

Some Problems with Using Ranks

Converting continuous measures into ranks (i.e. changing an interval or ratio scale to an ordinal scale) is a procedure that can be resorted by a researcher because the scatter plot reveals an unacceptable degree of nonlinearity. However, sometimes an investigator will choose to use rankings from the beginning of a study. This practice is often due to the absence of a suitable continuous measure. There is hesitancy among social and behavioral scientists to using ranks; ranks are less sensitive measures than interval or ratio scales. The reason for the insensitivity is the lack of uniformity between the ranks. For instance, the

distance between the ranks of, say, 2 and 3 may be very different from the distance between the ranks of 7 and 8. In addition, a high rank may not necessarily correspond to a large amount of the variable that is being ranked. For example, we might rank five comedians on how funny we find them. In our estimation, the comedian receiving the highest rank is funnier than the other four, but we may find none of them to be very funny. Correspondingly, a low rank may not mean there is only a small amount of the quality being measured. Using the same example, we may find the lowest-ranked comedian to be very funny, just not as funny as the other four.

Another disadvantage to using ranked data is the drop in power that occurs with the diminished quantitative sensitivity. In other words, an inferential test using a ranked version of a data set is less likely to reject the null hypothesis compared with a data set using an interval or ratio scale. Whenever there is an opportunity to use an interval or ratio scale that also meets all of the population assumptions of an inferential test, it should be taken. The statistical power will be greater.

Using Spearman's *r* to Test the Null Hypothesis

When using r_s to test the null hypothesis, use Table A.9 in the Appendix. This table specifies the critical values for the Spearman's r_s. Note that the critical value (r_{crit}) is found by entering the left column using the *number of pairs* of scores (not $n_p - 2$). We can then conduct a directional or nondirectional test of the null hypothesis by using the appropriate column. The null hypothesis states that the population correlation ρ is 0. A rejection of the null hypothesis is warranted when the observed r_s falls outside of $\pm r_{crit}$ found in Table A.9.

To demonstrate how to use the Spearman to test a null hypothesis, let us look at the dominance and popularity data that is found in Table 18.2. The r_s was found to be +.36. Setting alpha at .05, the r_{crit} for a nondirectional test, with $n_p = 10$, equals ±.648. The value of .36 does *not* fall outside of ±.648. Therefore, we should *not* reject the null hypothesis that $\rho = 0$. This means that we do not have evidence that popularity and dominance are related at the population level. Keep in mind that inferential hypothesis testing takes place only when we are interested in determining the characteristics of populations using sample data. To run an inferential test between the judges' rankings of bodybuilders would have little meaning. This data, after all, is not a sample drawn from a population.

18.4 The Point-Biserial Correlation Coefficient

The **point-biserial correlation** analysis is used when one variable is continuous and the second variable is dichotomous. Some examples of dichotomous variables are student/nonstudent, married/single, theist/nontheist, and resident/alien.

To use the point-biserial formula, the **dichotomous variable** should be *genuinely* dichotomous and not merely *artificially* dichotomous. An example of an artificial dichotomous variable would be to take the heights of research participants and split them into two groups, tall people and short people. The point-biserial correlation is not a good option for a variable that has an underlying continuity.

To compute the point-biserial correlation coefficient, each participant is assigned either a 0 or a 1 for the dichotomous variable. Assigning numbers to dichotomous groups is called **dummy coding**. For example, if one variable is biological sex, all males might be assigned a 0 and all females assigned a 1. We could assign 3's and 4's if we would like, but researchers typically use either 0's and 1's or 1's and 2's. The term *biserial* reflects the fact that there are two series of persons being observed on variable *Y*: those who are assigned a 0 on *X* and those assigned a 1 on *X*.

Suppose a researcher wants to examine the relationship between biological sex and assertiveness, with assertiveness assessed using an interval scale. For the *X* variable of biological sex, one series of participants (males) would receive a 0, and the other series of participants (females) would receive a 1. The *Y* variable is the continuous measure, "assertiveness." The point-biserial correlation would measure the strength of association between these two variables.

Formula for point-biserial correlation, r_{pb}

$$r_{pb} = \frac{M_{Y_1} - M_{Y_0}}{s_y}\sqrt{\frac{n_1 n_0}{n(n-1)}} \quad \text{(Formula 18.2)}$$

where

M_{y_1} = the mean of the continuous measure for just those participants assigned an *X* value of 1
M_{y_0} = the mean of the continuous measure for just those participants assigned an *X* value of 0
s_y = the standard deviation of *all* the scores on the continuous measure, i.e. irrespective of group designation
n_1 = the number of participants assigned a 1 for the *X* variable
n_0 = the number of participants assigned a 0 for the *X* variable
$n = n_1 + n_0$; total number of participants

■ **Question** *A clinical psychologist is interested in the relationship between biological sex and the fear of making a long-term commitment to a member of the opposite* biological *sex. A continuous measure of "fear of commitment" is given to biological males and females. Using the data presented in Table 18.4, compute the point-biserial correlation coefficient.*

Solution In Table 18.4, the numbers in the Biological Male and Biological Female columns are scores on the *Y*, continuous measure (*fear of commitment*). Note how the *X*, dichotomous variable (*biological sex*), has been dummy coded as biological male = 0 and biological female = 1. ■

Interpreting the Point-Biserial Correlation Coefficient

Students sometimes have difficulty interpreting the meaning of a point-biserial correlation coefficient, especially the direction of the relationship. As we examine the biological male and biological female columns in Table 18.4, remember that the scores across from one another are *not pairs* of scores. If the data were presented as pairs of scores, there would be two columns, one column with 0's

Table 18.4 Calculating the point-biserial correlation coefficient between sex and fear of commitment.

Biological Male: (0)		Biological Female: (1)	
P_1	22	P_8	13
P_2	14	P_9	16
P_3	20	P_{10}	11
P_4	8	P_{11}	12
P_5	11	P_{12}	4
P_6	9	P_{13}	3
P_7	9	P_{14}	6

$$r_{pb} = \frac{M_{Y_1} - M_{Y_0}}{s_y}\sqrt{\frac{n_1 n_2}{n(n-1)}}$$

$$M_{y_0} = 13.29$$

$$M_{y_1} = 9.29$$

$$s_y = 5.51$$

$$n_0 = 7$$

$$n_1 = 7$$

$$n = 14$$

$$r_{pb} = \frac{9.29 - 13.29}{5.51}\sqrt{\frac{(7)(7)}{14(14-1)}}$$

$$r_{pb} = \frac{-4.00}{5.51}\sqrt{\frac{49}{182}}$$

$$r_{pb} = -0.73\sqrt{0.269}$$

$$r_{pb} = -0.73(0.52)$$

$$r_{pb} = \mathbf{-.38}$$

and 1's, and one column with each participant's score on the continuous measure. Organized as pairs of scores, the data in Table 18.4 would look like this:

P	X	Y
P_1	0	22
P_2	0	14
P_3	0	20
⋮	⋮	⋮
P_8	1	13
P_9	1	16
P_{10}	1	11
⋮	⋮	⋮
P_{14}	1	6

We could, in fact, compute the correlation between a dichotomous and continuous measure using the Pearson raw score formula presented in Chapter 15; the resulting value would be the same. Indeed, just as the Spearman rank formula is a simplified version of the Pearson formula applied to ranks, the point-biserial formula is a version of the Pearson formula applied when one variable is dichotomous.[2]

When interpreting the direction of the correlation, pay attention to which group members received the lower of the two dummy codes and which received the higher of the two codes. Recall that a positive correlation means that lower numbers on the X variable are associated with lower numbers of the Y variable, and, of course, higher numbers of the X variable are associated with higher numbers of the Y variable. The reverse is true for a negative correlation: lower numbers on one variable are associated with higher numbers on the second variable. For the worked problem in Table 18.4, a negative correlation was obtained; biological males were assigned the lower number (0) and biological females the higher number (1). Therefore, the negative correlation means that higher "fear of commitment" scores are associated with biological males. Had the dummy codes been reversed and biological females assigned a 0 and biological males a 1, the correlation would have been +.38 instead of –.38. However, the interpretation of the correlation would have remained the same. When reporting the results of a point-biserial correlational analysis, be sure to include a specific interpretation of the finding. Simply stating that there is a +.38 or –.38

2 As an exercise, use the Pearson raw score formula with the data in Table 18.4.

correlation between biological sex and fear of commitment would confuse readers since they would not necessarily know how the dummy codes were assigned.

Using the Point-Biserial Correlation Coefficient to Test the Null Hypothesis

The point-biserial correlation coefficient can be used for testing the null hypothesis by using the same table of critical values as the Pearson *r* (Table A.7 in the Appendix). The degrees of freedom is $n - 2$, where n is the total number of participants. If r_{pb} falls outside of $\pm r_{crit}$, then the null hypothesis that $\rho = 0$ can be rejected. Test the correlation found between *biological sex* and *fear of commitment* using a nondirectional test with an alpha level of .05. Entering the appropriate column in Table A.7, note that the critical value for 12 *df* (14 – 2) is .532. Since –.38 does not fall outside of ±.532, do *not* reject the null hypothesis. In other words, we do not have statistical evidence that there is a relationship between *biological sex* and *fear of commitment*.

■ **Question** *A psychologist hypothesizes an association between Marital Status and Need for Achievement. A questionnaire measuring "Need for Achievement" is* administered *to married and single people. Higher scores indicate a greater need. As we examine the following data set, notice that there are more single than married people in the sample. This is perfectly fine (within reason); the point-biserial formula can work with unequal numbers of participants in each group. Married individuals are assigned a 0 and single individuals are assigned a 1. Test the null hypothesis that* $\rho = 0$, *set* $\alpha = .05$, *and interpret the correlation.*

Marital Status	Need for Achievement
0	3
0	7
1	12
1	16
1	24
0	11
1	15
0	10
0	11
1	18
1	22
0	9
1	19
1	17

Solution

$$r_{pb} = \frac{M_{Y_1} - M_{Y_0}}{s_y}\sqrt{\frac{n_1 n_2}{n(n-1)}}$$

$$M_{y_0} = 8.5$$

$$M_{y_1} = 17.9$$

$$s_y = 5.89$$

$$n_0 = 6$$

$$n_1 = 8$$

$$n = 14$$

$$r_{pb} = \frac{17.9 - 8.5}{5.89}\sqrt{\frac{(8)(6)}{14(14-1)}}$$

$$r_{pb} = \frac{9.4}{5.89}\sqrt{\frac{48}{182}}$$

$$r_{pb} = 1.60\sqrt{0.264}$$

$$r_{pb} = 1.60(0.51)$$

$$r_{pb} = +.\mathbf{82}$$

To test the null hypothesis, turn to Table A.7 in the Appendix, and find the critical value for an alpha of .05, with 12 *df*. The critical value is .532. The obtained correlation of .82 falls outside of ±.532; therefore, we can reject the null hypothesis that states $\rho = 0$. A positive correlation means that we have found evidence that singles (who were dummy coded with a 1) have a greater need for achievement compared with married individuals. ■

18.5 The Mann–Whitney *U* Test

The **Mann–Whitney *U* test** is the nonparametric alternative to the independent-samples *t* test. Any research design that has two independent groups of participants is a candidate for the use of a Mann–Whitney *U* test. The decision to use a *t* test or the Mann–Whitney *U* test is based on whether we believe we have met the statistical assumptions for a *t* test. If there is reason to suspect that the population distributions depart radically from normality, the variances of the populations are unequal, or the data is not measured on an interval or ratio scale, the Mann–Whitney *U* test is preferred to the *t* test. Therefore, the most common pathway to the Mann–Whitney *U* test is to collect data, discover that the statistical assumptions for a *t* test have been violated, *convert the raw scores into ranks*, and perform the Mann–Whitney *U* test on the ranked data.

The Rationale Underlying the Mann–Whitney *U* Test

The experimental situation appropriate for the use of the Mann–Whitney *U* test is one in which there are two independent groups of participants, with each participant providing a score. Once the decision is made to use the Mann–Whitney *U*, the scores from *both* groups are combined, forming one large group. The entire set of scores is listed from lowest to highest. Each score is then assigned its corresponding rank: typically, the lowest is ranked 1, the next lowest is ranked 2, and so on. However, just as with the Spearman, the ranking system can be reversed. It will not change the outcome of the test. Next, the participants' *ranks* are placed back into the original two groups of the design.

Now, suppose there is a treatment effect. How do we think the groups will be distributed across the span of the ranks? If there is an effect, we should find that one group will have many more ranks at the lower end of the scale in comparison with the other group. Suppose there is no treatment effect. In this situation, the group members should look like they have been evenly dispersed across the span of the ranks. Figure 18.2a illustrates a distribution of ranks

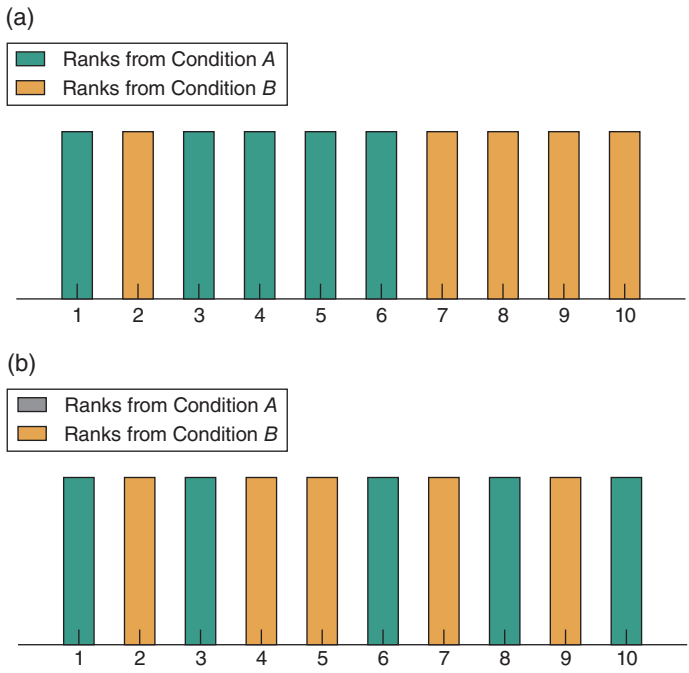


Figure 18.2 (a) More ranks of Condition *A* are shown at the lower end of the scale, and more ranks from Condition *B* are at the higher end of the scale. This indicates the presence of a treatment effect. (b) The ranks of Conditions *A* and *B* fail to show a systematic grouping in either end of the scale. When there is no treatment effect, the ranks of Conditions *A* and *B* are intermixed and evenly dispersed across the span of the ranks.

when there is a treatment effect. Note that Condition *A* has greater representation in the lower ranks than Condition *B*. The group members are not evenly dispersed across the span of the ranks. Figure 18.2b illustrates a case in which there is no treatment effect. The ranks from one group do not systematically fall into either end of the scale. The group members appear to be evenly dispersed across the span of the ranks. The Mann–Whitney *U* test helps us determine the likelihood that a particular arrangement of ranks can be explained by chance.

Calculating the Mann–Whitney *U* Without a Formula

Worked Example

As a means for demonstrating how to calculate the Mann–Whitney *U*, let us evaluate a hypothetical program for increasing vocabulary. Ten participants are randomly assigned to the experimental and control conditions (five participants in each group). After two days of training, all participants are tested on the number of words they can define.

Step 1. The vocabulary scores for each participant are listed according to experimental condition.

Condition *A* (treatment)	75, 4, 32, 140, 20
Condition *B* (control)	33, 49, 90, 100, 9

Step 2. Arrange all the scores from lowest to highest and rank them. Although the scores are combined into one list, we need to keep track of which scores come from which condition.

Score	4	9	20	32	33	49	75	90	100	140
Rank	1	2	3	4	5	6	7	8	9	10
Condition	*A*	*B*	*A*	*A*	*B*	*B*	*A*	*B*	*B*	*A*

Step 3. Now work only with the ranks and their respective group assignments. For each participant (rank) from Condition *A*, count the number of participants (ranks) from Condition *B* that are *above* that rank. The number of ranks above a given rank will be referred to as points (Gravetter & Wallnau, 2017).

Rank	1	2	3	4	5	6	7	8	9	10
Condition	*A*	*B*	*A*	*A*	*B*	*B*	*A*	*B*	*B*	*A*
Points for Condition *A*	5		4	4			2			0

Step 4. Add all the points for Condition *A*: 5 + 4 + 4 + 2 + 0 = 15. Fifteen is the *U* value for Condition *A*, symbolized as U_A.

Step 5. Steps 3 and 4 are repeated for the ranks of Condition *B*.

Rank	1	2	3	4	5	6	7	8	9	10
Condition	*A*	*B*	*A*	*A*	*B*	*B*	*A*	*B*	*B*	*A*
Points for Condition *B*		4			2	2		1	1	

$$U_B = 4 + 2 + 2 + 1 + 1 = 10$$

Step 6. Determine the Mann–Whitney *U*. The Mann–Whitney *U* value is the *smaller* of U_A and U_B, in this case, 10. As a computational check, the number of participants in Condition *A*, n_A, multiplied by the number of participants in Condition *B*, n_B, should equal $U_A + U_B$. Therefore, $U_A + U_B = 15 + 10 = n_A n_B = 5(5) = 25$.

Calculating the Mann–Whitney *U* with Formulas

The Mann–Whitney *U value* can be found another way, using formulas to determine U_A and U_B. With this method, all participants are rank ordered in the manner described in the preceding step list. For the participants in Condition *A*, ΣR_A is computed (the sum of the *ranks*, not the sum of points). Then, the sum of the ranks is computed for the participants in Condition *B*, ΣR_B. The following are formulas for U_A and U_B. Remember that n_A and n_B refer to the number of participants in Conditions *A* and *B*, respectively.

Formula for computing U_A

$$U_A = n_A n_B + \frac{n_A(n_A + 1)}{2} - \Sigma R_A \quad \text{(Formula 18.3)}$$

Formula for computing U_B

$$U_B = n_A n_B + \frac{n_B(n_B + 1)}{2} - \Sigma R_B \quad \text{(Formula 18.4)}$$

Let us use the last worked problem to illustrate how these formulas are applied.

Rank	1	2	3	4	5	6	7	8	9	10
Condition	*A*	*B*	*A*	*A*	*B*	*B*	*A*	*B*	*B*	*A*

$$\Sigma R_A = 1 + 3 + 4 + 7 + 10 = 25$$

$$n_A = 5$$

$$n_B = 5$$

$$U_A = n_A n_B + \frac{n_A(n_A+1)}{2} - \Sigma R_A$$

$$U_A = 5(5) + \frac{5(5+1)}{2} - 2$$

$$U_A = 25 + \frac{5(6)}{2} - 25$$

$$U_A = 25 + 15 - 25$$

$$U_A = \mathbf{15}$$

Note that U_A is 15, the same value computed using the points method. Now use Formula 18.4 to determine U_B:

$$\Sigma R_B = 2 + 5 + 6 + 8 + 9 = 30$$

$$U_B = n_A n_B + \frac{n_B(n_B+1)}{2} - \Sigma R_B$$

$$U_B = 5(5) + \frac{5(5+1)}{2} - 30$$

$$U_B = 25 + \frac{5(6)}{2} - 30$$

$$U_B = 25 + 15 - 30$$

$$U_B = \mathbf{10}$$

Again, note that the value of U_B (10) is the same whether the point or formula method is used. Recall that the Mann–Whitney U is the smaller of U_A and U_B: $U = 10$.

Hypothesis Testing and the Mann–Whitney *U*

The Null and Alternative Hypotheses

If the null hypothesis is true, the two samples are taken from a single population. Under this condition, the distribution of ranks for Conditions *A* and *B* should not show a systematic difference. As noted previously, and illustrated in Figure 18.2, the ranks of both conditions will be *highly* intermixed when the null hypothesis is true. However, hypothesis testing is probabilistic. Therefore, it is possible for the distribution of ranks to show a systematic ordering, even when the null hypothesis is true. As with other inferential tests, sampling error can lead us to commit a type I error by erroneously rejecting a true null hypothesis. The null hypothesis is typically a statement about the equivalence of population distributions:

H_0 : The population distribution of *A* = the population distribution of *B*.

The alternative hypothesis is a statement about the nonequivalence of the population distributions:

H_1 : The population distribution of $A \neq$ to the population distribution of B.

Finding the Critical Value for *U*

The sampling distribution of U is based on the number of participants in each group, n_A and n_B. Tables A.10 and A.11 in the Appendix contains the critical values for various alpha levels and all combinations of n_A and n_B, provided the largest sample size of either group does not exceed 20 (more on this point later). Critical values are provided in lightface and boldface for directional and nondirectional tests, respectively. A dash mark in the table indicates that no decision is possible at the stated level of significance given those values of n_A and n_B. For two-tailed tests, Table A.10 is used when alpha is set at .02 or .01. Table A.11 is consulted when conducting a two-tailed test and alpha is set at .10 or .05.

■ **Question** *What is the critical value for a two-tailed test when $\alpha = .05$, $n_A = 9$, and $n_B = 12$?*

Solution 26. ■

Comparing *U* with U_{crit}

In all previously discussed significance tests, the null hypothesis is rejected when the obtained statistic *is equal to or great than* (in terms of absolute value) the critical score. The Mann–Whitney U test is different. To reject the null hypothesis, U must be *equal to or smaller than* the critical value. To help explain why this is the case, consider the following example in which the distributions of Conditions A and B depart maximally.

Rank	1	2	3	4	5	6	7	8	9	10	11
Condition	*A*	*A*	*A*	*A*	*A*	*B*	*B*	*B*	*B*	*B*	*B*
Points	6	6	6	6	6	0	0	0	0	0	0

Recall that U is the smaller value of U_A and U_B. Since U_B is 0, $U = 0$. This shows that the strongest possible evidence for rejecting the null hypothesis occurs when $U = 0$. The smaller the value of U, the more likely it is to support the rejection of the null hypothesis.

Let us test the U we calculated from the vocabulary study. The critical value for $n_A = 5$, $n_B = 5$, $\alpha = .05$, and two-tailed test is 2. The U value is 10. Since 10 is not smaller than 2, the null hypothesis is *not* rejected. We should interpret this as failure to find statistical evidence suggesting the training program influenced the strength of a participant's vocabulary.

Hypothesis Testing with a Large Sample Size

By inspecting Tables A.10 and A.11, we can see that the Mann–Whitney U table does not provide critical values when either n_A or n_B is greater than 20. When either sample size exceeds 20, the sampling distribution of U approximates a normal distribution. In this instance, the standard normal curve can be used to identify critical values. These values have become familiar to us: ±1.96 and ±2.58 for a nondirectional test when alpha is .05 or .01, respectively. When using a large sample size, the U value is transformed to a z value, z_U, which is compared with the desired critical value of z. The old rule of comparison now applies. If z_U falls outside of $\pm z_{crit}$, the null hypothesis can be rejected. Formula 18.5 is used to transform U to z_U.

The U to z_U transformation formula

$$z_U = \frac{U - (n_A n_B/2)}{\sqrt{\frac{n_A n_B (n_A + n_B + 1)}{12}}} \quad \text{(Formula 18.5)}$$

Formula 18.5 would not be applied to the data from the vocabulary study because neither group has more than 20 participants, but for the sake of illustration, we will use this data to illustrate the workings of the transformation formula:

$$z_U = \frac{10 - 5(5)/2}{\sqrt{\frac{5(5)(5+5+1)}{12}}}$$

$$z_U = \frac{-2.5}{\sqrt{275/12}}$$

$$z_U = \mathbf{-0.52}$$

The z_{crit} value for $\alpha = .05$, two-tailed test, is ±1.96. The z_U of –0.52 does not fall outside of ±1.96. Therefore, the null hypothesis is not rejected.

Ranking Tied Scores

When using the Mann–Whitney U test, resolving tied scores is handled in the same manner as the Spearman rank correlation analysis. Briefly, tied scores are resolved by taking the average of the ranks to which the scores need to be assigned; in this way, each tied score is assigned the same averaged rank.

Several tied ranks *within* a condition do not present a problem for the Mann–Whitney U test. However, when a rank from Condition A is tied with a rank from Condition B, and the number of ties is large, the Mann–Whitney U test becomes excessively conservative. A correction factor can be applied in these

instances, but the procedure is rather complex. More information can be found by consulting nonparametric statistics texts (e.g. Siegel & Castellan, 1988).

18.6 The Wilcoxon Signed-Ranks Test

The **Wilcoxon signed-ranks test** is the nonparametric alternative to the dependent-samples *t* test. The research context is a repeated-measures design in which one group of participants receives two treatments. The treatments can be two experimental conditions, an experimental and control condition, or a pretest and posttest. The Wilcoxon signed-ranks test assumes that the dependent variable is a continuous measure, even though the analysis is performed on ranks. As we might expect, no population assumptions are needed, and the null hypothesis states that there is no treatment effect.

Calculating the Wilcoxon *T*

The Wilcoxon signed-ranks test is performed on the rankings of *difference scores*. In a repeated-measures design, a difference score is a single participant's score in Condition *B*, subtracted from their score in Condition *A*. The difference scores are ranked, from smallest to largest, based on the *absolute value* of the scores. In this way, a score of –72 is ranked higher (i.e. given a larger rank value) than a score of 2. Next, the ranks of the positive-difference scores are placed in one group, and the ranks of the negative-difference scores are placed in a second group. The ranks of the positive-difference scores are summed, ΣR_{pos}, and the ranks of the negative-difference scores are summed, ΣR_{neg}. Of the values of ΣR_{pos} and ΣR_{neg}, the one that is smaller is the Wilcoxon statistic, *T*. To test the null hypothesis, the *T* value is compared with a critical value found in Table A.12 in the Appendix. The steps for calculating *T* are shown in the following worked example.

Worked Example

A cognitive psychologist would like to compare two techniques for enhancing the recollection of nonsense syllables. In Condition *A*, participants are told to study a list of syllables by repeating them over and over (repetition). In Condition *B*, the same participants are told to examine a different list of nonsense syllables and to try to associate them with a common word (association). Half of the participants receive the repetition method first; the remaining half receive the association method first. The dependent variable is the number of nonsense syllables correctly recalled. Assume there is some reason to suspect that the population assumptions for a paired-observations *t* test have been violated and that the Wilcoxon signed-ranks test is the analysis of choice.

Number correct

	Condition			
Participant	*A*	*B*	Difference	Rank
P_1	32	27	+5	6
P_2	40	44	−4	−5
P_3	12	12	0	1.5
P_4	2	16	−14	−10
P_5	56	53	+3	4
P_6	16	6	+10	8
P_7	29	22	+7	7
P_8	49	20	+29	11
P_9	20	21	−1	−3
P_{10}	15	15	0	−1.5
P_{11}	13	2	+11	9

Step 1. Arrange the data in a table and compute a difference score for each participant.

Step 2. Arrange the difference scores from smallest to largest, and rank these scores based on their absolute values. Handle tied ranks in the usual manner; take the average of the ranks. Notice that ranks associated with negative-difference scores have a negative sign in front of them. This is simply to remind us which ranks are assigned to the positive group and which ranks are assigned to the negative group.

Step 3. Group all the ranks associated with a positive-difference score. Form a second group of ranks that correspond to negative-difference scores. If there is a tie between two participants with *difference scores of* 0, assign one rank to the first group and the other rank to the second group. Note that two participants received a difference score of 0, and both were assigned a rank of 1.5; one is placed in group one and the other in group two. If there is an odd number of ties, discard one of them, and divide the remaining tied ranks equally among the groups. *This method only applies to ties based on difference scores of* 0. If two positive-difference scores are the same, their ranks are averaged, and both ranks are assigned to the positive group. The same rule holds for ties based on negative-difference scores.

Step 4. Sum the ranks of each group. When adding, *do not consider ranks with negative signs as negative numbers*. Again, the negative signs before ranks are only there to aid us in arranging the ranks into their appropriate groups.

$$\Sigma R_{pos} : 1.5 + 4 + 6 + 7 + 8 + 9 + 11 = 46.5$$

$$\Sigma R_{neg} : 1.5 + 3 + 5 + 10 = 19.5$$

Step 5. The value T is the smaller of $\Sigma R_{pos}$and ΣR_{neg}. Therefore, T = **19.5**.

Hypothesis Testing and the Wilcoxon Signed-Ranks Test

The Null and Alternative Hypotheses

Similar to the Mann–Whitney U test, the null and alternative hypotheses for the Wilcoxon signed-ranks test are statements regarding the equivalence and nonequivalence of the population distributions:

H_0 : The population distribution of A = the population distribution of B.

The alternative hypothesis is a statement about the nonequivalence of the population distributions:

H_1 : The population distribution of $A \neq$ the population distribution of B.

Finding the Critical Value for *T* and Deciding Whether to Reject the Null Hypothesis

The critical values for the T statistic are found in Table A.12 of the Appendix. T_{crit} is found by locating the number in the left column that corresponds to the number of participants in the study. Move to the column that specifies the desired alpha level. The null hypothesis is rejected if T is less than or equal to T_{crit}.

For the worked example, it was found that T = 19.5. Since there were eleven participants in the study, the critical value, assuming α = .05, is 10. The obtained value of T is *greater* than T_{crit}: 19.5 > 10. Therefore, do not reject the null hypothesis. In conclusion, there was no statistical evidence found to suggest differential effectiveness between the repetition and the association techniques.

The Wilcoxon Signed-Ranks Test with Large Samples

Table A.12 in the Appendix provides critical values for the Wilcoxon T statistic for sample sizes up to 50. When a sample size is greater than 50, T is transformed to a z value. The z_{obt} is compared with a critical value of z found using the z table (Table A.1 of the Appendix). We have made extensive use of the z table throughout this text. Recall that when the z table is used to determine critical values, the z_{crit} that corresponds to α = .05 is ±1.96, when α = .01, z_{crit} = ±2.58, and when α = .10, z_{crit} = ±1.645, for two-tailed tests. Formula 18.6 transforms the Wilcoxon T statistic into a z value.

Formula for the Wilcoxon signed-ranks test for large sample sizes

$$z_{obt} = \frac{T - [n(n+1)/4]}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \quad \text{(Formula 18.6)}$$

where

n = the number of participants in the analysis

When using Formula 18.6, all participants that have a difference score that is 0 are excluded from the analysis. This practice will not affect z_{crit}, but the n value needed for calculation will be the number of participants that are actually *used* in the analysis. When performing a nondirectional test of the null hypothesis, if z_{obt} falls outside of $\pm z_{crit}$, the H_0 is rejected.

Let us run through the computational steps involved in Formula 18.6. Assume T equals 17 and $n = 60$:

$$z_{obt} = \frac{17 - [60(60+1)/4]}{\sqrt{\frac{60(60+1)(2\times 60+1)}{24}}}$$

$$z_{obt} = \frac{17 - 3660/4}{\sqrt{(3660)(121)/24}}$$

$$z_{obt} = \frac{-898}{\sqrt{442\,860/24}}$$

$$z_{obt} = \frac{-898}{\sqrt{18\,452.50}}$$

$$z_{obt} = \frac{-898}{135.84}$$

$$z_{obt} = \mathbf{-6.61}$$

If alpha is set at .05, the critical value for z is ±1.96. The obtained value of z is –6.61; it falls outside the critical value of ±1.96. Consequently, the null hypothesis is rejected.

Box 18.1 reports an interesting study that examines how infants' attention to an adult's speech is influenced by how closely the adult's lip movements

Box 18.1 Do Infants Notice the Difference Between Lip Movement and Speech Sounds?

It is a common tactic, for those who are in the process of losing their hearing, to rely increasingly on observing the lip movements of speakers to help in the processing of auditory information. Actually, most everyone uses this tactic when they find themselves in a situation where there is a high amount of ambient noise. In fact, even under normal listing conditions, we naturally use lip movements to help process speech. This process is so automatic that we are scarcely aware of it. However, we need only to watch a dubbed foreign film to realize how much we take for granted the congruence between lip movements and speech.

Developmental psychologists have found that our awareness of the synchrony between speech sounds and lip movements is evident by the age of 6 months. Researchers are frequently interested in the earliest age a behavioral skill emerges (e.g. Lewkowicz & Hansen-Tift, 2012). Dodd (1979) conducted one of the first investigations in this area. In this study, 10- to 16-week-old babies were found to be able to tell the difference between speakers who show *synchronous* speech and lip movements and speakers who show *asynchronous* speech and lip movements.

Study Method

Infants were placed in a soundproof room with a window through which they could see an adult speaking to them through a microphone. In the Synchrony condition, the sound of the adult's voice was direct, a perfect congruence between lip movements and auditory sound. In the Asynchrony condition, the same infants viewed an adult speaking to them with a 400-millisecond delay between lip movements and speech sounds. The question was, "Are infants this young able to tell the difference between synchronous and asynchronous speech?" If infants are unable to tell the difference, irrespective of experimental condition, they should attend to the speaker approximately the same amount of time. On the other hand, if these infants are able to discriminate between the two conditions, it should be reflected in the amount of time they spend looking at the adult in each condition. Since the point of the study was to see if the infants could make the discrimination, no prediction was made regarding which condition would lead to a greater amount of attentional deployment. In fact, one could speculate that the children would spend more time attending to the synchronous adult because it is a familiar experience. Equally plausible, however, is the hypothesis that children would attend more to the asynchronous adult due to novelty. Clearly, this repeated-measures design requires a nondirectional hypothesis test.

The dependent variable was the percentage of time the child spent looking at the speaker.

Data Analysis and Results

Although the percentage of time looking at the speaker is a reasonable way of measuring attention, Dodd was not so confident in treating the dependent variable as an interval or ratio scale of measurement. There was uncertainty, for example, about whether the difference between attending 20 and 30% of the time reflects the same amount of difference as attending 80 and 90% of the time. Dodd had more confidence that the *rankings* of the time differences in looking indicated an *order* of difference in attending. As a result, Dodd decided to use a Wilcoxon signed-ranks test instead of a dependent-samples *t* test. Since the calculated *T* of 5 is *less than* the T_{crit} of 13, the null hypothesis was rejected. Therefore, we should conclude that infants 10–16 weeks old are able to tell the difference between speech and lip movements that are congruent as opposed to incongruent.

Percentage of time attending

Participant	Asynchrony	Synchrony	Difference	Rank
P_1	50.4	20.3	30.1	10
P_2	87.0	17.0	70.0	12
P_3	25.1	6.5	18.6	6
P_4	28.5	25.0	3.5	3
P_5	26.9	5.4	21.5	8
P_6	36.6	29.2	7.4	5
P_7	1.0	2.9	–1.9	–1
P_8	43.8	6.6	37.2	11
P_9	44.2	15.8	28.4	9
P_{10}	10.4	8.3	2.1	2
P_{11}	29.9	34.0	–4.1	–4
P_{12}	27.7	8.0	19.7	7

$$\Sigma R_{pos} = 10 + 12 + 6 + 3 + 8 + 5 + 11 + 9 + 2 + 7 = 73$$

$$\Sigma R_{neg} = 1 + 4 = 5$$

$$\alpha = 0.05$$

$$T_{crit} = 13$$

$$T = \mathbf{5}$$

correspond to their spoken words. The author uses a repeated-measures design and applies the Wilcoxon signed-ranks test to the ranked data.

18.7 Using Nonparametric Tests

In comparison with parametric tests, nonparametric tests have certain advantages. Nonparametric tests can and should be used when population assumptions for a parametric test are grossly violated. In addition, nonparametric tests require only that data be scaled according to ranks (or, for chi-square, is categorical). Nonetheless, despite these advantages, nonparametric tests are used less often than parametric tests. Why?

First, nonparametric tests tend to be less powerful than parametric tests. The probability of detecting a treatment effect is lower, all other things being equal. In other words, there is a greater probability of making a Type II error when using a nonparametric test on ordinal data than when using a parametric test on interval or ratio data.

Second, when analyzing complex factorial designs, parametric tests, like two- and three-way ANOVAs, generate much more information than any nonparametric test.

Third, two-sample parametric tests analyze population differences between means. Two population distributions can vary in a number of ways: central tendency, variability, skewness, and so on. To test the null hypothesis that $\mu_1 = \mu_2$ requires that other aspects of the population distributions be similar. Since nonparametric procedures do not require these assumptions, they are less specific in what they tell us.

Fourth, statisticians remind us that parametric tests are relatively robust with respect to the violation of population assumptions. Researchers are told that they should become suspicious of the use of parametric analyses when there are *gross* violations of population assumptions: population distributions that depart *radically* from normality and violations of homogeneity of variances. However, *how* gross do the violations have to be to justify using the less powerful nonparametric tests? There are no clear-cut rules for when to transform a continuous measure-dependent variable into ranks. Given this lack of clarity, researchers tend to lean heavily toward the application of parametric analyses. Of course, when data are collected using an ordinal or nominal scale, a nonparametric test is the only option.

As we approach the end of this textbook, Box 18.2 presents a philosophical reflection on the health and current state of the scientific endeavor.

Box 18.2 Is the Scientific Method Broken? The Limitations of Science

Throughout this book we have periodically stopped to look more closely at some commonly discussed problems in the world of science; in particular, we have tried to understand better the current reproducibility crisis that is afflicting many of the behavioral and social sciences. Let us finish this series by stepping back to look a bit more philosophically at the scientific endeavor as a whole. What can we hope to accomplish with the help of science, and what, if anything, lies on the outside? This has sometimes been referred to as science's *demarcation* problem – drawing the line between what *is* science and what *is not* science.

In Chapter 1, we were told that the scientific method addresses a limited set of questions. Causal explanations, when methodologically warranted, are to be understood in a limited way, not as complete, final explanations. Unfortunately, not all claims made by scientists reflect this modesty. In other words, not all pronouncements made by scientists are, in fact, statements of *science*. Some claims made in the name of science are clearly outside this demarcation line. Perhaps the human desire to win arguments tempts us to use periodically the justly earned authority of science regarding topics within its domain to declare matter-of-factly something true that we merely *want* to be true sitting outside its domain. These over-reaches, when exposed, can leave the general public, who is listening in on the conversation, with the impression that science might be broken. A lack of proper modesty by some popularizers of science and the zealotry of others whose primary concern is advocacy for a particular political position can severely damage science's public reputation.

A related problem is the claim made by some that science is the *only* avenue to truth. For instance, the famed philosopher of science, Bertrand Russell (1936/1997), once famously wrote, "… what science cannot discover, mankind cannot know." This position has come to be known as "scientism." This view, however, does not withstand scrutiny. In fact, the statement itself is circular. After all, this claim is not a statement of science, so, if it is true, by its own pronouncement, it cannot be known to be true.

Of course, the scientific method does an excellent job of exploring the physical machinery of reality including social reality. However, it is powerless to answer even simple existential questions like "why am I here?" and "what is the meaning of life?" When discussing the demarcation problem, British biologist/theologian Alister McGrath (2015) refers to Frank Rhodes famous question regarding a boiling kettle. Rhodes asks us to imagine that we find a kettle sitting on a gas ring; and upon closer inspection, we see that the kettle is boiling. "Why is the kettle boiling?" we may ask. Well, one answer addressing the mechanics of the phenomenon would be that there is a heat transfer taking place between

the underlying burner and the bottom of the copper kettle. This transfer then continues on to the water inside the kettle. The additional energy excites the liquid water molecules and eventually changes their physical state being released into the air as steam. This explanation has been revealed to us over the past few hundred years through careful scientific analysis. Rhodes, however, introduces a second explanation. He says, "it is boiling because I want a cup of tea." Now, Rhodes asks, "Which answer is right?" He goes on to say, "[n]ow these are different answers…But both are true, both are complementary and not competitive. One answer is appropriate within a particular frame of reference, the other within another frame of reference. There is a sense in which each is incomplete without the other." Indeed, the answers work together to give a richer explanation of the phenomenon. One further observation might be that we have been asking the *metaphysical* question of "why" much longer than we have been asking the *mechanical* question of "how." Yet, the recent advances in supplying the specific mechanical answers, while interesting and helpful in many ways, have not served to bring us any closer to answering the metaphysical question.

The scientific method, once given the time needed to work through human foibles and limitations, ends up doing a very good job of describing the interworking parts of much of physical reality. However, even here, on the edges of physical reality, we find limitations. For instance, no one understands what gravity actually *is*. We know how to measure it, and we have learned in careful detail how it works, but that is where our understanding stops. What exactly is it? We also do not know what energy is; or time, space, life, and consciousness. Many secondary questions related to these fundamental features can be asked and seem to have been sufficiently answered by using careful scientific investigation. However, this gained information, as helpful as it is, only serves to deepen the mystery around these basic features of our reality. It is interesting to note that these fundamental questions have been around long before the scientific method was established; in fact, they give every indication they will be perennially tied to the human experience. The famed NASA astronomer Robert Jastrow (1992), when speaking of those who wish to give too much credibility to science, once wrote, "For the scientist who has lived by his faith in the power of reason, the story ends like a bad dream. He has scaled the mountain of ignorance; he is about to conquer the highest peak; as he pulls himself over the final rock, he is greeted by a band of theologians who have been sitting there for centuries." Perhaps science is not broken after all; perhaps its' boundaries just need to be more properly considered.

18.8 How to Present Formally the Conclusions for Various Nonparametric Tests

The proper reporting of Spearman and point-biserial correlations is similar to the reporting of Pearson correlations (see Section 15.6, for review). Simply identify which correlation has been run prior to the presentation of the findings. For example, "A Spearman correlation was used to analyze the relationship between Ingratiation and Need for Approval. The analysis found statistical evidence of a positive correlation, $r(8) = .98, p < .05$." This finding suggests those who more often engage in ingratiating behavior also have a higher need for approval.

The proper reporting of Mann–Whitney *U*'s and Wilcoxon *T*'s (or *z*'s) is similar to the proper reporting of *t* tests. Identify the dependent variable and the two conditions on which the dependent variable differs. If measures of centrality are needed, the medians should be used for nonparametric analyses. Make sure to use cautious language (i.e. statistical evidence suggests) and then present the statistic symbol followed by the observed value and then either "$p < .05$" or "*n.s.*," depending on whether or not the null hypothesis can be rejected. For example, "A Mann-Whitney *U* analysis found evidence suggesting Treatment *B* was more effective than Treatment *A*, $U = 10, p < .05$." There is usually no need to report a critical *U* value.

The proper reporting of a failure to reject the null hypothesis from a nonparametric test can be presented in a similar way to previous inferential tests. For example, "A Spearman correlational analysis did not find statistical evidence of a relationship between Ingratiation and Need for Approval, $r(8) = .32$, *n.s.*"

Recall that Section 8.8, contains information about several other common principles for reporting statistical findings. Please consult this portion of the text for more general information about the proper reporting of statistical findings.

Summary

Statisticians have developed an array of nonparametric inferential tests that have become useful additions to the more standard parametric tests. This chapter addressed four of these nonparametric tests.

The Spearman rank correlation is a simplified Pearson formula applied to ordinal data. This analysis can be used when the scatter diagram of *X* and *Y* shows a nonlinear but monotonic relationship. Converting scores to ranks often "straightens out" the scatter plot, thus allowing for the use of the Spearman rank correlation analysis. The Spearman rank analysis can also be used when data are collected using an ordinal scale, one example being the application of r_s to judges' rankings of some variable.

The point-biserial correlation is used when one variable is genuinely dichotomous and the second variable is continuous. Examples of dichotomous variables are employed/unemployed, resident/alien, and single/married. The null hypothesis for both the Spearman and point-biserial correlation analyses is typically $\rho = 0$.

The Mann–Whitney *U* test is the nonparametric alternative to the independent-samples *t* test. When the population assumptions for using a *t* test are violated, the Mann–Whitney *U* test can be applied to data transformed to ranks. The null hypothesis states that the population distributions are the same. The alternative hypothesis states that the population distributions are not the same. If the group members are evenly dispersed across the span of the ranks, the null hypothesis cannot be rejected. As the members of one group tend to cluster near one end of the continuum, the null hypothesis can be rejected.

The Wilcoxon signed-ranks test is the nonparametric alternative to the dependent-samples *t* test. The Wilcoxon test is performed on the ranks of *difference* scores. Similar to the Mann–Whitney *U* test, the null hypothesis states that the population distribution of ranks is the same for both treatment conditions.

Although nonparametric tests have their place in inferential statistics, the decision to use a nonparametric test is not made lightly. Nonparametric tests are not as powerful as their parametric counterparts. They should only be used when data are collected using a nominal or ordinal scale or when the statistical assumptions for using a parametric test are grossly violated.

Using Microsoft® Excel and SPSS® to Calculate Various Nonparametrics

Excel

There are no specific Excel functions for the nonparametric tests covered in this chapter. However, Excel, just like a handheld calculator, can be used to generate the necessary values to perform each of the tests.

General instructions for data entry into Excel can be found in Appendix C.

SPSS

General instructions for inputting data into SPSS can be found in Appendix C.

Data Entry for Spearman and Point-Biserial Correlations

In SPSS, each row of the data file represents a participant. Since bivariate data is used in calculating the Spearman or point-biserial *r*, create a series of variables within **Variable View** corresponding to the variables measured. Then, go to **Data View**, and input the data, being careful to keep the values from each

	biological_sex	popularity_rank	kindness
1	1	11.0	4
2	2	6.0	6
3	1	16.0	7
4	1	5.0	3
5	2	7.5	6
6	1	7.5	8
7	2	1.0	2
8	2	20.0	2
9	2	13.0	6
10	1	16.0	9
11	2	17.0	5
12	1	18.5	7
13	2	18.5	1
14	2	8.0	5
15	1	11.0	7
16	2	14.0	4
17	1	11.0	6
18	2	22.0	9
19	1	3.0	5
20	1	21.0	7
21	2	9.0	6
22	1	15.0	8

Figure 18.3 An example of entered data for a Spearman and point-biserial correlation using SPSS.

participant within a given row. See Figure 18.3 for an example. Figure 18.3 shows three variables; the first is dichotomous, the second is ranked, and the third is ambiguous. SPSS will convert this third variable into ranks if we select the Spearman correlation to be run. In this data set, a Spearman would be used to analyze "popularity rank" and "kindness," and a point-biserial correlation would be used to analyze "biological_sex" and "kindness."

Data Analysis

1) Click **Analyze** on the tool bar, select **Correlate**, and then click **Bivariate**.
2) Use the arrow key to move the variables of interest into the **Variables** box.
3) The default correlation coefficient calculated is the Pearson; leave this box checked if a point biserial is to be run, unclick this box, and click the

Spearman box if a Spearman is to be run. Make a selection regarding the critical *r* value to be calculated – one-tailed or two-tailed.

4) If descriptive statistics are of interest, open the **Options** box in the upper right corner, and click the **Means and standard deviations** option.
5) Click **Ok**.
6) If descriptives were asked for, the first box will present the means, standard deviations, and sample size of all selected variables. The next box is the correlation grid box simply labeled **Correlations** (or **Nonparametric Correlations** if the Spearman was run). Each variable is listed both down the left-side column and across the top of the grid. (SPSS can run multiple correlations at once.) Down the diagonal spine of the correlation grid will be the value 1, representing the correlation between a variable and itself. The correlations of interest can be found by locating the coordinate between one variable on the left-hand column and the other across the top row. The table is redundant showing each correlation from each perspective. Within each correlation box can also be found the probability of getting a Spearman or point-biserial *r* of that size if $\rho = 0$ [**Sig. (2-tailed)**] as well as a count of the number of paired scores (**N**). If the significance value is equal to or less than .05, there is statistical evidence to reject the null hypothesis. (See Figure 18.4 for a worked example of both a Spearman [top] and a point-biserial [bottom] analysis.)

Correlations

			popularity_rank	kindness
Spearman's rho	popularity_rank	Correlation coefficient	1.000	.276
		Sig. (2-tailed)	.	.214
		N	22	22
	kindness	Correlation coefficient	.276	1.000
		Sig. (2-tailed)	.214	.
		N	22	22

Correlations

		kindness	biological_sex
kindness	Pearson correlation	1	–.399
	Sig. (2-tailed)		.066
	N	22	22
biological_sex	Pearson correlation	–.399	1
	Sig. (2-tailed)	.066	
	N	22	22

Figure 18.4 Output tables from worked examples using SPSS to run a Spearman and point-biserial correlation.

Data Entry for Mann–Whitney *U*

In SPSS, each row of the data file represents a participant. Since both samples in a Mann–Whitney *U* test have different participants, all of the dependent variable data from both samples will need to be placed in one column. Within **Variable View**, label this variable appropriately. However, also create a second variable that will allow the user to identify which data go with which group. A typical label for this variable might be "condition." Then, go to **Data View**. Input the sample data to the appropriate column, and use a nominal variable in the "condition" column to distinguish the two samples (either "0" and "1" or "1" and "2" are typical). (See Figure 18.5 for an example.)

Figure 18.5 An example of entered data for a Mann–Whitney *U* test using SPSS.

	popularity_rank	condition
1	20.0	1
2	5.0	2
3	21.0	2
4	16.0	2
5	1.0	1
6	17.0	2
7	22.0	1
8	2.0	1
9	6.5	2
10	18.0	1
11	13.0	2
12	3.0	2
13	15.0	1
14	6.5	1
15	8.0	2
16	19.0	1
17	10.0	2
18	23.0	2
19	4.0	1
20	13.0	1
21	13.0	1
22	11.0	1
23	9.0	1

Data Analysis

1) Click Analyze on the tool bar, select **nonparametric tests**, then **legacy dialogs**, and then **2 Independent Samples…**.
2) Highlight the dependent variable column label in the left box, and click the arrow to move it into the **Test Variable(s)** box. Move the "condition" variable to the **Grouping Variable** box.
3) Because there may be more than two conditions identified under our grouping variable, click **Define Groups** to identify which two groups we want to compare. Place the nominal values used to distinguish the groups into the two group boxes – one in each. Click **Continue**.
4) Click **OK**.
5) The output will generate two boxes. The first box will identify how many scores were in the sample (**N**) as well as the mean rank and the sum of the ranks. The second box will identify, among other things, the Mann–Whitney *U* value (**Mann–Whitney U**) as well as the probability of getting that value if the null hypothesis of identical populations is true (**Asymp. Sig. (2-tailed)**). It does not show us U_{crit}. As with previous inferential tests run in SPSS, either we can find U_{crit} ourselves, or we can look at the given significance level to see if that value is equal to or lower than .05. If it is, we can reject the null. If it is not, we need to fail to reject the null hypothesis. (See Figure 18.6 for a worked example.)

Ranks

	Condition	*N*	Mean rank	Sum of ranks
popularity_rank	1	13	11.81	153.50
	2	10	12.25	122.50
	Total	23		

Test statistics[a]

	popularity_rank
Mann-Whitney U	62.500
Wilcoxon W	153.500
Z	–.155
Asymp. sig. (2-tailed)	.877
Exact sig. [2*(1-tailed Sig.)]	.879[b]

[a] Grouping variable: condition
[b] Not corrected for ties

Figure 18.6 Output tables from a worked example using SPSS to run a Mann–Whitney *U* test.

Data Entry for Wilcoxon

In SPSS, each row of the data file represents a participant. Since each participant is being measured twice, we will need two columns to hold the raw data. Within **Variable View**, label the two column headings using terms that will distinguish between the two conditions of the study (e.g. Pre/Post, Exp/Control, Cond1/Cond2, etc.). Then, go to **Data View**. Input the sample data to the appropriate column, being careful to keep the data from each participant within the same row, as this will be essential for creating the proper difference score. (See Figure 18.7 for a worked example.)

Figure 18.7 An example of entered data for a Wilcoxon signed-ranks test using SPSS.

	pretest	posttest
1	4	5
2	7	8
3	3	3
4	8	8
5	9	9
6	3	3
7	4	4
8	8	9
9	7	8
10	6	7
11	3	3
12	9	9
13	3	4
14	2	2
15	8	8
16	9	9
17	10	9
18	5	6
19	6	8
20	7	7
21	3	4
22	7	8
23	.	.

Data Analysis

1) Click **Analyze** on the tool bar, select **Nonparametric Tests**, then **Legacy Dialogs**, and then click **2 Related Samples**.
2) Highlight one variable, and use the right arrow key to move it into the **Variable1** box. Move the other variable to the **Variable2** box in the same manner. (Disregard the new row of boxes that are added underneath. These are for running more than one dependent-samples *t* test at a time.) Leave the default option (Wilcoxon) checked. (If descriptives are wanted, click **Options** and then check **Descriptives** and then **Continue**.)
3) Click **OK**.
4) The output will generate two boxes. The first will identify the number of observations (**N**), the mean ranks (**Mean Rank**), and the sum of the ranks (**Sum of Ranks**) for the negative ranks, positive ranks, ties, and total. The second box will generate the test statistic – in this case the Wilcoxon *z*. As always, either we need to go and find the appropriate critical value, or we need to look at the **Asymp. Sig. (2-tailed)** value to see how likely it is to get a Wilcoxon *z* score of that size if the populations are identical. (See Figure 18.8 for a worked example.)

Ranks

		N	Mean rank	Sum of ranks
Posttest - pretest	Negative ranks	1[a]	5.50	5.50
	Positive ranks	10[b]	6.05	60.50
	Ties	11[c]		
	Total	22		

[a] posttest < pretest

[b] posttest > pretest

[c] posttest = pretest

Test statistics[a]

	Posttest - pretest
Z	–2.673[b]
Asymp. sig. (2-tailed)	.008

[a] Wilcoxon signed ranks test

[b] Based on negative ranks

Figure 18.8 Output tables from a worked example using SPSS to run a Wilcoxon signed-ranks test.

Key Formulas

Formula for Spearman rank correlation, r_s

$$r_s = 1 - \frac{6\Sigma D^2}{n_p\left(n_p^2 - 1\right)} \quad \text{(Formula 18.1)}$$

Formula for point-biserial correlation, r_{pb}

$$r_{pb} = \frac{M_{Y_1} - M_{Y_0}}{s_y}\sqrt{\frac{n_1 n_0}{n(n-1)}} \quad \text{(Formula 18.2)}$$

Formula for computing U_A

$$U_A = n_A n_B + \frac{n_A(n_A+1)}{2} - \Sigma R_A \quad \text{(Formula 18.3)}$$

Formula for computing U_B

$$U_B = n_A n_B + \frac{n_B(n_B+1)}{2} - \Sigma R_B \quad \text{(Formula 18.4)}$$

The U to z_U transformation formula

$$z_U = \frac{U - (n_A n_B/2)}{\sqrt{\frac{n_A n_B(n_A + n_B + 1)}{12}}} \quad \text{(Formula 18.5)}$$

Formula for the Wilcoxon signed-ranks test for large sample sizes

$$z_{obt} = \frac{T - [n(n+1)/4]}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \quad \text{(Formula 18.6)}$$

Key Terms

Parametric tests
Spearman rank correlation
Point-biserial correlation
Dichotomous variable
Dummy coding
Mann–Whitney *U* test
Wilcoxon signed-ranks test

Questions and Exercises

1 When we need to convert scaled data into ranks, how is it done?

2 How does the transformation of scaled data into ranked data address the problem of outliers?

3 Which of the following nonparametric tests should be used for a study that aims at investigating the relationship between people's running time, as determined by their finishing position in a race, and their T-shirt size (YL, S, M, L, XL, XXL).
a Spearman rank correlation
b Point-biserial correlation
c Mann–Whitney *U* test
d Wilcoxon signed-ranks test

4 Which of the following nonparametric tests would best be used to analyze the data from a study that aims at investigating a null hypothesis of no population differences on memory recall between participants who are taught to use the "peg-word" mnemonic device system and other participants who are taught to use the "narrative" mnemonic device system?
a Spearman rank correlation
b Point-biserial correlation
c Mann–Whitney *U* test
d Wilcoxon signed-ranks test

5 Which nonparametric test involves "tagging" scores from different samples so that they can be returned to their samples after they have been converted into ranks?

6 If given the luxury of choosing between a parametric and a nonparametric analytical tool, which one should be chosen, and why?

7 The following represents a bivariate distribution, using continuous measures. Convert these scores to ranks. Assign a rank of 1 to the lowest score.

X	Y
3	7
2	2
4	4
9	12
8	8
4	2

8 For the data presented in Problem 7, assume that *X* represents an experimental group and *Y* represents a control group. We plan to conduct a Mann–Whitney *U* test. Convert the scores to ranks.

9 For the data in Problem 7, assume that a repeated-measures design is used and each row is data from a given participant. We plan to conduct a Wilcoxon signed-ranks test. Convert each participant's scores into a rank.

10 Convert the following scores into ranks. Assume they represent a bivariate distribution, both variables using a continuous measure. For both variables, assign the highest score a value of 1.

X	Y
77	45
54	45
96	83
12	37
73	93
76	14
56	52
96	85
68	62
15	19

11 Perform a Spearman rank correlation on the ranked data from Problem 10. Can the null hypothesis be rejected if $\alpha = .05$ for a two-tailed test? Would it have changed the outcome if we assigned the lowest score a value of 1?

12 Convert the following scores into ranks. Assume they represent a bivariate distribution, both variables using a continuous measure. For both variables, assign the lowest score a value of 1.

X	Y
11	19
14	10
16	15
11	15
15	16
16	11
18	14
11	19
12	19
17	12

13 Perform a Spearman rank correlation on the ranked data from Problem 12. Can the null hypothesis be rejected if $\alpha = .05$ for a two-tailed test?

14 Perform a point-biserial correlation on the following data.

Biological Males	Biological Females
2	4
5	11
9	10
3	7
10	7
3	7
7	12
7	14
9	

15 A researcher is interested in seeing if there is a relationship between "Need for Affiliation" and "Fear of Criticism." Questionnaires that measure each trait are administered to eight participants. For the following data set:

a Draw the scatter plot of the continuous measures.
b Convert the scores to ranks.
c Draw the scatter plot based on ranks.
d Compute the Spearman rank correlation.
e State the null and alternative hypotheses.
f Use $\alpha = .05$ to see if the null hypothesis can be rejected (two tailed).

	Need for Affiliation		Fear of Criticism	
Participant	Score	Rank	Score	Rank
P_1	16		40	
P_2	14		35	
P_3	14		30	
P_4	12		18	
P_5	10		14	
P_6	8		13	
P_7	9		12	
P_8	4		4	

16 A social psychologist hypothesizes a relationship between Physical Attractiveness and Popularity. Ten high school students are ranked on each variable. For the following ranks:

a Compute the Spearman rank correlation.
b State the null and alternative hypotheses.
c Use $\alpha = .05$ to see if the null hypothesis can be rejected (two tailed).
d Interpret the findings.

	Physical Attractiveness	Popularity
Participant	Rank	Rank
P_1	1	1
P_2	2	3
P_3	5	2
P_4	3	4
P_5	4	5
P_6	7	7
P_7	9	6
P_8	6	8
P_9	8	9
P_{10}	10	10

17 A sociologist is interested in the relationship between political affiliation and attitudes toward military intervention in Central America. The measure of attitudes is continuous, with 1 meaning "no intervention" and 10 meaning "aggressive intervention."

a Compute r_{pb}.
b State the null and alternative hypotheses.
c Use $\alpha = .05$ to see if the null hypothesis can be rejected (two tailed).
d Interpret the correlation.
e State how much of the variance in attitudes is due to political affiliation.

Democrat	Republican
1	10
4	7
7	8
3	6
2	9
1	5
1	10

18 A teacher developed a mathematical ability test and believes that the answer to one particular question is correlated with the total score on the test. The teacher assigns a 0 if the answer is correct and a 1 if the answer is incorrect.

a Calculate r_{pb}.

b State the null and alternative hypotheses.

c Use $\alpha = .05$ to see if the null hypothesis can be rejected (two tailed).

d Interpret the correlation.

Participant	Question	Test Score
1	0	36
2	0	39
3	1	16
4	1	14
5	0	22
6	1	26
7	1	9
8	1	7
9	0	30
10	1	11

19 In a dog show, rankings are based on Body Shape and Posture. For the following results:

a Compute r_s.

b State the null and alternative hypotheses.

c Use $\alpha = .05$ to see if the null hypothesis can be rejected (two tailed).

d Interpret the findings.

Dog	Posture	Body Shape
1	1	2
2	2	1
3	3	3
4	7	5.5
5	9	7
6	4	9
7	5	5.5
8	6	8
9	8	4

20 Twelve medical students are ranked on their clinical and written examination performance over the past year.
a Calculate the Spearman rank correlation.
b State the null and alternative hypotheses.
c Use $\alpha = .05$ to see if the null hypothesis can be rejected (two tailed).
d Interpret the findings.

Student	Clinical	Written
1	4	2
2	12	10
3	1	1
4	7	5
5	8	8
6	2	3
7	11	9
8	3	4
9	9	7
10	6	6
11	5	11
12	10	12

21 A child psychologist hypothesizes a relationship between when a child first walks (months) and whether the child has an older sibling. For the following data set:
a Calculate r_{pb}.
b State the null and alternative hypotheses.
c Use $\alpha = .05$ to see if the null hypothesis can be rejected (two tailed).
d Interpret the correlation.

Older Sibling (0)	No Older Sibling (1)
10.9	11.6
11.2	13.7
11.4	15.2
12.4	10.9
10.3	16.0
10.0	15.8
12.0	12.8
11.9	10.8
13.2	14.7
11.4	15.0

22 A psychologist hypothesizes a relationship between expressed gender and attitudes toward state-mandated paid maternity leave. The range of values measuring attitudes is from 1 – strongly opposed to 10 – strongly in favor. Males = 0 and females = 1.

a Compute the appropriate correlation coefficient.
b State the null and alternative hypotheses.
c Use $\alpha = .05$ to see if the null hypothesis can be rejected (two tailed).
d Interpret the correlation.

Participant	Expressed Gender	Attitudes about Paid Maternity Leave
1	0	3
2	0	6
3	1	9
4	0	4
5	1	1
6	1	10
7	1	8
8	0	3
9	1	5
10	0	9

23 Eight students are ranked by a faculty member based on their performance in a Statistics class. A year later, the same students are ranked on the quality of their Senior Thesis.

a Compute r_s.
b State the null and alternative hypotheses.
c Use $\alpha = .05$ to see if the null hypothesis can be rejected (two tailed).
d State the amount of variance in the quality of the papers that can be attributed to performance in the statistics class.

Statistics	Senior Thesis
1	1
3	2
5	4
8	8
6	6
7	3
2	7
4	5

24 A clinical psychology program has two training tracks: Behavioral Therapy and Psychoanalysis. First-year graduate students are randomly assigned to these tracks. A professor wonders if there is a difference between the tracks in how well students learn basic interviewing skills. After one year of training, the professor ranks *all* the students as to how well they demonstrate fundamental interviewing techniques. For the following data set, perform the appropriate nonparametric test. Higher ranks indicate better interviewing skills.

a State the null and alternative hypotheses.
b Compute the appropriate test statistic.
c Use $\alpha = .05$ to see if the null hypothesis can be rejected (two tailed).
d Interpret the findings.

Behavioral Therapy	Psychoanalysis
12	10
9	2
11	3
8	1
4	5
7	6

25 When conducting a Mann–Whitney *U* test when one of the sample sizes is greater than 20, what should we do, and why?

26 Suppose we are about to perform a Wilcoxon signed-ranks test and we notice that three participants out of 20 have a difference score of 0. What should we do when it comes time to separate positive and negative ranks?

27 Assume that we are about to perform a Mann–Whitney *U* test. We look at the ordering of ranks between the two groups and observe that the rankings indicate the null hypothesis to be as wrong as it can possibly be. What number will the smaller $\sum R$ be? What will *U* equal?

28 Perform a Mann–Whitney *U* test on the following ranked data. Even though sample sizes are small, use the z_U formula. Let $\alpha = .05$.

R_A	R_B
1	2
3	6
4	7
5	8
9	10

29 A high school counselor wonders if the type of music played during lunch hour influences the speed with which students eat. In one condition, soft, new-age music is piped through the sound system. In a second condition, Hip Hop is played. Assume that relevant aspects of the setting are controlled (e.g. menu, portions, seating arrangements, etc.). Six students are observed for 5 minutes, and the dependent variable is the average number of bites per minute. Since this is a repeated-measures design, half of the participants listen to the New Age music first, and half of the participants listen to Hip Hop first. The dependent variable is measured during each experimental condition. Perform a Wilcoxon signed-ranks test on the following data. Set alpha at .05.

	Music	
Participant	**New Age**	**Hip Hop**
P_1	2	6
P_2	1	6
P_3	3	2
P_4	4	8
P_5	3	6
P_6	1	5

30 Assume that the data from Problem 29 are obtained using an independent-groups design. Perform a Mann–Whitney U test, with $\alpha = .05$.

31 For each of the following situations, specify which statistic we should compute.

a Design: independent groups
Data: ordinal

b Design: repeated measures
Data: ordinal

c Design: correlational
Data: nonlinear relationship between X and Y

d Design: correlational
Data: X is dichotomous, Y is continuous

e Design: correlational
Data: X and Y are ordinal

32 Perform a Wilcoxon signed-ranks test using the formula for large samples. $T = 14$ and $n = 55$. Set alpha at .05 and test H_0.

Computer Work

33 (This problem is similar to the Chapter 9, Problem 33. However, different data will be used.) We observe that people seem to be happier when they are wearing a new article of clothing. To test this, we provide a small random sample of our students with a new T-shirt and instruct them to wear the shirts all day. At the end of the day, independent judges ranked the happiness (or what social psychologists call "subjective well-being") of each participant. A control group of students is also judged in terms of happiness, but without the experimental manipulation. The ratings are below. Lower scores indicate greater happiness. Conduct a Mann–Whitney *U* test. Use a two-tailed test and set α=.05.

Happiness	
New T-shirt	**Control**
8	9
10	6
19	22
7	11
1	23
4	24
5	12
3	14
13	2
20	15
21	18
16	17

34 A cardiologist is testing the effectiveness of Propranolol versus a Diuretic for lowering systolic blood pressure. Seven participants are started on propranolol, and eight participants are started on a diuretic. After 90 days, all participants switch to the other drug. The systolic blood pressure readings for each participant are provided in the following table. (This data set could be analyzed with a dependent-samples *t* test. However, for the sake of practice, use the corresponding nonparametric test.)

a State the null and alternative hypotheses.
b Compute the appropriate nonparametric test statistic.
c What is the critical value for α = .05 (two-tailed test)?
d Interpret the findings.

Participant	Propranolol	Diuretic
P_1	127	140
P_2	116	130
P_3	120	150
P_4	132	132
P_5	110	111
P_6	125	120
P_7	131	138
P_8	129	148
P_9	134	149
P_{10}	119	118
P_{11}	116	149
P_{12}	119	121
P_{13}	116	124
P_{14}	144	152
P_{15}	115	126

35 Using the data from Problem 22, assume that the design is an independent-groups design. Perform a Mann–Whitney U test on the data. Set alpha at .05.

36 A clinical psychologist who works with alcoholics is interested in the effects of Antabuse (a substance that leads to nausea and illness when alcohol is ingested) versus Vitamin B_1 in preventing relapse. Three adult volunteers take Antabuse for three months and then switch to vitamin B_1 for three months. Three other participants begin with the vitamin and then switch to Antabuse. Total alcohol-free days (out of 91) for each participant are listed below. Conduct a Wilcoxon signed-ranks test. Use a two-tailed test to see if the null hypothesis can be rejected. Set $\alpha = .05$.

Participant	Antabuse	Vitamin B_1
P_1	80	75
P_2	65	57
P_3	91	89
P_4	52	50
P_5	45	35
P_6	81	62

Part 7 Review

Nonparametric Tests

Review of Concepts Presented in Part 7

As with previous section reviews, the purpose here is to revisit both the similar concepts that hold together the statistical tests presented in Chapters 17 and 18 and the concepts that distinguish them one from another. First, let us look at the primary similarity holding all of these procedures together; all of the tests are nonparametric. Nonparametric tests do not make assumptions about population parameters and do not require descriptive statistics that derive from interval- or ratio-scaled data, like means and standard deviations. These nonparametric tests allow the researcher to investigate null hypotheses for frequency count data as well as ordinal-scaled data. Furthermore, nonparametric tests can be used to analyze interval or ratio data that do not meet the population assumptions required by parametric tests. Methodological assumptions of representativeness and independent observations are still present, but if there are gross violations of normality or homogeneity of variance, the standard parametric tests are invalid and nonparametric tests are used as a substitute. With the added flexibility of nonparametrics, however, there is a trade-off. These tests are typically less powerful than their corresponding parametric counterparts. Nonetheless, if the data are not measured on an interval or ratio scale, or if a statistical assumption cannot be met, nonparametric procedures are a welcomed analytical tool. As a result, nonparametrics are typically not a researchers "Plan A" for analysis; however, they are a grateful "Plan B" to be used when needed.

As we might expect, there are many differences between the six nonparametric procedures presented in these final two chapters. A quick review may be helpful. The first distinction to be made is based on the type of data gathered. The two

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

chi-square tests are used when frequency count data has been collected. In these situations, participants are not measured in terms of "how much" but rather in terms of "how many" – how many participants fit into this category versus other categories. These measures are of the all-or-nothing kind. It may help to think of a "taste-test" situation. Participants sample several, say, soft drinks and then select the one they most prefer. The data being gathered is not how much the participant likes each option but rather which one of the options is selected. The two different chi-squares can be distinguished based on the design of the study. If there is only one dimension or factor, the goodness-of-fit chi-square is the test to be run. If, however, conditions vary across two dimensions or factors, then the chi-square test for independence is the analytical tool needed.

The null hypothesis for a goodness-of-fit test depends on what "nothing" or "no difference" means for a particular investigation. It may mean no difference between the various conditions, it may mean no change from the last time the data was collected, or it may be no difference from a theoretically derived set of expected frequencies. This needs to be carefully determined as the expected frequencies required for calculation of the chi-square statistic are derived from a proper understanding of the null hypothesis. The null hypothesis for a chi-square test for independence, however, is always the same – no relationship between the two factors. In this test there are no predictions based on the relative frequency of counts across a given dimension or factor; rather the null hypothesis states that there is no relationship between the two factors. (This is similar to investigating an interaction effect in a two-way factorial design.) The procedure for both chi-square tests is similar; compare the observed frequency counts with the expected frequency counts derived from a null hypothesis. The greater the cumulative disparity across the various conditions, the more likely the null hypothesis is false.

Chapter 18 brings us to two additional correlation procedures to be added to the Pearson *r* presented in Chapter 15. The Pearson *r* assumes both variables are measured on an interval or ratio scale, the data are normally distributed, and the two variables are not curvilinearly related. The Spearman rank correlation can be used if either or both variables are measured ordinally – if there is a violation of normality in the data or if there is a monotonic (i.e. does not reverse direction) curvilinear relationship between the variables. The Spearman converts the raw data into ranks and generates an *r* based on this ranked data. The point-biserial correlation is used when one variable is continuous and the other variable is dichotomous, that is, an either/or measure. Data from the continuous measure are not altered, and the dichotomous data are dummy coded into a "1" and "0." When using this procedure, it is important to keep in mind which variable is given the value of "1" and which is given the value of "0." This will be needed for proper interpretation of any finding.

The final two tests introduced in Chapter 18 are the nonparametric alternatives to the independent- and dependent-samples *t* tests. The Mann–Whitney *U* test is used when there are two independent groups being compared. The test

requires the data from both groups to be compiled into one and then ranked from lowest to highest. There are two ways to determine the *U* statistic. One uses a point system based on how many values from the other group fall below each given value, and the other is a rank-counting procedure based on this organization of all of the scores in the study. A *U* is generated for each group. If the two groups are well interspersed, the two *U* values will be fairly large and roughly equal. If, however, the values from one group tend to fall toward one end of the continuum and the values from the other group tend toward the other end, the *U* values will be very different, and one of them will be quite low. The null is rejected if the lowest *U* value falls *below* the critical *U* value as determined by Table A.10 found in the appendix.

The Wilcoxon *T* is determined in a similar way, except the ranked scores will be the difference scores in a dependent-samples research situation. In this procedure, differences are found between each pair of scores – subtracting the second value from the first. The valence of the difference score is temporarily set aside, and the difference scores are ranked. Then, the rank values are separated based on the valence of the difference score into two values, a sum of the negative ranks and a sum of the positive ranks. If the null hypothesis of no difference is true, then some of the differences will be positive and some will be negative. The sum of the ranks for both groups will be rather similar, and neither sum will be very small. If, however, the null hypothesis of no difference is false, either the positive or the negative ranks will be much smaller than the other. Once again, the *smaller* sum value of the ranks is used as the observed value, and it is compared with a critical score found in Table A.11 of the appendix. For large sample sizes, a *z* formula can be used to test the null hypothesis for either an independent-samples or dependent-samples research situation.

Now that we are at the end of the text, we can be presented with an opportunity to test our ability to connect the appropriate statistical tool to the appropriate research analysis situation. Understandably, the exercises at the end of each particular chapter only require the use of the test(s) found and studied within that chapter for solution. The questions and exercises at the end of chapters are designed to get us familiar with using the tools immediately just described. They are not designed to challenge our diagnostic skills (i.e. knowing which test among many to use for a given situation). The following review section, however, is designed to help us develop these diagnostic skills.

The exercises below will help us review the conceptual differences between the various nonparametric tests explored in Chapters 17 and 18, as well as the *t* tests, ANOVAs, and bivariate analyses introduced in the preceding chapters. The hypothesis testing exercises will not identify which test is appropriate for the described scenario. We will need to use the available information presented in the exercise to make that determination. (Note: Most of the exercises below involving data can be solved either with or without the use of statistical software.)

Questions and Exercises

1 What is the difference between a parametric test and a nonparametric test?

2 Which type of test (parametric or nonparametric) is to be preferred and why?

3 Match the appropriate statistical tool at the bottom with each of the following descriptors. (Identify the statistical tool with the assigned number.)

a Categorical data across one factor
b Independent-group design (2 cells) with ordinal data
c Independent-group design (2 cells) with ratio data – all assumptions met
d Bivariate data, one variable being on an ordinal scale
e Categorical data across two factors
f Repeated-measures design (2 cells) – all assumptions are met
g Repeated-measures design (2 cells), but homogeneity of variance assumption grossly violated
h Use of bivariate data to predict unknown value

1) Wilcoxon
2) Mann–Whitney *U*
3) Spearman correlation
4) Chi-square "test for independence"
5) Chi-square "goodness-of-fit" test
6) Regression
7) Pearson correlation
8) Independent-samples t test
9) Dependent-samples t test
10) Point-biserial correlation

4 Find a measure of relationship between eating behavior and books read per year. Can the null hypothesis of no relationship be rejected?

Vegetarian	Carnivore
14	23
17	27
18	20
11	35
9	16
14	15
5	7

5 (This problem uses the data and research scenario from the Part 4 Review, question #10.) A researcher is interested in the effect of emotion on concentration. A two-sample study is designed in which anger is induced in one sample by having a confederate provoke an argument in the lab waiting room. The control group does not undergo this mood induction. Both samples are then tested on a computer stunt driving game, and the number of times the participant runs the vehicle into an object (crashes) is counted. Suppose there is reason to believe the assumption of normality is grossly violated in the population data from which these samples are drawn. Please choose the proper nonparametric test to see if the null of no difference can be rejected. Use a two-tailed test and set $\alpha = .05$. Compare this answer with Part 4 Review, question #10.

Angry group	Control group
6	6
9	5
13	8
11	6
5	9
10	7

6 A researcher believes that marital status is a determining factor in the kinds of pets people have. Data regarding marital status is collected from a number of pet owners, with the following results (assume that each person has only one pet). Can the null hypothesis of no relationship between marital status and type of pet be rejected?

	Dog	Cat	Bird	Fish	Snake
Married	47	25	5	10	2
Single	22	40	8	20	2
Divorced/widowed	18	43	14	15	1

7 A sleep researcher believes that people will experience a different number of dreams depending on the temperature of the room in which they are sleeping. Adult volunteers are asked to sleep for ten nights in an 80 °F room and for ten nights in a 65 °F room. The temperature is alternated randomly to

prevent habituation. The total number of dreams reported by each participant is given below. The researcher believes the tendency to recall dreams is terribly skewed, with most people hardly remembering any and only a few people who claim to remember them frequently. For this reason, the sleep researcher suggests using a nonparametric test. Please choose the proper nonparametric test to see if the null of no difference can be rejected. Use a two-tailed test and set $\alpha = .05$.

Participant	80 °F room	65 °F room
P_1	5	10
P_2	7	7
P_3	15	20
P_4	12	18
P_5	10	16
P_6	8	8

8 A school psychologist would like to determine whether there are differences in reading preferences among the students in a particular junior high school. The number of each type of book checked out of the school library (in both hardbound as well as electrical versions) is tallied over a six-month period. Perform the appropriate test to see if the null hypothesis of no differences between book type can be rejected.

Nonfiction	Sports	Romance	Science fiction	Classics
68	75	55	50	52

9 A researcher would like to examine the correlation between stress and reading comprehension. The researcher randomly selects a sample of nine first-year college students. Participants are asked to rate their current level of stress on a 1–10 scale, with 1 = no stress and 10 = extreme stress. They are then given a short story to read and a 15-item comprehension test upon completion of the story. Stress levels and test scores are listed below. What test should be used to measure the relationship and test the null hypothesis of no relationship? Why that particular test? Make a decision regarding the null.

Stress level	Test score
2	3
6	11
9	10
3	6
10	9
3	4
8	12
7	14
9	8

10 A biology professor theorizes that caffeinated sodas cause more burping than noncaffeinated sodas. A large sample of students are gathered and randomly assigned to drink sodas with or without caffeine. The professor then waits 10 minutes and classifies each student as having either burped or not burped during that time period. What test should be used?

11 Suppose the biology professor mentioned in the previous question decided to count the number of burps each student generated. Would the test used to analyze the data change? If so, what would be the appropriate test?

12 A university administrator looked at the number of senior students majoring in sociology and found 21 out of 24 to be biological females. Assuming the student body at this university is roughly equal in terms of biological sex, what test would the administrator use to see if this academic program is overrepresented with biological females?

13 A team of social science researchers wants to know more about social media usage. They are interested in which platforms university students prefer, what type of students use them heavily, and for what social purposes they are used.

a Think of a research situation that would employ a chi-square test for independence.

b Think of a research situation that would employ an independent-samples *t* test.

c Think of a research situation that would employ a Wilcoxon signed-ranks test.

14 In fantasy baseball, groups of pretend owners conduct a draft in which they can "buy" baseball players to fill out their roster for an upcoming season. These made-up teams are then compared based on the individual performance of each team member of each team. At the end of the season, team winners are declared based on the cumulative performance of the roster of players. This practice has become quite popular – and many leagues of friends have stayed together for 20 years or more. Because of some small rule differences between the two leagues within professional baseball, players drafted from National League teams tend to have better defensive numbers, while players drafted from American League teams tend to have better offensive numbers. This opens the door to an interesting question: Is there an advantage to drafting players from one league or the other (all other things being equal) in a fantasy league situation? One way to look at this issue might be to see which made-up teams have won the fantasy league in the past – teams with predominantly National League players or teams with predominantly American League players. If we had access to this data, describe a methodological situation that would employ the following analytical tools.

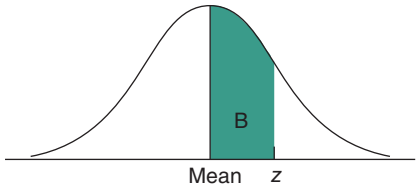
a A chi-square goodness-of-fit test
b A point-biserial correlation
c A Spearman rank correlation

15 A social worker wants to see if there is a relationship between literacy and marital status among indigent mothers in a given city. Access is gained to what is believed to be a random sampling of these individuals and gains information from each one regarding their literacy (illiterate or literate) and their marital status (single, married, divorced/widowed). What analytical tool should be used to see if there is a relationship?

16 Suppose the social worker from the previous question decides to measure literacy by giving each participant a test that scores the degree of literacy possessed by an individual; it is claimed to be a continuous measure. Furthermore, suppose it is decided to simplify the marital status dimension by simply noting if each person claims to have a significant other or not. Now which analytical tool would be best to use?

Appendix A

Statistical Tables



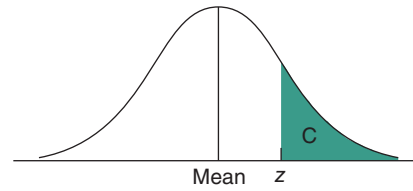


Table A.1 *z* Table.

(A)	*(B)*	*(C)*	*(A)*	*(B)*	*(C)*	*(A)*	*(B)*	*(C)*
z	Area between mean and *z*	Area beyond *z*	*z*	Area between mean and *z*	Area beyond *z*	*z*	Area between mean and *z*	Area beyond *z*
.00	.0000	.5000	.10	.0398	.4602	.20	.0793	.4207
.01	.0040	.4960	.11	.0438	.4562	.21	.0832	.4168
.02	.0080	.4920	.12	.0478	.4522	.22	.0871	.4129
.03	.0120	.4880	.13	.0517	.4483	.23	.0910	.4090
.04	.0160	.4840	.14	.0557	.4443	.24	.0948	.4052
.05	.0199	.4801	.15	.0596	.4404	.25	.0987	.4013
.06	.0239	.4761	.16	.0636	.4364	.26	.1026	.3974
.07	.0279	.4721	.17	.0675	.4325	.27	.1064	.3936
.08	.0319	.4681	.18	.0714	.4286	.28	.1103	.3897
.09	.0359	.4641	.19	.0753	.4247	.29	.1141	.3859

(*Continued*)

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

Table A.1 (Continued)

(A)	(B)	(C)	(A)	(B)	(C)	(A)	(B)	(C)
z	Area between mean and *z*	Area beyond *z*	*z*	Area between mean and *z*	Area beyond *z*	*z*	Area between mean and *z*	Area beyond *z*
.30	.1179	.3821	.66	.2454	.2546	1.02	.3461	.1539
.31	.1217	.3783	.67	.2486	.2514	1.03	.3485	.1515
.32	.1255	.3745	.68	.2517	.3686	1.04	.3508	.1492
.33	.1293	.3707	.69	.2549	.2451	1.05	.3531	.1469
.34	.1331	.3669	.70	.2580	.2420	1.06	.3554	.1446
.35	.1368	.3632	.71	.2611	.2389	1.07	.3577	.1423
.36	.1406	.3594	.72	.2642	.2358	1.08	.3599	.1401
.37	.1443	.3557	.73	.2673	.2327	1.09	.3621	.1379
.38	.1480	.3520	.74	.2704	.2296	1.10	.3643	.1357
.39	.1517	.3483	.75	.2734	.2266	1.11	.3665	.1335
.40	.1554	.3446	.76	.2764	.2236	1.12	.3686	.1314
.41	.1591	.3409	.77	.2794	.2206	1.13	.3708	.1292
.42	.1628	.3372	.78	.2823	.2177	1.14	.3729	.1271
.43	.1664	.3336	.79	.2852	.2148	1.15	.3749	.1251
.44	.1700	.3300	.80	.2881	.2119	1.16	.3770	.1230
.45	.1736	.3264	.81	.2910	.2090	1.17	.3790	.1210
.46	.1772	.3228	.82	.2939	.2061	1.18	.3810	.1190
.47	.1808	.3192	.83	.2967	.2033	1.19	.3830	.1170
.48	.1844	.3156	.84	.2995	.2005	1.20	.3849	.1151
.49	.1879	.3121	.85	.3023	.1977	1.21	.3869	.1131
.50	.1915	.3085	.86	.3051	.1949	1.22	.3888	.1112
.51	.1950	.3050	.87	.3078	.1922	1.23	.3907	.1093
.52	.1985	.3015	.88	.3106	.1894	1.24	.3925	.1075
.53	.2019	.2981	.89	.3133	.1867	1.25	.3944	.1056
.54	.2054	.2946	.90	.3159	.1841	1.26	.3962	.1038
.55	.2088	.2912	.91	.3186	.1814	1.27	.3980	.1020
.56	.2123	.2877	.92	.3212	.1788	1.28	.3997	.1003
.57	.2157	.2843	.93	.3238	.1762	1.29	.4015	.0985
.58	.2190	.2810	.94	.3264	.1736	1.30	.4032	.0968
.59	.2224	.2776	.95	.3289	.1711	1.31	.4049	.0951
.60	.2257	.2743	.96	.3315	.1685	1.32	.4066	.0934
.61	.2291	.2709	.97	.3340	.1660	1.33	.4082	.0918
.62	.2324	.2676	.98	.3365	.1635	1.34	.4099	.0901
.63	.2357	.2643	.99	.3389	.1611	1.35	.4115	.0885
.64	.2389	.2611	1.00	.3413	.1587	1.36	.4131	.0869
.65	.2422	.2578	1.01	.3438	.1562	1.37	.4147	.0853

(*Continued*)

Table A.1 (Continued)

(A)	(B)	(C)	(A)	(B)	(C)	(A)	(B)	(C)
z	Area between mean and *z*	Area beyond *z*	*z*	Area between mean and *z*	Area beyond *z*	*z*	Area between mean and *z*	Area beyond *z*
1.38	.4162	.0838	1.73	.4582	.0418	2.08	.4812	.0188
1.39	.4177	.0823	1.74	.4591	.0409	2.09	.4817	.0183
1.40	.4192	.0808	1.75	.4599	.0401	2.10	.4821	.0179
1.41	.4207	.0793	1.76	.4608	.0392	2.11	.4826	.0174
1.42	.4222	.0778	1.77	.4616	.0384	2.12	.4830	.0170
1.43	.4236	.0764	1.78	.4625	.0375	2.13	.4834	.0166
1.44	.4251	.0749	1.79	.4633	.0367	2.14	.4838	.0162
1.45	.4265	.0735	1.80	.4641	.0359	2.15	.4842	.0158
1.46	.4279	.0721	1.81	.4649	.0351	2.16	.4846	.0154
1.47	.4292	.0708	1.82	.4656	.0344	2.17	.4850	.0150
1.48	.4306	.0694	1.83	.4664	.0336	2.18	.4854	.0146
1.49	.4319	.0681	1.84	.4671	.0329	2.19	.4857	.0143
1.50	.4332	.0668	1.85	.4678	.0322	2.20	.4861	.0139
1.51	.4345	.0655	1.86	.4686	.0314	2.21	.4864	.0136
1.52	.4357	.0643	1.87	.4693	.0307	2.22	.4868	.0132
1.53	.4370	.0630	1.88	.4699	.0301	2.23	.4871	.0129
1.54	.4382	.0618	1.89	.4706	.0294	2.24	.4875	.0125
1.55	.4394	.0606	1.90	.4713	.0287	2.25	.4878	.0122
1.56	.4406	.0594	1.91	.4719	.0281	2.26	.4881	.0119
1.57	.4418	.0582	1.92	.4726	.0274	2.27	.4884	.0116
1.58	.4429	.0571	1.93	.4732	.0268	2.28	.4887	.0113
1.59	.4441	.0559	1.94	.4738	.0262	2.29	.4890	.0110
1.60	.4452	.0548	1.95	.4744	.0256	2.30	.4893	.0107
1.61	.4463	.0537	1.96	.4750	.0250	2.31	.4896	.0104
1.62	.4474	.0526	1.97	.4756	.0244	2.32	.4898	.0102
1.63	.4484	.0516	1.98	.4761	.0239	2.33	.4901	.0099
1.64	.4495	.0505	1.99	.4767	.0233	2.34	.4904	.0096
1.65	.4505	.0495	2.00	.4772	.0228	2.35	.4906	.0094
1.66	.4515	.0485	2.01	.4778	.0222	2.36	.4909	.0091
1.67	.4525	.0475	2.02	.4783	.0217	2.37	.4911	.0089
1.68	.4535	.0465	2.03	.4788	.0212	2.38	.4913	.0087
1.69	.4545	.0455	2.04	.4793	.0207	2.39	.4916	.0084
1.70	.4554	.0446	2.05	.4798	.0202	2.40	.4918	.0082
1.71	.4564	.0436	2.06	.4803	.0197	2.41	.4920	.0080
1.72	.4573	.0427	2.07	.4808	.0192	2.42	.4922	.0078

(*Continued*)

Table A.1 (Continued)

(A)	(B)	(C)	(A)	(B)	(C)	(A)	(B)	(C)
z	Area between mean and *z*	Area beyond *z*	*z*	Area between mean and *z*	Area beyond *z*	*z*	Area between mean and *z*	Area beyond *z*
2.43	.4925	.0075	2.73	.4968	.0032	3.03	.4988	.0012
2.44	.4927	.0073	2.74	.4969	.0031	3.04	.4988	.0012
2.45	.4929	.0071	2.75	.4970	.0030	3.05	.4989	.0011
2.46	.4931	.0069	2.76	.4971	.0029	3.06	.4989	.0011
2.47	.4932	.0068	2.77	.4972	.0028	3.07	.4989	.0011
2.48	.4934	.0066	2.78	.4973	.0027	3.08	.4990	.0010
2.49	.4936	.0064	2.79	.4974	.0026	3.09	.4990	.0010
2.50	.4938	.0062	2.80	.4974	.0026	3.10	.4990	.0010
2.51	.4940	.0060	2.81	.4975	.0025	3.11	.4991	.0009
2.52	.4941	.0059	2.82	.4976	.0024	3.12	.4991	.0009
2.53	.4943	.0057	2.83	.4977	.0023	3.13	.4991	.0009
2.54	.4945	.0055	2.84	.4977	.0023	3.14	.4992	.0008
2.55	.4946	.0054	2.85	.4978	.0022	3.15	.4992	.0008
2.56	.4948	.0052	2.86	.4979	.0021	3.16	.4992	.0008
2.57	.4949	.0051	2.87	.4979	.0021	3.17	.4992	.0008
2.58	.4951	.0049	2.88	.4980	.0020	3.18	.4993	.0007
2.59	.4952	.0048	2.89	.4981	.0019	3.19	.4993	.0007
2.60	.4953	.0047	2.90	.4981	.0019	3.20	.4993	.0007
2.61	.4955	.0045	2.91	.4982	.0018	3.21	.4993	.0007
2.62	.4956	.0044	2.92	.4982	.0018	3.22	.4994	.0006
2.63	.4957	.0043	2.93	.4983	.0017	3.23	.4994	.0006
2.64	.4959	.0041	2.94	.4984	.0016	3.24	.4994	.0006
2.65	.4960	.0040	2.95	.4984	.0016	3.30	.4995	.0005
2.66	.4961	.0039	2.96	.4985	.0015	3.40	.4997	.0003
2.67	.4962	.0038	2.97	.4985	.0015	3.50	.4998	.0002
2.68	.4963	.0037	2.98	.4986	.0014	3.60	.4998	.0002
2.69	.4964	.0036	2.99	.4986	.0014	3.70	.4999	.0001
2.70	.4965	.0035	3.00	.4987	.0013	3.80	.49993	.00007
2.71	.4966	.0034	3.01	.4987	.0013	3.90	.49995	.00005
2.72	.4967	.0033	3.02	.4987	.0013	4.00	.49997	.00003

Column *A* lists the *z* score values. Column *B* provides the proportion of area between the mean and the *z* score value. Column *C* provides the proportion of area beyond the *z* score.

Note: Because the normal distribution is symmetrical, areas for negative *z* scores are the same as those for positive *z* scores.

Table A.2 *t* Table.

		α values for two-tailed test				
	.20	*.10*	*.05*	*.02*	*.01*	*.001*
		α values for one-tailed test				
df	*.10*	*.05*	*.025*	*.01*	*.005*	*.0005*
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Source: Table III of Fisher and Yates': *Statistical Tables for Biological, Agricultural and Medical Research*, published by Longman Group UK, London (previously published by Oliver & Boyd Ltd, Edinburgh) and by permission of the authors and publishers.

To be significant the *t* obtained from the data must be equal to or larger than the value shown in the table.

Table A.3 Power table (finding power).

	One-tailed test *(α)*					One-tailed test *(α)*			
	.05	*.025*	*.01*	*.005*		*.05*	*.025*	*.01*	*.005*
	Two-tailed test *(α)*					Two-tailed test *(α)*			
δ	*.10*	*.05*	*.02*	*.01*	*δ*	*.10*	*.05*	*.02*	*.01*
.0	.10*	.05*	.02	.01	2.5	.80	.71	.57	.47
.1	.10*	.05*	.02	.01	2.6	.83	.74	.61	.51
.2	.11*	.05	.02	.01	2.7	.85	.77	.65	.55
.3	.12*	.06	.03	.01	2.8	.88	.80	.68	.59
.4	.13*	.07	.03	.01	2.9	.90	.83	.72	.63
.5	.14	.08	.03	.02	3.0	.91	.85	.75	.66
.6	.16	.09	.04	.02	3.1	.93	.87	.78	.70
.7	.18	.11	.05	.03	3.2	.94	.89	.81	.73
.8	.21	.13	.06	.04	3.3	.96	.91	.83	.77
.9	.23	.15	.08	.05	3.4	.96	.93	.86	.80
1.0	.26	.17	.09	.06	3.5	.97	.94	.88	.82
1.1	.30	.20	.11	.07	3.6	.97	.95	.90	.85
1.2	.33	.22	.13	.08	3.7	.98	.96	.92	.87
1.3	.37	.26	.15	.10	3.8	.98	.97	.93	.89
1.4	.40	.29	.18	.12	3.9	.99	.97	.94	.91
1.5	.44	.32	.20	.14	4.0	.99	.98	.95	.92
1.6	.48	.36	.23	.16	4.1	.99	.98	.96	.94
1.7	.52	.40	.27	.19	4.2	.99	.99	.97	.95
1.8	.56	.44	.30	.22	4.3	**	.99	.98	.96
1.9	.60	.48	.33	.25	4.4		.99	.98	.97
2.0	.64	.52	.37	.28	4.5		.99	.99	.97
2.1	.68	.56	.41	.32	4.6		**	.99	.98
2.2	.71	.59	.45	.35	4.7			.99	.98
2.3	.74	.63	.49	.39	4.8			.99	.99
2.4	.77	.67	.53	.43	4.9			.99	.99
					5.0			**	.99
					5.1				.99
					5.2				**

* Values inaccurate for *one-tailed* test by more than .01.
** The power at and below this point is greater than .995.

Table A.4 Power table (finding delta).

	One-tailed test *(α)*			
	.05	***.025***	***.01***	***.005***
	Two-tailed test *(α)*			
Power	***.10***	***.05***	***.02***	***.01***
.25	.97	1.29	1.65	1.90
.50	1.64	1.96	2.33	2.58
.60	1.90	2.21	2.58	2.83
.67	2.08	2.39	2.76	3.01
.70	2.17	2.48	2.85	3.10
.75	2.32	2.63	3.00	3.25
.80	2.49	2.80	3.17	3.42
.85	2.68	3.00	3.36	3.61
.90	2.93	3.24	3.61	3.86
.95	3.29	3.60	3.97	4.22
.99	3.97	4.29	4.65	4.90
.999	4.37	5.05	5.42	5.67

Tables A.3 and A.4 from *Introductory Statistics for the Behavioral Sciences*, 3rd ed., by J. Welkowitz, R Ewen and J. Cohen, Copyright © 1982 by Harcourt Brace Jovanovich, Inc., reprinted by permission of the publisher.

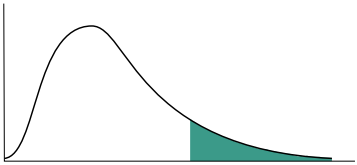


Table A.5 *F* table.

	Degrees of freedom: numerator														
***df*: denominator**	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***14***	***16***	***20***
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248
	4052	**4999**	**5403**	**5625**	**5764**	**5859**	**5928**	**5981**	**6022**	**6056**	**6082**	**6106**	**6142**	**6169**	**6208**
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44
	98.49	**99.00**	**99.17**	**99.25**	**99.30**	**99.33**	**99.34**	**99.36**	**99.38**	**99.40**	**99.41**	**99.42**	**99.43**	**99.44**	**99.45**
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66
	34.12	**30.82**	**29.46**	**28.71**	**28.24**	**27.91**	**27.67**	**27.49**	**27.34**	**27.23**	**27.13**	**27.05**	**26.92**	**26.83**	**26.69**
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80
	21.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.98**	**14.80**	**14.66**	**14.54**	**14.45**	**14.37**	**14.24**	**14.15**	**14.02**
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56
	16.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.45**	**10.27**	**10.15**	**10.05**	**9.96**	**9.89**	**9.77**	**9.68**	**9.55**
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87
	13.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.26**	**8.10**	**7.98**	**7.87**	**7.79**	**7.72**	**7.60**	**7.52**	**7.39**
7	5.59	4.47	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**7.00**	**6.84**	**6.71**	**6.62**	**6.54**	**6.47**	**6.35**	**6.27**	**6.15**

8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.19**	**6.03**	**5.91**	**5.82**	**5.74**	**5.67**	**5.56**	**5.48**	**5.36**
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.62**	**5.47**	**5.35**	**5.26**	**5.18**	**5.11**	**5.00**	**4.92**	**4.80**
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**	**4.78**	**4.71**	**4.60**	**4.52**	**4.41**
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**	**4.46**	**4.40**	**4.29**	**4.21**	**4.10**
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**	**4.22**	**4.16**	**4.05**	**3.98**	**3.86**
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.44**	**4.30**	**4.19**	**4.10**	**4.02**	**3.96**	**3.85**	**3.78**	**3.67**
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**	**3.86**	**3.80**	**3.70**	**3.62**	**3.51**
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**	**3.73**	**3.67**	**3.56**	**3.48**	**3.36**
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28
	8.53	**6.23**	**5.29**	**4.77**	**4.44**	**4.20**	**4.03**	**3.89**	**3.78**	**3.69**	**3.61**	**3.55**	**3.45**	**3.37**	**3.25**
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.93**	**3.79**	**3.68**	**3.59**	**3.52**	**3.45**	**3.35**	**3.27**	**3.16**
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.85**	**3.71**	**3.60**	**3.51**	**3.44**	**3.37**	**3.27**	**3.19**	**3.07**
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.77**	**3.63**	**3.52**	**3.43**	**3.36**	**3.30**	**3.19**	**3.12**	**3.00**

(*Continued*)

Table A.5 (Continued)

	Degrees of freedom: numerator														
***df*: denominator**	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***14***	***16***	***20***
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.71**	**3.56**	**3.45**	**3.37**	**3.30**	**3.23**	**3.13**	**3.05**	**2.94**
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09
	8.02	**5.78**	**4.87**	**4.37**	**4.04**	**3.81**	**3.65**	**3.51**	**3.40**	**3.31**	**3.24**	**3.17**	**3.07**	**2.99**	**2.88**
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18	2.13	2.07
	7.94	**5.72**	**4.82**	**4.31**	**3.99**	**3.76**	**3.59**	**3.45**	**3.35**	**3.26**	**3.18**	**3.12**	**3.02**	**2.94**	**2.83**
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14	2.10	2.04
	7.88	**5.66**	**4.76**	**4.26**	**3.94**	**3.71**	**3.54**	**3.41**	**3.30**	**3.21**	**3.14**	**3.07**	**2.97**	**2.89**	**2.78**
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.50**	**3.36**	**3.25**	**3.17**	**3.09**	**3.03**	**2.93**	**2.85**	**2.74**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.46**	**3.32**	**3.21**	**3.13**	**3.05**	**2.99**	**2.89**	**2.81**	**2.70**
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.42**	**3.29**	**3.17**	**3.09**	**3.02**	**2.96**	**2.86**	**2.77**	**2.66**
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97
	7.68	**5.49**	**4.60**	**4.11**	**3.79**	**3.56**	**3.39**	**3.26**	**3.14**	**3.06**	**2.98**	**2.93**	**2.83**	**2.74**	**2.63**
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96
	7.64	**5.45**	**4.57**	**4.07**	**3.76**	**3.53**	**3.36**	**3.23**	**3.11**	**3.03**	**2.95**	**2.90**	**2.80**	**2.71**	**2.60**
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.33**	**3.20**	**3.08**	**3.00**	**2.92**	**2.87**	**2.77**	**2.68**	**2.57**
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.30**	**3.17**	**3.06**	**2.98**	**2.90**	**2.84**	**2.74**	**2.66**	**2.55**

32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.91
	7.50	**5.34**	**4.46**	**3.97**	**3.66**	**3.42**	**3.25**	**3.12**	**3.01**	**2.94**	**2.86**	**2.80**	**2.70**	**2.62**	**2.51**
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.89
	7.44	**5.29**	**4.42**	**3.93**	**3.61**	**3.38**	**3.21**	**3.08**	**2.97**	**2.89**	**2.82**	**2.76**	**2.66**	**2.58**	**2.47**
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98	1.93	1.87
	7.39	**5.25**	**4.38**	**3.89**	**3.58**	**3.35**	**3.18**	**3.04**	**2.94**	**2.86**	**2.78**	**2.72**	**2.62**	**2.54**	**2.43**
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85
	7.35	**5.21**	**4.34**	**3.86**	**3.54**	**3.32**	**3.15**	**3.02**	**2.91**	**2.82**	**2.75**	**2.69**	**2.59**	**2.51**	**2.40**
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.84
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**3.12**	**2.99**	**2.88**	**2.80**	**2.73**	**2.66**	**2.56**	**2.49**	**2.37**
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82
	7.27	**5.15**	**4.29**	**3.80**	**3.49**	**3.26**	**3.10**	**2.96**	**2.86**	**2.77**	**2.70**	**2.64**	**2.54**	**2.46**	**2.35**
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92	1.88	1.81
	7.24	**5.12**	**4.26**	**3.78**	**3.46**	**3.24**	**3.07**	**2.94**	**2.84**	**2.75**	**2.68**	**2.62**	**2.52**	**2.44**	**2.32**
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.80
	7.21	**5.10**	**4.24**	**3.76**	**3.44**	**3.22**	**3.05**	**2.92**	**2.82**	**2.73**	**2.66**	**2.60**	**2.50**	**2.42**	**2.30**
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96	1.90	1.86	1.79
	7.19	**5.08**	**4.22**	**3.74**	**3.42**	**3.20**	**3.04**	**2.90**	**2.80**	**2.71**	**2.64**	**2.58**	**2.48**	**2.40**	**2.28**
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90	1.85	1.78
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.18**	**3.02**	**2.88**	**2.78**	**2.70**	**2.62**	**2.56**	**2.46**	**2.39**	**2.26**
55	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93	1.88	1.83	1.76
	7.12	**5.01**	**4.16**	**3.68**	**3.37**	**3.15**	**2.98**	**2.85**	**2.75**	**2.66**	**2.59**	**2.53**	**2.43**	**2.35**	**2.23**
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86	1.81	1.75
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.95**	**2.82**	**2.72**	**2.63**	**2.56**	**2.50**	**2.40**	**2.32**	**2.20**

(*Continued*)

Table A.5 (Continued)

	Degrees of freedom: numerator														
***df*: denominator**	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*14*	*16*	*20*
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90	1.85	1.80	1.73
	7.04	**4.95**	**4.10**	**3.62**	**3.31**	**3.09**	**2.93**	**2.79**	**2.70**	**2.61**	**2.54**	**2.47**	**2.37**	**2.30**	**2.18**
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72
	7.01	**4.92**	**4.08**	**3.60**	**3.29**	**3.07**	**2.91**	**2.77**	**2.67**	**2.59**	**2.51**	**2.45**	**2.35**	**2.28**	**2.15**
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82	1.77	1.70
	6.96	**4.88**	**4.04**	**3.56**	**3.25**	**3.04**	**2.87**	**2.74**	**2.64**	**2.55**	**2.48**	**2.41**	**2.32**	**2.24**	**2.11**
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68
	6.90	**4.82**	**3.98**	**3.51**	**3.20**	**2.99**	**2.82**	**2.69**	**2.59**	**2.51**	**2.43**	**2.36**	**2.26**	**2.19**	**2.06**
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83	1.77	1.72	1.65
	6.84	**4.78**	**3.94**	**3.47**	**3.17**	**2.95**	**2.79**	**2.65**	**2.56**	**2.47**	**2.40**	**2.33**	**2.23**	**2.15**	**2.03**
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.76	1.71	1.64
	6.81	**4.75**	**3.91**	**3.44**	**3.14**	**2.92**	**2.76**	**2.62**	**2.53**	**2.44**	**2.37**	**2.30**	**2.20**	**2.12**	**2.00**
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80	1.74	1.69	1.62
	6.76	**4.71**	**3.88**	**3.41**	**3.11**	**2.90**	**2.73**	**2.60**	**2.50**	**2.41**	**2.34**	**2.28**	**2.17**	**2.09**	**1.97**
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78	1.72	1.67	1.60
	6.70	**4.66**	**3.83**	**3.36**	**3.06**	**2.85**	**2.69**	**2.55**	**2.46**	**2.37**	**2.29**	**2.23**	**2.12**	**2.04**	**1.92**
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76	1.70	1.65	1.58
	6.66	**4.62**	**3.80**	**3.34**	**3.04**	**2.82**	**2.66**	**2.53**	**2.43**	**2.34**	**2.26**	**2.20**	**2.09**	**2.01**	**1.89**
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57
	6.64	**4.60**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**	**2.24**	**2.18**	**2.07**	**1.99**	**1.87**

Source: Reproduced by permission from *Statistical Methods*, 8th ed., by G. W. Snedecor and W. G. Cochran. © 1956 by The Iowa State University Press.
Table entries in lightface type are critical values for the .05 level of significance. Boldface type values are for the .01 level of significance.

Table A.6 The critical values for studentized range statistic (q), $\alpha = .05$.

	k = number of treatments								
***df*: error term**	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99
	5.70	**6.98**	**7.80**	**8.42**	**8.91**	**9.32**	**9.67**	**9.97**	**10.24**
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49
	5.24	**6.33**	**7.03**	**7.56**	**7.97**	**8.32**	**8.61**	**8.87**	**9.10**
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16
	4.95	**5.92**	**6.54**	**7.01**	**7.37**	**7.68**	**7.94**	**8.17**	**8.37**
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92
	4.75	**5.64**	**6.20**	**6.63**	**6.96**	**7.24**	**7.47**	**7.68**	**7.86**
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74
	4.56	**5.43**	**5.96**	**6.35**	**6.66**	**6.91**	**7.13**	**7.33**	**7.51**
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60
	4.48	**5.27**	**5.77**	**6.14**	**6.43**	**6.67**	**6.88**	**7.05**	**7.21**
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49
	4.39	**5.15**	**5.62**	**5.97**	**6.25**	**6.48**	**6.67**	**6.84**	**6.99**
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39
	4.32	**5.05**	**5.50**	**5.84**	**6.10**	**6.32**	**6.51**	**6.67**	**6.81**
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32
	4.26	**4.96**	**5.40**	**5.73**	**5.98**	**6.19**	**6.37**	**6.53**	**6.67**
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25
	4.21	**4.89**	**5.32**	**5.63**	**5.88**	**6.08**	**6.26**	**6.41**	**6.54**
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20
	4.17	**4.84**	**5.25**	**5.56**	**5.80**	**5.99**	**6.16**	**6.31**	**6.44**
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15
	4.13	**4.79**	**5.19**	**5.49**	**5.72**	**5.92**	**6.08**	**6.22**	**6.35**
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11
	4.10	**4.74**	**5.14**	**5.43**	**5.66**	**5.85**	**6.01**	**6.15**	**6.27**
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07
	4.07	**4.70**	**5.09**	**5.38**	**5.60**	**5.79**	**5.94**	**6.08**	**6.20**
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04
	4.05	**4.67**	**5.05**	**5.33**	**5.55**	**5.73**	**5.89**	**6.02**	**6.14**
20	2.95	3.58	3.96	4.23	4.44	4.62	4.77	4.90	5.01
	4.02	**4.64**	**5.02**	**5.29**	**5.51**	**5.69**	**5.84**	**5.97**	**6.09**

(*Continued*)

Table A.6 (Continued)

	k = number of treatments								
***df*: error term**	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.91
	3.96	**4.55**	**4.91**	**5.17**	**5.37**	**5.54**	**5.69**	**5.81**	**5.92**
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82
	3.89	**4.45**	**4.80**	**5.05**	**5.24**	**5.40**	**5.54**	**5.65**	**5.76**
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73
	3.82	**4.37**	**4.70**	**4.93**	**5.11**	**5.26**	**5.39**	**5.50**	**5.60**
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65
	3.76	**4.28**	**4.59**	**4.82**	**4.99**	**5.13**	**5.25**	**5.36**	**5.45**
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56
	3.70	**4.20**	**4.50**	**4.71**	**4.87**	**5.01**	**5.12**	**5.21**	**5.30**
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47
	3.64	**4.12**	**4.40**	**4.60**	**4.76**	**4.88**	**4.99**	**5.08**	**5.16**

Table entries in lightface type are critical values for the .05 level of significance. Boldface type values are for the .01 level of significance.

Table A.7 Pearson *r* table.

	α levels for two-tailed test				
	.10	***.05***	***.02***	***.01***	***.001***
	α levels for one-tailed test				
(df = n_p − 2)	***.05***	***.025***	***.01***	***.005***	***.0005***
1	.98769	.99692	.999507	.999877	.9999988
2	.90000	.95000	.98000	.990000	.99900
3	.8054	.8783	.93433	.95873	.99116
4	.7293	.8114	.8822	.91720	.97406
5	.6694	.7545	.8329	.8745	.95074
6	.6215	.7067	.7887	.8343	.92493
7	.5822	.6664	.7498	.7977	.8982
8	.5494	.6319	.7155	.7646	.8721

Table A.7 (Continued)

	α levels for two-tailed test				
	.10	*.05*	*.02*	*.01*	*.001*
	α levels for one-tailed test				
(df = n_p − 2)	*.05*	*.025*	*.01*	*.005*	*.0005*
9	.5214	.6021	.6851	.7348	.8371
10	.4973	.5760	.6581	.7079	.8233
11	.4762	.5529	.6339	.6835	.8010
12	.4575	.5324	.6120	.6614	.7800
13	.4409	.5139	.5923	.6411	.7603
14	.4259	.4973	.5742	.6226	.7420
15	.4124	.4821	.5577	.6055	.7246
16	.4000	.4683	.5425	.5897	.7084
17	.3887	.4555	.5285	.5751	.6932
18	.3783	.4438	.5155	.5614	.6787
19	.3687	.4329	.5034	.5487	.6652
20	.3598	.4227	.4921	.5368	.6524
25	.3233	.3809	.4451	.4869	.5974
30	.2960	.3494	.4093	.4487	.5541
35	.2746	.3246	.3810	.4182	.5189
40	.2573	.3044	.3578	.3932	.4896
45	.2428	.2875	.3384	.3721	.4648
50	.2306	.2732	.3218	.3541	.4433
60	.2108	.2500	.2948	.3248	.4078
70	.1954	.2319	.2737	.3017	.3799
80	.1829	.2172	.2565	.2830	.3568
90	.1726	.2050	.2422	.2673	.3375
100	.1638	.1946	.2301	.2540	.3211

Source: Table VII of Fisher and Yates': *Statistical Tables for Biological, Agricultural and Medical Research* published by Longman Group UK, London (previously published by Oliver and Boyd Ltd., Edinburgh) and by permission of the authors and publishers.
To be significant the *r* obtained from the data must be equal to or larger than the value shown in the table.

Table A.8 Chi-square table.

	α Levels				
df	***.10***	***.05***	***.02***	***.01***	***.001***
1	2.71	3.84	5.41	6.64	10.83
2	4.60	5.99	7.82	9.21	13.82
3	6.25	7.82	9.84	11.34	16.27
4	7.78	9.49	11.67	13.28	18.46
5	9.24	11.07	13.39	15.09	20.52
6	10.64	12.59	15.03	16.81	22.46
7	12.02	14.07	16.62	18.48	24.32
8	13.36	15.51	18.17	20.09	26.12
9	14.68	16.92	19.68	21.67	27.88
10	15.99	18.31	21.16	23.21	29.59
11	17.28	19.68	22.62	24.72	31.26
12	18.55	21.03	24.05	26.22	32.91
13	19.81	22.36	25.47	27.69	34.53
14	21.06	23.68	26.87	29.14	36.12
15	22.31	25.00	28.26	30.58	37.70
16	23.54	26.30	29.63	32.00	39.25
17	24.77	27.59	31.00	33.41	40.79
18	25.99	28.87	32.35	34.80	42.31
19	27.20	30.14	33.69	36.19	43.82
20	28.41	31.41	35.02	37.57	45.32
21	29.62	32.67	36.34	38.93	46.80
22	30.81	33.92	37.66	40.29	48.27
23	32.01	35.17	38.97	41.64	49.73
24	33.20	36.42	40.27	42.98	51.18
25	34.38	37.65	41.57	44.31	52.62
26	35.56	38.88	42.86	45.64	54.05
27	36.74	40.11	44.14	46.96	55.48
28	37.92	41.34	45.42	48.28	56.89
29	39.09	42.56	46.69	49.59	58.30
30	40.26	43.77	47.96	50.89	59.70

Source: Table IV of Fisher and Yates': *Statistical Tables for Biological Agricultural and Medical Research*, published by Longman Group UK, London (previously published by Oliver and Boyd Ltd., Edinburgh) and by permission of the authors and publishers.
To be significant the χ^2 obtained from the data must be equal to or larger than the value shown in the table.

Table A.9 Spearman r_s table.

Number of pairs, n_p	Level of significance for a one-tailed test			
	.05	*.025*	*.01*	*.005*
	Level of significance for a two-tailed test			
	.10	*.05*	*.02*	*.01*
5	.900	1.000	1.000	
6	.829	.886	.943	1.000
7	.714	.786	.893	.929
8	.643	.738	.833	.881
9	.600	.700	.783	.833
10	.564	.648	.745	.794
11	.536	.618	.709	.755
12	.503	.587	.671	.727
13	.484	.560	.648	.703
14	.464	.538	.622	.675
15	.443	.521	.604	.654
16	.429	.503	.582	.635
17	.414	.485	.566	.615
18	.401	.472	.550	.600
19	.391	.460	.535	.584
20	.380	.447	.520	.570
21	.370	.435	.508	.556
22	.361	.425	.496	.544
23	.353	.415	.486	.532
24	.344	.406	.476	.521
25	.337	.398	.466	.511
26	.331	.390	.457	.501
27	.324	.382	.448	.491
28	.317	.375	.440	.483
29	.312	.368	.433	.475
30	.306	.362	.425	.467
32	.296	.350	.412	.452

(*Continued*)

Table A.9 (Continued)

Number of pairs, n_p	Level of significance for a one-tailed test			
	.05	*.025*	*.01*	*.005*
	Level of significance for a two-tailed test			
	.10	*.05*	*.02*	*.01*
34	.287	.340	.399	.439
36	.279	.330	.388	.427
38	.271	.321	.378	.415
40	.264	.313	.368	.405
42	.257	.305	.359	.395
44	.251	.298	.351	.386
46	.246	.291	.343	.378
48	.240	.285	.336	.370
50	.235	.279	.329	.363
52	.231	.274	.323	.356
54	.226	.268	.317	.349
56	.222	.264	.311	.343
58	.218	.259	.306	.337
60	.214	.255	.300	.331
70	.198	.235	.278	.307
80	.185	.220	.260	.287
90	.174	.207	.245	.271
100	.165	.197	.233	.257

If obtained value of r_s is equal to or greater than tabled value for the appropriate alpha, reject H_0.
Glasser, G. J., & Winter, R. F. (1961). "Critical values of the coefficient of rank correlation for Testing the hypothesis of independence," *Biometrika*, 48, 444. Reprinted by permission of the Biometrika Trustees.

Table A.10 Mann–Whitney *U* table, critical values for a one-tailed test at $\alpha = .01$ (roman type) and $\alpha = .005$ (boldface type) and for a two-tailed test at $\alpha = .02$ (roman type) and $\alpha = .01$ (boldface type).[a]

	n_A																			
n_B	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—[b]	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	—	—	—	—	—	0	0	0	0	0	0	1	1
													—	—	—	—	—	—	**0**	**0**
3	—	—	—	—	—	—	0	0	1	1	1	2	2	2	3	3	4	4	4	5
							—	—	**0**	**0**	**0**	**1**	**1**	**1**	**2**	**2**	**2**	**2**	**3**	**3**
4	—	—	—	—	0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10
					—	**0**	**0**	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**5**	**5**	**6**	**6**	**7**	**8**
5	—	—	—	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
				—	**0**	**1**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
6	—	—	—	1	2	3	4	6	7	8	9	11	12	13	15	16	18	19	20	22
				0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**9**	**10**	**11**	**12**	**13**	**15**	**16**	**17**	**18**
7	—	—	0	1	3	4	6	7	9	11	12	14	16	17	19	21	23	24	26	28
			—	**0**	**1**	**3**	**4**	**6**	**7**	**9**	**10**	**12**	**13**	**15**	**16**	**18**	**19**	**21**	**22**	**24**
8	—	—	0	2	4	6	7	9	11	13	15	17	20	22	24	26	28	30	32	34
			—	**1**	**2**	**4**	**6**	**7**	**9**	**11**	**13**	**15**	**17**	**18**	**20**	**22**	**24**	**26**	**28**	**30**
9	—	—	1	3	5	7	9	11	14	16	18	21	23	26	28	31	33	36	38	40
			0	**1**	**3**	**5**	**7**	**9**	**11**	**13**	**16**	**18**	**20**	**22**	**24**	**27**	**29**	**31**	**33**	**36**
10	—	—	1	3	6	8	11	13	16	19	22	24	27	30	33	36	38	41	44	47
			0	**2**	**4**	**6**	**9**	**11**	**13**	**16**	**18**	**21**	**24**	**26**	**29**	**31**	**34**	**37**	**39**	**42**
11	—	—	1	4	7	9	12	15	18	22	25	28	31	34	37	41	44	47	50	53
			0	**2**	**5**	**7**	**10**	**13**	**16**	**18**	**21**	**24**	**27**	**30**	**33**	**36**	**39**	**42**	**45**	**48**
12	—	—	2	5	8	11	14	17	21	24	28	31	35	38	42	46	49	53	56	60
			1	**3**	**6**	**9**	**12**	**15**	**18**	**21**	**24**	**27**	**31**	**34**	**37**	**41**	**44**	**47**	**51**	**54**
13	—	0	2	5	9	12	16	20	23	27	31	35	39	43	47	51	55	59	63	67
		—	**1**	**3**	**7**	**10**	**13**	**17**	**20**	**24**	**27**	**31**	**34**	**38**	**42**	**45**	**49**	**53**	**56**	**60**
14	—	0	2	6	10	13	17	22	26	30	34	38	43	47	51	56	60	65	69	73
		—	**1**	**4**	**7**	**11**	**15**	**18**	**22**	**26**	**30**	**34**	**38**	**42**	**46**	**50**	**54**	**58**	**63**	**67**
15	—	0	3	7	11	15	19	24	28	33	37	42	47	51	56	61	66	70	75	80
		—	**2**	**5**	**8**	**12**	**16**	**20**	**24**	**29**	**33**	**37**	**42**	**46**	**51**	**55**	**60**	**64**	**69**	**73**
16	—	0	3	7	12	16	21	26	31	36	41	46	51	56	61	66	71	76	82	87
		—	**2**	**5**	**9**	**13**	**18**	**22**	**27**	**31**	**36**	**41**	**45**	**50**	**55**	**60**	**65**	**70**	**74**	**79**
17	—	0	4	8	13	18	23	28	33	38	44	49	55	60	66	71	77	82	88	93
		—	**2**	**6**	**10**	**15**	**19**	**24**	**29**	**34**	**39**	**44**	**49**	**54**	**60**	**65**	**70**	**75**	**81**	**86**
18	—	0	4	9	14	19	24	30	36	41	47	53	59	65	70	76	82	88	94	100
		—	**2**	**6**	**11**	**16**	**21**	**26**	**31**	**37**	**42**	**47**	**53**	**58**	**64**	**70**	**75**	**81**	**87**	**92**
19	—	1	4	9	15	20	26	32	38	44	50	56	63	69	75	82	88	94	101	107
		0	**3**	**7**	**12**	**17**	**22**	**28**	**33**	**39**	**45**	**51**	**56**	**63**	**69**	**74**	**81**	**87**	**93**	**99**
20	—	1	5	10	16	22	28	34	40	47	53	60	67	73	80	87	93	100	107	114
		0	**3**	**8**	**13**	**18**	**24**	**30**	**36**	**42**	**48**	**54**	**60**	**67**	**73**	**79**	**86**	**92**	**99**	**105**

[a] To be significant for any given n_A, n_B, the obtained *U* must be *equal to* or *less than* the value shown in table.
[b] Dashes in the body of the table indicate that no decision is possible at the stated level of significance.
Source: Table B.9a and B.9b are from *Statistics: An Introduction*, 3rd. ed., R Kirk © 1990 Holt, Rinehart and Winston, Inc., reprinted by permission of the publisher.

Table A.11 Mann–Whitney *U* table, critical values for a one-tailed test at $\alpha = .05$ (roman type) and $\alpha = .025$ (boldface type) and for a two-tailed test at $\alpha = .10$ (roman type) and $\alpha = .05$ (boldface type).

	n_A																			
n_B	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0	0
																			—	—
2	—	—	—	—	0	0	0	1	1	1	1	2	2	2	3	3	3	4	4	4
					—	—	—	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**2**
3	—	—	0	0	1	2	2	3	3	4	5	5	6	7	7	8	9	9	10	11
	—	—	—	—	**0**	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**	**5**	**5**	**6**	**6**	**7**	**7**	**8**
4	—	—	0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
	—	—	—	**0**	**1**	**2**	**3**	**4**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**11**	**12**	**13**	**13**
5	—	0	1	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
	—	—	**0**	**1**	**2**	**3**	**5**	**6**	**7**	**8**	**9**	**11**	**12**	**13**	**14**	**15**	**17**	**18**	**19**	**20**
6	—	0	2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
	—	—	**1**	**2**	**3**	**5**	**6**	**8**	**10**	**11**	**13**	**14**	**16**	**17**	**19**	**21**	**22**	**24**	**25**	**27**
7	—	0	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
	—	—	**1**	**3**	**5**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**	**22**	**24**	**26**	**28**	**30**	**32**	**34**
8	—	1	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
	—	0	**2**	**4**	**6**	**8**	**10**	**13**	**15**	**17**	**19**	**22**	**24**	**26**	**29**	**31**	**34**	**36**	**38**	**41**
9	—	1	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
	—	0	**2**	**4**	**7**	**10**	**12**	**15**	**17**	**20**	**23**	**26**	**28**	**31**	**34**	**37**	**39**	**42**	**45**	**48**
10	—	1	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58	62
	—	0	**3**	**5**	**8**	**11**	**14**	**17**	**20**	**23**	**26**	**29**	**33**	**36**	**39**	**42**	**45**	**48**	**52**	**55**
11	—	1	5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65	69
	—	0	**3**	**6**	**9**	**13**	**16**	**19**	**23**	**26**	**30**	**33**	**37**	**40**	**44**	**47**	**51**	**55**	**58**	**62**
12	—	2	5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72	77
	—	1	**4**	**7**	**11**	**14**	**18**	**22**	**26**	**29**	**33**	**37**	**41**	**45**	**49**	**53**	**57**	**61**	**65**	**69**
13	—	2	6	10	15	19	24	28	33	37	42	47	51	56	61	65	70	75	80	84
	—	1	**4**	**8**	**12**	**16**	**20**	**24**	**28**	**33**	**37**	**41**	**45**	**50**	**54**	**59**	**63**	**67**	**72**	**76**
14	—	2	7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87	92
	—	1	**5**	**9**	**13**	**17**	**22**	**26**	**31**	**36**	**40**	**45**	**50**	**55**	**59**	**64**	**67**	**74**	**78**	**83**
15	—	3	7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	94	100
	—	**1**	**5**	**10**	**14**	**19**	**24**	**29**	**34**	**39**	**44**	**49**	**54**	**59**	**64**	**70**	**75**	**80**	**85**	**90**
16	—	3	8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	101	107
	—	**1**	**6**	**11**	**15**	**21**	**26**	**31**	**37**	**42**	**47**	**53**	**59**	**64**	**70**	**75**	**81**	**86**	**92**	**98**
17	—	3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	109	15
	—	**2**	**6**	**11**	**17**	**22**	**28**	**34**	**39**	**45**	**51**	**57**	**63**	**67**	**75**	**81**	**87**	**93**	**99**	**105**
18	—	4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	116	123
	—	**2**	**7**	**12**	**18**	**24**	**30**	**36**	**42**	**48**	**55**	**61**	**67**	**74**	**80**	**86**	**93**	**99**	**106**	**112**
19	0	4	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	16	123	130
	—	**2**	**7**	**13**	**19**	**25**	**32**	**38**	**45**	**52**	**58**	**68**	**72**	**78**	**85**	**92**	**99**	**106**	**13**	**119**
20	0	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138
	—	**2**	**8**	**13**	**20**	**27**	**34**	**41**	**48**	**55**	**62**	**69**	**76**	**83**	**90**	**98**	**105**	**112**	**119**	**127**

Table A.12 Wilcoxon signed-ranks table.[a]

	Level of significance for a one-tailed test					Level of significance for a one-tailed test			
	.05	*.25*	*.01*	*.005*		*.05*	*.25*	*.01*	*.005*
	Level of significance for a two-tailed test					Level of significance for a two-tailed test			
n	*.10*	*.05*	*.02*	*.01*	*n*	*.10*	*.05*	*.02*	*.01*
5	0	—	—	—	28	130	116	101	91
6	2	0	—	—	29	140	126	110	100
7	3	2	0	—	30	151	137	120	109
8	5	3	1	0	31	163	147	130	118
9	8	5	3	1	32	175	159	140	128
10	10	8	5	3	33	187	170	151	138
11	13	10	7	5	34	200	182	162	148
12	17	13	9	7	35	213	195	173	159
13	21	17	12	9	36	227	208	185	171
14	25	21	15	12	37	241	221	198	182
15	30	25	19	15	38	256	235	211	194
16	35	29	23	19	39	271	249	224	207
17	41	34	27	23	40	286	264	238	220
18	47	40	32	27	41	302	279	252	233
19	53	46	37	32	42	319	294	266	247
20	60	52	43	37	43	336	310	281	261
21	67	58	49	42	44	353	327	296	276
22	75	65	55	48	45	371	343	312	291
23	83	73	62	54	46	389	361	328	307
24	91	81	69	61	47	407	378	345	322
25	100	89	76	68	48	426	396	362	339
26	110	98	84	75	49	446	415	379	355
27	119	107	92	83	50	466	434	397	373

Source: Statistics: An Introduction, 3rd. ed., R. Kirk, © 1990 Holt, Rinehart and Winston, Inc., reprinted by permission of the publisher.

[a] The obtained T is significant at a given level if it is *equal to* or *less than* the value shown in the table.

Appendix B

Answers to Questions and Exercises

Chapter 1

1 **a** Observation
b Hypothesis
c Theory
d Theory
e Observation
f Hypothesis

2 **d** It is the only one of the four definitions that reflects a concrete way of measuring the concept for the purposes of a scientific study.

3 Here are some examples:
Description (1) How frequently are tiny homes being constructed, and is that frequency changing? (2) How many different types of tiny homes are there? (3) What is the ratio of different types of tiny homes? (4) What are the defining features of a tiny home? (5) Is tiny home living best understood as a temporary situation or as a lifestyle?
Correlation (1) Is there a relationship between tiny homeowners and various socioeconomic variables? (2) Is there a relationship between parts of the country and tiny home frequency? (3) Do tiny homeowners tend to have similar political viewpoints? Understanding (1) Are tiny homes more attractive as space is dedicated to specific functionality (e.g. bathroom, bedroom, kitchen) or to general living space? (2) Are tiny homes more attractive if they are mobile or anchored?
Comment: because the topic of interest is a major life decision (what kind of home to buy), it is hard to manipulate and study it experimentally. Some topics of study are limited because manipulation is either ethically

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

dubious, logically impossible, or logistically impractical. For instance, one interesting question would be to see if tiny house living changes one's political and social attitudes over time. A quasi-experimental idea might entail measuring participant's social and political attitudes prior to moving in to a tiny home and then some years later to measure change. This difference could be compared with changes in participants who, over the same period of time, are not living in a tiny home. This is only quasi-experimental, however, because participants are not being randomly assigned to the two conditions. They are choosing for themselves to either live in a tiny home or not.

4 **a** Independent variable: Vitamin E or amount of Vitamin E. Dependent variable: Time participants spent riding the bicycle. There are three levels of the independent variable (20 units; 60 units; placebo).

b Independent variable: Educational programs. Dependent variable: There are two dependent variables – comprehension and reading speed.

c Independent variable: Type of justification for behaving counter-attitudinally (insufficient, $1; sufficient, $50). Dependent variable: Amount of attitude change.

d Independent variable: Amount of natural light. Dependent variable: Number of widgets made.

5 Examples of quantitative independent variables: (1) 0 gummy bears, 2 gummy bears, 4 gummy bears; (2) no verbal praise, some verbal praise, a lot of verbal praise; (3) 10 minutes of screen time, 20 minutes of screen time. (Studies can have different numbers of levels of the independent variable.)

Examples of qualitative independent variables: (1) verbal praise, sugar snack, salty snack; (2) verbal praise, physical affection; (3) gifted toy, screen time, snack.

It would be important in many of these scenarios to "hold constant" the amount of physical affection offered to the child. For example, in both example "c's," the parent should either give no physical affection or give the same amount of physical affection in all conditions. This holds this variable constant and removes it from consideration if differences between conditions are found.

6 **a** The letter on the can is confounded with the types of root beer. A&W always has the letter *A* and Stewart's always has the letter *B*. Is it the taste that participants are responding to, or are they simply showing a preference for the letter *A* over *B*? This is admittedly a "stretch," but sometimes differences can be traced to seemingly innocuous differences such as this – see Box 1.1.

b The abnormal behavior of the mice might be due to the loud blast of noise alone and not because it is coupled with a difficult choice with significant consequences.

c The presence of the radar units could have caused motorists to drive more carefully. Without using the radar, it is possible that just reducing the speed limit (even if motorists obey the new speed limit) would not have an effect on traffic accidents.

d Participants are not randomly assigned to the two pain control conditions. It is possible that those participants who select the headphones are different in some way from those participants who select novocaine. Perhaps the "headphones" participants are less anxious about having work done on their teeth and therefore report less pain due to less initial anxiety.

7 Yes. It is possible that hearing a fast heart rate causes participants' actual heart rates to rise and hearing a low heart rate causes participants' heart rates to lower. The experimental effect, therefore, might be due to the difference in actual heart rates. Therefore, the variable "belief" may be confounded with the true level of heart rate. The researcher should record all participants' heart rates during the course of the experiment to make sure there is no change in actual heart rates as the various visual and auditory stimuli are presented.

8 **b**

9 Since participants were not randomly assigned to the "title-page" vs. "no title-page" conditions, it is possible that higher grades are associated with "title-page" papers not because the professor is biased in favor of them, but rather because the students with better writing skills in general have simply learned to include them.

10 "a" is for sure. "b" is most likely, unless there is some way of assigning ID numbers that would suggest that different types of students get different types of numbers. "c" is probably not a good method – to many reasons to think different types of students use different means of signing up. "d" is probably not a good system either – color preference could be associated with personality and temperament differences. "e" also has potential problems – if we are mainly using collegiate freshman, the younger 18 year olds are going to end up in one group, and the older 19 year olds will end up in the other.

11 "a" is problematic in many ways – one being that not everyone is likely to eat at the cafeteria; for instance, perhaps all students with extracurricular

activities eat later in the evening after practice and rehearsal. "b" is probably a good method even though the alphabetical order is set – still it is hard to argue that the resulting sample would not be representative of the students as a whole. "c" is very problematic because participants are selecting themselves – the resulting sample is very likely to misrepresent the larger population in terms of extroversion, amount of free time, helpfulness, etc. "d" is problematic because our classes will contain students who are largely from our major and at our academic level (freshman, sophomore, etc.). "e," although awkward and time consuming, would probably be an excellent way to generate a random sample of the student population.

12 **a** *Experiment*: First, select a method for inducing pain. Next, randomly assign participants to at least two levels of pain induction that differ in intensity (e.g. placing arms in buckets of water with different temperatures – both cold, but one colder than the other). For the dependent variable, select a known method for measuring anxiety, perhaps a self-report anxiety questionnaire and/or psychophysiological recordings. (As a check on the experimental manipulation, it would be a good idea to ask for pain ratings from the participants to document that the groups differ in their perception of pain regarding the two levels of the pain stimulus.)

Correlational design: Have participants experience a painful stimulus. Do not manipulate the level of pain. Measure each participant's pain perception and anxiety level.

b *Experiment*: Randomly assign participants to a "high frequency of exercise" condition and a "low frequency of exercise" condition. After a predetermined length of time, say, three months, obtain a measure of resting heart rate. It would be a good idea to document that the two groups do not differ in resting heart rate before the exercise program begins. In addition, we can use more than two experimental conditions. We could also have a group that is not asked to exercise. (Of course, we would have to document that this group actually exercises less than the low frequency group.)

Correlational design: Randomly select a group of participants. Find out how much participants exercise and measure their resting heart rate. Note that participants are not randomly assigned to different exercise conditions.

c *Experiment*: Need for achievement is a personality (participant) variable. There is no way to manipulate it and use it as an independent variable in an experiment. However, it would be possible to think of it as the dependent variable and see if the number of hours worked per week leads to changes in need for achievement. Here we would randomly

assign participants to different experimental conditions that differ in the amount of hours participants are required to work. At the end of some predetermined amount of time, the groups are compared on a measure of need for achievement.

Correlational design: Take a random sample of participants, and have them record the number of hours they work per week over a one-month period. Also, measure each participant's need for achievement.

d *Experiment*: Preschool children would be randomly assigned to attend or not attend day care. Measure all children's level of social skills when they are in first grade. Obviously, there are real-world problems with implementing this design. What if the parent of a child assigned not to attend day care wants their child to attend day care, and vice versa.

Correlational design: Take a random sample of children, making sure that the sample includes some children who will attend preschool and some children who will not attend preschool. Measure their social skills in first grade. Another approach is to take a random sample of first-grade students, measure social skills, and identify which students attended preschool and which did not.

Chapter 2

1 **a** The answer is "ratio" or "unknown"; it depends on the features of the original scale. The change between numbers on an interval or ratio scale is a ratio measure. A change from a 5 to a 7 (2 units) is half as large as a change from a 5 to a 9 (4 units), regardless of where the zero is anchored – assuming all intervals are conserved. If attitude was initially measured on an ordinal scale, however, then changing from one value to another cannot be meaningfully compared with other changes between values.

b Nominal

c Ordinal

d Ratio

e Ratio (80 heartbeats per minute are twice as many as 40)

f Interval or ordinal. (Assuming that need for approval has no meaningful zero point, it is definitely not a ratio scale. However, can we be sure that a Likert-type scale has constant intervals?)

g Ratio

h Nominal

i Nominal

j Nominal

k Ordinal

2 **a** *Nominal* scale examples include "pass/no pass" or "good student" and "bad student."

Ordinal scale examples include grade in a given class (A, B, C, D, F) or a set of categories like great, good, so-so, poor, and terrible.

Interval scale examples are hard to think of – debatably GPA is an interval measure since a "0" GPA is more easily thought of as a grouping of very poor students as opposed to a single place on a scale. For instance, one student may fail every class but barely, while another fails every class dreadfully. Both get a "0" GPA. A student with a GPA of "1" is not necessarily half of a student with a GPA of "2." Furthermore, an argument can be made that the spacing between integers is conserved, at least in terms of the number of academic points needed to move from one notch on the scale up to another.

Ratio scale examples include the number correct on a given measure or perhaps the number of degrees earned (1 degree is half of 2).

b *Nominal* scale examples include a "student–athlete/non-student–athlete," a self-reported "athlete/nonathlete," and an independent judge report of "athlete/nonathlete."

Ordinal scale examples include a Likert scale question like "how athletic are you?" (very, somewhat, not really, not at all), "what place did you earn in the tournament?" (first, second…), or "what rank one has on a tennis team?"

Interval scale examples are hard to think of – but one might include Likert scale questions, depending upon how one argues the Likert scale is to be interpreted.

Ratio scale examples include the number of trophies won, how far one can throw a javelin, and how fast one can run a mile.

c *Nominal* scale examples include categorizations like "creative/not creative."

Ordinal scale examples include a ranking by independent judges of creative products, a grade in a class involving creativity, and a Likert scale question like "how creative are you?" (very, somewhat, so-so, not very, not at all).

Interval scale examples are hard to think of – but one might include Likert scale questions, depending upon how one argues the Likert scale is to be interpreted.

Ratio scale examples include how many art awards a person has won and the number of judges out of 10 who classify a person as "artistic."

d *Nominal* scale examples include categorizations like "high amount" vs. "low amount."

Ordinal scale examples include Likert scale questions like "how much food did you eat today?" (a lot, an average amount, a little) or an independent judge rankings.

Interval scale examples are hard to think of – but one might include Likert scale questions, depending upon how one argues the Likert scale is to be interpreted; weighing the plate might be considered interval if one does not take out the weight of the plate itself.

Ratio scales examples include the weight of food (minus the plate), the number of calories eaten, or even the difference between pre-meal weight and after-meal weight.

e Nominal scale examples include categorizations like "large family" vs. "small family."

Ordinal scale examples include Likert scale questions like "how big is your extended family?" (very, somewhat, so-so, not very, not at all) or some family-sized ranking system or the size of the banquet hall that needs to be reserved for a reunion (small, medium, large, extra large).

Interval scale examples are hard to think of – but one might include Likert scale questions, depending upon how one argues the Likert scale is to be interpreted.

Ratio scale examples include a total count of all extended family (once properly defined) or the number of cousins one has.

3

	Width	*LL*	Midpoint	*UL*
a.	3	0.5	2	3.5
b.	6	4.5	7.5	10.5
c.	5	−8.5	−6	−3.5
d.	5	−2.5	0	2.5
e.	3	1.000	2.5	4.000
f.	26	24.5	37.5	50.5

4 and 5 Part *a*. Only the top and bottom four numbers of the distribution are provided here (and also in part *b*).

LL	*X*	*UL*	*f*	*cf*
97.5	98	98.5	1	36
96.5	97	97.5	0	35
96.5	96	96.5	0	35
94.5	95	95.5	0	35
⋮	⋮	⋮	⋮	⋮

(*Continued*)

(Continued)

LL	*X*	*UL*	*f*	*cf*
43.5	44	44.5	0	3
42.5	43	43.5	2	3
41.5	42	42.5	0	1
40.5	41	41.5	1	1

Part *b*.

LL	*X*	*UL*	*f*	*cf*
95.5	96–98	98.5	1	36
92.5	93–95	95.5	0	35
89.5	90–92	92.5	3	35
86.5	87–89	89.5	3	32
⋮	⋮	⋮	⋮	⋮
47.5	48–50	50.5	1	5
44.5	45–47	47.5	1	4
41.5	42–44	44.5	2	3
38.5	39–41	41.5	1	1

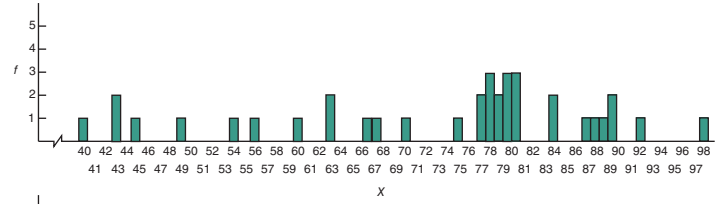
Part *c*.

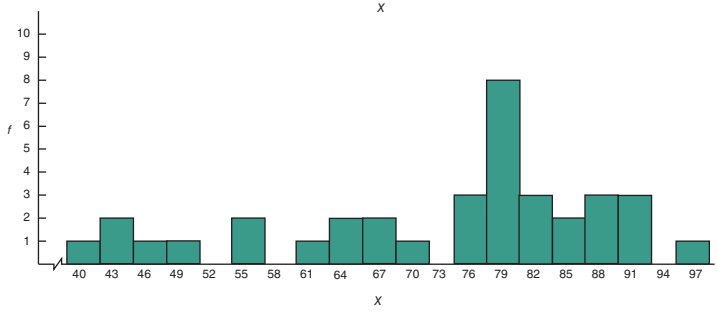
LL	*X*	*UL*	*f*	*cf*
89.5	90–99	99.5	4	36
79.5	80–89	89.5	11	32
69.5	70–79	79.5	9	21
59.5	60–69	69.5	5	12
49.5	50–59	59.5	2	7
39.5	40–49	49.5	5	5

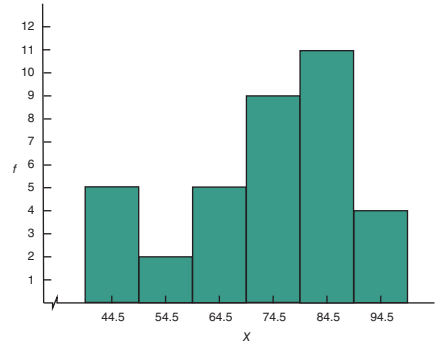
Part *d*.

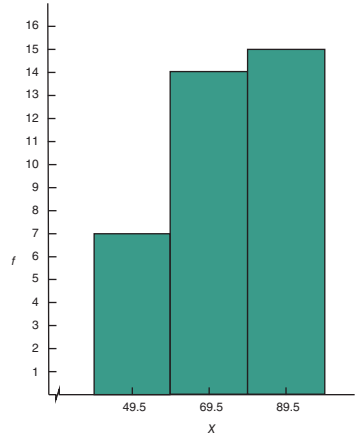
LL	*X*	*UL*	*f*	*cf*
79.5	80–99	99.5	15	36
59.5	60–79	79.5	14	21
39.5	40–59	59.5	7	7

6 Refer to graph.

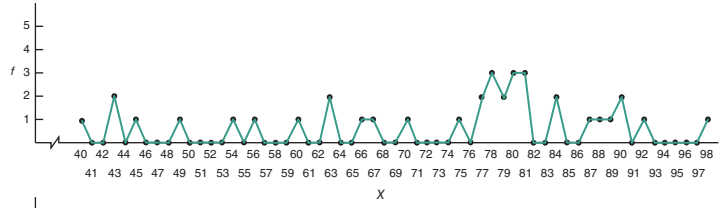


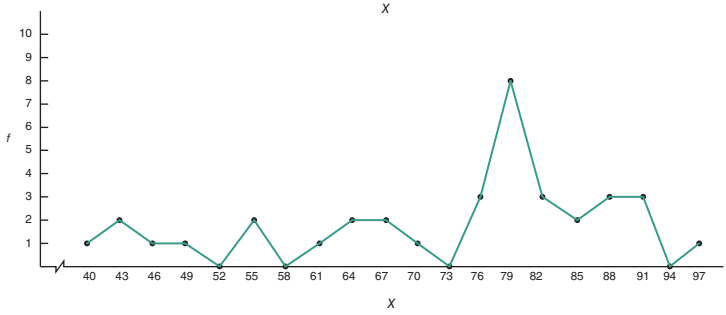


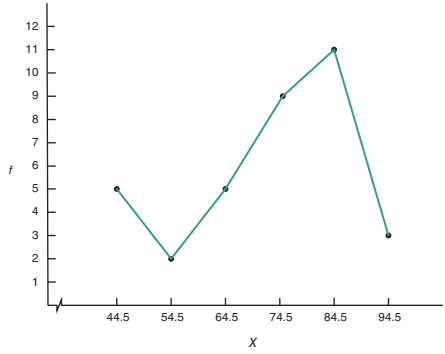


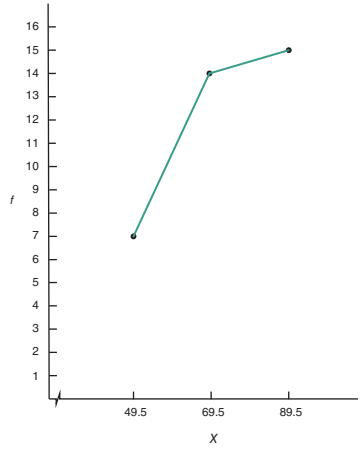


7 Refer to graph.







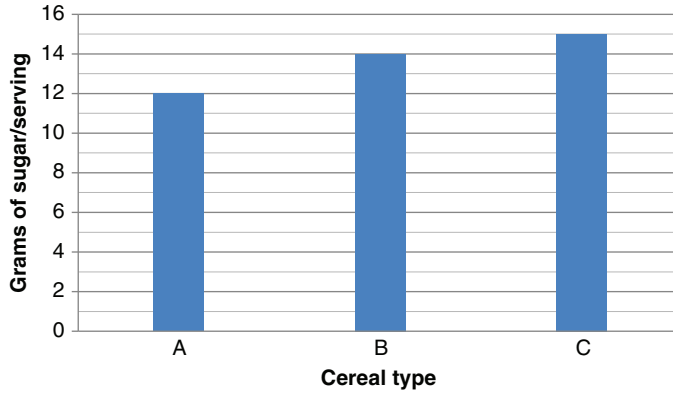


8 Here are two examples: pounds of garbage per week for a given family (it makes sense to think that most weeks would produce garbage amounts that are roughly similar to most other weeks – with some weeks being a bit less and some weeks being a bit more) and the total inches of snowfall for a given northern American city (it makes sense to think that most years produce a total snowfall that is very similar to most other years – with some years being a bit less and some being a bit more).

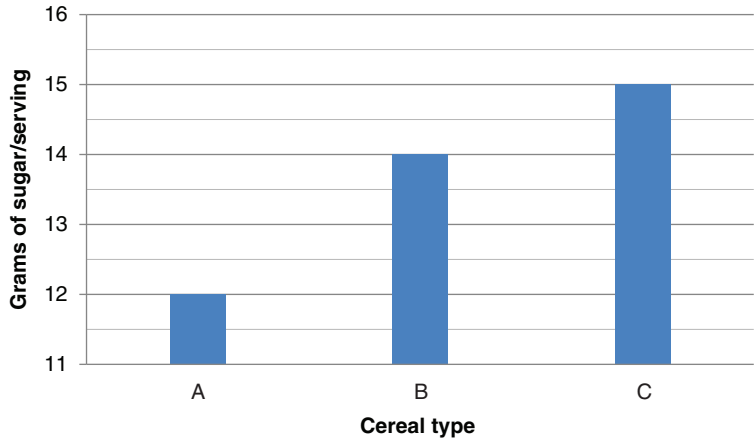
9 Here are two examples: GPA's at a given university (it makes sense to suggest that most GPA's will cluster around 3.25 or 3.3 with some being greater, but many more will trail off down the scale to about a 1.75 or 1.5 – of course at some point, students are put on academic probation or not allowed to return to school – so this distribution might have an artificial bottom point) and free-throw percentages of professional basketball players (it makes sense to think that most professionals have a high rate of success for these uncontested shots, but some do struggle – and since the ceiling of 100% cannot be exceeded, the dispersion is much more likely to stretch out toward the lower percentages).

10 Here are two examples: yards rushed by an NFL running back (it makes sense that most rushing attempts produce just a few yards gained, sometimes even the loss of a few years; but there will be a good number of rushes that will amount to significant yardage gains) and completion times for a triathlon (it makes sense to suggest that there will be a few winners with shorter times, but then a bulk of competitors will finish soon thereafter; however, there will be people straggling in for hours after the bulk of runners have finished).

11 Here is a picture of a graph drawn in Excel that faithfully represents the relationship between the cereal types in terms of sugar per serving.



And here is a graph drawn in Excel that truncates the Y axis without labeling it, potentially leaving the viewer with a false understanding of the relationship between the cereal types in terms of sugar per serving.

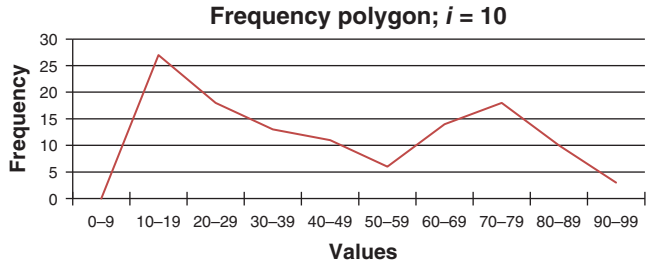


12 For space purposes, a simple frequency distribution is not provided.

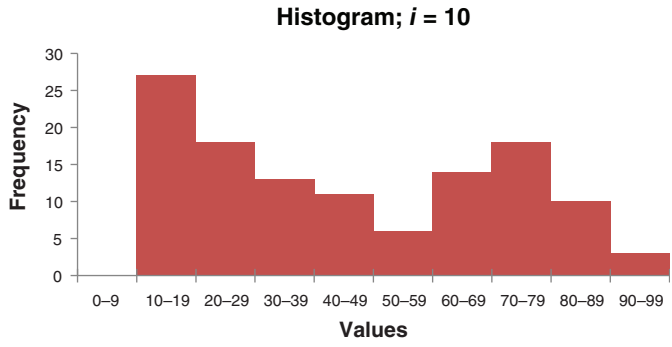
Here is a grouped frequency distribution – with a column for cumulative frequency.

LL	*X*	*UL*	*f*	cum *f*
89.5	90–99	99.5	3	120
79.5	80–89	89.5	10	117
69.5	70–79	79.5	18	107
59.5	60–69	69.5	14	89
49.5	50–59	59.5	6	75
39.5	40–49	49.5	11	69
29.5	30–39	39.5	13	58
19.5	20–29	29.5	18	45
9.5	10–19	19.5	27	27

Here is a frequency polygon of the data – using an $i = 10$.



Here is a histogram of the data – using an $i = 10$.



Chapter 3

1 It must be symmetrical.

2 The mean will still be 15. The formula for a sample mean, though using different symbols, is functionally equivalent to the formula for a population mean.

3 **a** 2
b –1
c –10
d 0
e –11
f –0.5

4 **a** −2
b 2
c 9
d −0.5
e −29
f 2.5

5

a.	$M = 6$	Median = 6	Mode = 8
b.	$M = 4.86$	Median = 4	Mode = 4
c.	$M = 8.71$	Median = 9	Mode = 10
d.	$M = 3.86$	Median = 4	Mode = 1 and 4
e.	$M = 5.67$	Median = 6.5	Mode = 8
f.	M = 8.22	Median = 9	Mode = 5

For "f," remember to rearrange the numbers from lowest to highest.

6 **a**

Distribution A $\Sigma(X - M)$		Distribution B $\Sigma(X - M)$	
$X - M$	x	$X - M$	x
3 – 6	−3	2 – 4.86	−2.86
3 – 6	−3	4 – 4.86	−0.86
4 – 6	−2	4 – 4.86	−0.86
5 – 6	−1	4 – 4.86	−0.86
6 – 6	0	6 – 4.86	+1.14
8 – 6	+2	7 – 4.86	+2.14
8 – 6	+2	7 – 4.86	+2.14
8 – 6	+2		$\Sigma x = -0.02$ (will be 0 without rounding error)
9 – 6	+3		
	$\Sigma x = 0$		

b

Distribution A $\Sigma(X - \text{Median})$	Distribution B $\Sigma(X - \text{Median})$
$X - \text{Median}$	$X - \text{Median}$
3 – 6 = −3	2 – 4 = −2
3 – 6 = −3	4 – 4 = 0
4 – 6 = −2	4 – 4 = 0
5 – 6 = −1	4 – 4 = 0

(Continued)

Distribution A $\Sigma(X - \text{Median})$	Distribution B $\Sigma(X - \text{Median})$
X – Median	***X – Median***
6 – 6 = 0	6 – 4 = +2
8 – 6 = +2	7 – 4 = +3
8 – 6 = +2	7 – 4 = +3
8 – 6 = +2	$\Sigma(X - \text{Median}) = +6$
9 – 6 = +3	
$\Sigma(X - \text{Median}) = 0$	

c

Distribution A $\Sigma(X - \text{Mode})$	Distribution B $\Sigma(X - \text{Mode})$
X – Mode	***X – Mode***
3 – 8 = –5	2 – 4 = –2
3 – 8 = –5	4 – 4 = 0
4 – 8 = –4	4 – 4 = 0
5 – 8 = –3	4 – 4 = 0
6 – 8 = –2	6 – 4 = +2
8 – 8 = 0	7 – 4 = +3
8 – 8 = 0	7 – 4 = +3
8 – 8 = 0	$\Sigma(X - \text{Mode}) = +6$
9 – 8 = +1	
$\Sigma(X - \text{Mode}) = -18$	

The preceding exercise illustrates that $\Sigma(X - M)$ equals 0. The $\Sigma(X - \text{Median})$ or $\Sigma(X - \text{Mode})$ will only equal 0 if the distribution is perfectly symmetrical or by pure coincidence. Neither distribution A nor B is normal. The fact that $\Sigma(X - \text{Median}) = 0$ for distribution A is coincidental.

7 **a** $M = 8.65$
b Mode = 9

8 **a** $M = 16.29$
b Mode = 15 and 16

9 Median = 6.5 + 0.5 = 7

10 Grand Mean = 4318/30 = 143.93

11 Grand Mean = 102/17 = 6

12
a Positively skewed
b Negatively skewed
c Symmetrical, unimodal
d Symmetrical, bimodal
e Negatively skewed
f Positively skewed

13 Each one needs to think of their own examples, but here are a couple to point us in the right direction. (1) Height measurements for a basketball team composed of only guards and centers. (The shorter guards would all be clustered around a smaller height measure, and the taller centers would all be clustered around a larger number.) (2) The running times for Olympians running the 100-m dash. (The men's times would be clustered around just under 10 seconds, while the women's times would be clustered around 11 seconds.)

14 $M = \dfrac{552 + 551 + 448}{107} = 14.50$

15
a M = 104.80
b Median = 101
c Yes, it is also 101

16 The mean has the biggest difficulty with extreme scores. Because the mean takes into account the distance from each score to the middle, extreme scores, especially for small data sets, can generate a number that seems to be far away from the bulk of the scores.

17 The median. Neither an ordinal scale nor the concept of the median makes any assumptions about the uniformity of the intervals between values.

18 The mode. Whenever the data are in the form of *how many* (i.e. a nominal scale) rather than *how much* (i.e. an interval or ratio scale), the mode is the only appropriate measure of central tendency.

19 101. If the mean of the original distribution is 100, the sum of the 10 scores must be 1000. If one of those numbers goes from 80 to 90, the new sum must be 1010. There are still a total of 10 numbers, so the new mean is now 101.

20 11.3 If the mean of the original distribution is 12, the nine scores center on 12. This means the nine scores added together must equal 108. If we add 5 to that number and then divide it by 10 (since our n is now 9 + 1), the new mean will be 11.3

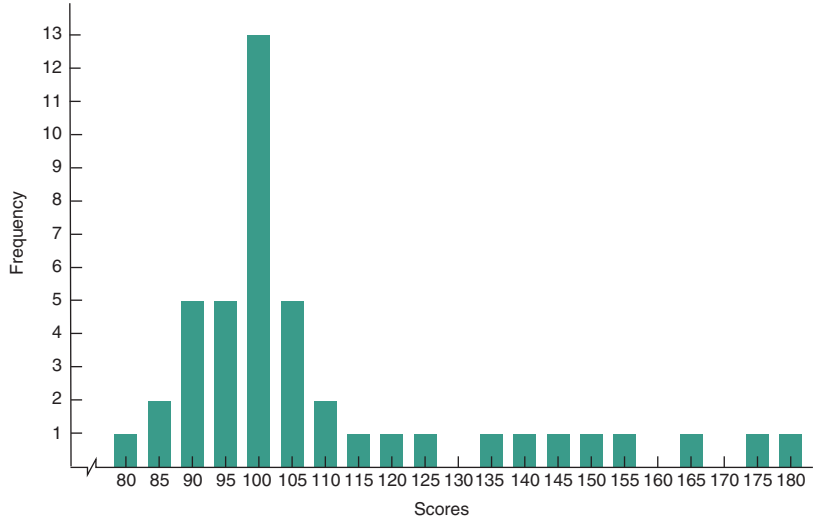
21 43. The first set of 17 numbers must sum to 425, since the mean is 25. If the new distribution of 18 numbers now has a mean of 26, then the new sum of the scores is 468. The difference between the two sums is 43, which means the value of the new number is 43.

22 27. The first set of 6 scores must sum to 150, since the mean is 25. If one score is removed, we now have 5 scores summing to 135, which means the value of the new mean is 27.

23 Fifty-two minutes is the best guess. As sample size increases, sample statistics better approximate population statistics. The sample including 500 people is most likely the most accurate.

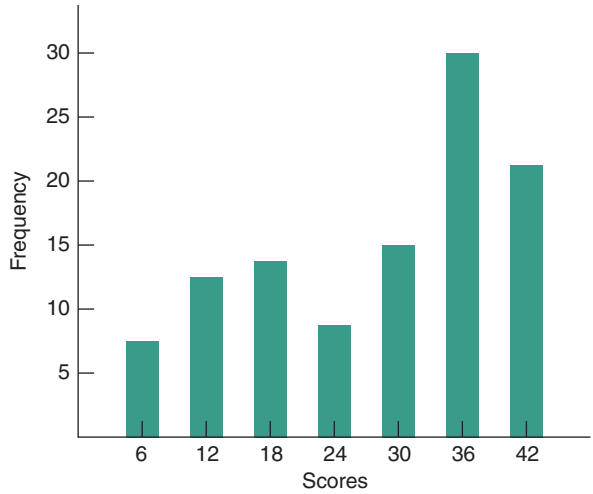
24 M = 109.73; median = 100.50; mode = 100

The following histogram reveals a positively skewed distribution, which is consistent with the fact that the mean is greater than the median. (Note: Some computer printouts represent histograms as bar graphs in that adjacent bars do not share a common border, which is the case for the next two graphs.)



25 $\mu = 28.12$; median = 32; mode = 35

The histogram shows this distribution to be negatively skewed, consistent with the fact that the mean is smaller than the median.



Chapter 4

1 Cannot say which would have the *larger* variance, but $n = 60$ would likely give a more accurate estimate of the population variance. A sample size of 60 would likely have the larger range.

2 The variance cannot be estimated based on the size of M.

3 **a** $M = 103.50$
b $s = 5.24$

4 **a** Range = 8 – 1 = 7
b $s^2 = 4.27$
c $s = 2.07$

5 **a** $IQR = 111 - 81 = 30$
b $SIQR = (111 - 81)/2 = 15$

6 **a** 3
b 12.8
c 2.58

7 **d**

8 All scores are 50. There is no deviation off that mean value.

9 **a** σ
b s
c σ
d s

10 The denominator in the sample formula has a correction factor $(n - 1)$. This is to make the resulting number and unbiased estimate of the population variance. Otherwise, a sample variance would be a biased estimate of the population variance; it would most likely be too small. Reducing the numerator by one helps correct this problem.

11 $\sigma = 4.83$ and $\sigma^2 = 23.35$

12 $s_A^2 = 7.125$ $s_B^2 = 5.90$

13 No. The 68-95-99.7 rule assumes a normal distribution.

14 Affected by extreme scores.

15 The standard deviation is the easiest to interpret because it is in the original units of the measured variable.

16 $\sigma = 2.66$

17 According to the 68-95-99.7 rule, values 40–60 would encapsulate the middle 68% of scores, 30–70 would encapsulate the middle 95% of scores, and 20–80 would encapsulate the middle 99.7% of scores.

18 According to the 68-95-99.7 rule, values 48–52 would encapsulate the middle 68% of scores, 46–54 would encapsulate the middle 95%, and 44–56 would encapsulate the middle 99.7%.

19 According to the 68-95-99.7 rule, the standard deviation must be 12 (the middle 68% are contained within plus and minus 12 points off the mean). This means the variance of this distribution must be 144.

20 A distribution with a variance of 100 has a standard deviation of 10 (the square root of the variance). According to the 68-95-99.7 rule, the values 120 and 160 must be plus and two standard deviations away from the

mean, respectively. If the standard deviation is 10, that means the distribution must have a mean of 140.

21 **a** 15
b 250
c 30
d 500

22 $M = 14.50$ $s^2 = 1.60$

23

	Experimental (biofeedback)		Control	
a.	Pre:	$s = 12.95$	Pre:	s = 9.11
b.	Post:	$s = 10.89$	Post:	s = 11.19

24

	Technique *A*	Technique *B*
a.	$M_{Pre} = 3.50$	$M_{Pre} = 3.00$
b.	$s^2_{Pre} = 1.67$	$s^2_{Pre} = 0.67$
c.	$s_{Pre} = 1.29$	$s_{Pre} = 0.82$
d.	$M_{Post} = 5.50$	$M_{Post} = 3.5$
e.	$s^2_{Post} = 1.67$	$s^2_{Post} = 1.67$
f.	$s_{Post} = 1.29$	$s_{Post} = 1.29$

25

$X + 10$	$X - 10$	$X(10)$	$X/10$
$\mu = 60$	$\mu = 40$	$\mu = 500$	$\mu = 5$
$\sigma^2 = 25$	$\sigma^2 = 25$	$\sigma^2 = 2500$	$\sigma^2 = 0.25$

26 The coach would need to transform the data by dividing each value recorded by three, thereby turning feet measurements into yards.

27 25 minutes. As sample size increases, sample statistics better approximate population statistics. The sample including 500 people is most likely the most accurate.

28 Range = 15 $IQR = 6$ $SIQR = 3$

29 Range = 40 $IQR = 19.75$ $SIQR = 9.88$

30 $M = 7.07$ $s^2 = 17.61$ $s = 4.20$ Range = 14

31 $M = 96.52$ $s^2 = 138.56$ $s = 11.77$ Range = 39

32 $M = 0.1675$ $s^2 = 0.0017$ $s = 0.0415$ Range = 0.1343

33 $M = 965.24$ $s^2 = 13\,856.20$ $s = 117.71$ Range = 390

34 $M = 2\,895.72$ $s^2 = 41\,568.6$ $s = 203.88$ Range = 1170

35 $M = 321.75$ $s^2 = 1539.58$ $s = 39.24$ Range = 130

Chapter 5

1

a $PR = \frac{23 + 0.5(12)}{49} \cdot 100 = 0.59(100) = 59$

b $PR = \frac{42 + 0.5(4)}{49} \cdot 100 = 0.90(100) = 90$

c $PR = \frac{13 + 0.5(10)}{49} \cdot 100 = 0.37(100) = 37$

d $PR = \frac{6 + 0.5(7)}{49} \cdot 100 = 0.19(100) = 19$

2 A z score represents the number of standard deviations a raw score is away from the mean.

3 A positive z score means the corresponding raw score is larger than the mean, while a negative z score means the corresponding raw score is smaller than the mean.

4 Answers will vary. Variables that may be normally distributed include the number of slices of pizza eaten in a month by university students, the body

weight of biological male students, the level of extroversion of all students, etc.

5 Answers will vary. Variables that might not be normally distributed include the miles away from home for university students (positively distributed), the driving speed of cars on an interstate relative to the speed limit (positively distributed), the number of credits taken per semester by university students (negatively distributed), the number of miles on car odometers (positively distributed), etc.

6 $\mu = 7.67$ $\sigma = 2.56$

X	z
4	−1.43
5	−1.04
7	−0.26
9	+0.52
10	+0.91
11	+1.30

7 $z = \dfrac{11 - 14}{4} = \dfrac{-3}{4} = = -0.75$

8 $X = 25 + 0.36(3) = 25 + 1.08 = 26.08$

9 $z = \dfrac{140 - 130}{13} = \dfrac{10}{13} = 0.77$ **Answer: 0.22 or 22%**

10 $z = \dfrac{27 - 34}{3} = -2.33$ **Answer: 0.99%**

11 $0.025 + 0.025 = 2.5\% + 2.5\% = 5\%$

12 $0.40 + 0.40 = 40\% + 40\% = 80\%$

13 **a** $z = -2.0$
b $z = -0.30$
c $z = +0.47$

d $z = +1.48$
e $z = -1.20$
f $z = +0.54$
g $z = +0.09$
h $z = -0.19$
i $z = +0.80$

14 $0.3849 + 0.2123 = 0.5972$

15 $0.3051 + 0.1141 = 0.4192$

16 $0.1587 + 0.1587 = 0.3174 = 31.74\%$

17 (z's = ±1) *Answer*: $0.3413 + 0.3413 = 0.6826$

18 z's = ±1.28 (from table)
$X = 70 \pm 1.28(7) = 61$ and 79
79 = best students and <61 worst students

19 $\mu = 5.67$ $\sigma = 2.36$

X	*z*
2	−1.56
4	−0.71
5	−0.28
6	+0.14
8	+0.99
9	+1.41

20 We can find the percentile rank by converting to z scores and using the z table.

a.	$z = -1.25$	PR = 10.56
b.	$z = +1.25$	PR = 98.78
c.	$z = 0$	PR = 50
d.	$z = +0.50$	PR = 69.15
e.	$z = -0.25$	PR = 40.13

21

a.	$z = 1.645$ (from table)	$X = 78 + 1.645(7) = 90$ (rounded)
b.	$z = 0.84$ (from table)	$X = 78 + 0.84(7) = 84$ (rounded)
c.	$z = -0.52$ (from table)	$X = 78 + (-0.52)(7) = 74$ (rounded)
d.	$z = -0.13$ (from table)	$X = 78 + (-0.13)(7) = 74$ (rounded)

22
a 11.51%
b 50% + 14.06% = 64.06%
c 28.77%
d 21.48%
e 13.57%
f 50% + 33.65% = 83.65%
g 44.83%

23
a $z = \dfrac{38-56}{5} = -3.6$ **Answer: 0.02%**
b $z = \pm 0.39$
X's = 56 ± (0.39)(5) = 54.05 and 57.95, so "*C*" category is roughly 54–58.
c $z = 1.28$
$X = 56 + 1.28(5) = 62.4$. The "*A*" category is (rounding) 62 and up.

24
a $z = 0.84$ (from table)
b No, we need a mean and standard deviation.

25 A z score is the number of standard deviations between a raw score and the mean. If a raw score that is 10 points below the mean corresponds to a z score of −2.50, the standard deviation must be 4. −2.5(4) = −10.

26 If a raw score that is 5 points above the mean corresponds to a z score of 2.00, the standard deviation must be 2.5. 2(2.5) = 5.

27 If a raw score of 51 corresponds to a z score of −1.00, then 51 is one standard deviation below the mean; therefore, the mean must be 65. $X = \mu + (z\sigma)$ or 51 = 65 + (−1∗15).

28 If a raw score of 31 corresponds to a z score of 2.00, then 31 is two standard deviations above the mean; therefore, the mean must be 21. $X = M + (zs)$ or 31 = 21 + (2∗5).

29 A z score is the number of standard deviations between a raw score and the mean. If the mean is 60 and a raw score of 61 corresponds to a z of 0.20, then the standard deviation must be 5. $X = \mu + z\sigma$ or 61 = 60 + (0.2∗5).

30 If the mean is 75 and a raw score of 60 corresponds to a z of –2.00, then the standard deviation must be 7.5. $X = M + zs$ or $60 = 75 + (-2*7.5)$.

31 If a raw score of 35 corresponds to a z score of –1.00 (which means it is one standard deviation below the mean) and a raw score of 40 corresponds to a z score of –0.50 (which means it is one-half of a standard deviation below the mean), then s must be two times the distance between 35 and 40, that is, 10. And if 35 is 1 s below the mean and 40 is 0.50 s below, $M = 45$. (Hint: if we are having trouble, draw it out.)

32 If a raw score of 72 corresponds to a z score of 0.20 (which means it is 0.2 standard deviation above the mean) and a raw score of 84 corresponds to a z score of 0.80 (which means it is 0.8 standard deviations above the mean), the distance between 72 and 84 must be 0.6 σ away from each other. Therefore, $\sigma = 12/0.6 = 20$. Further, this means $\mu = 68$. (Hint: if we are having trouble, draw it out.)

33 If a raw score of 16 corresponds to a z score of –2.00 (which means it is two standard deviation below the mean) and a raw score of 23.5 corresponds to a z score of 3.00 (which means it is three standard deviations above the mean), then the distance between the scores must be 5 s away from each other. Therefore, $s = 7.5/5 = 1.5$. Further, this means $M = 19$. (Hint: if we are having trouble, draw it out.)

34 If a raw score of 77 corresponds to a z score of 2.50 (which means it is 2.5 standard deviation above the mean) and a raw score of 41 corresponds to a z score of –5.00 (which means it is 5 standard deviations below the mean), the distance between 77 and 41 must be 7.5 σ away from each other. Therefore, $\sigma = 36/7.5 = 4.8$. Further, this means $\mu = 65$. (Hint: if we are having trouble, draw it out.)

35 Bottom 20% corresponds to a z score of –0.84; $X = \mu + z\sigma$ or $X = 25\,000 + (-0.84*6\,000) = 19\,960$.
Bottom 40% corresponds to a z score of –0.25; $X = 25\,000 + (-0.25*6\,000) = 23\,500$.
Bottom 60% corresponds to a z score of 0.25; $X = 25\,000 + (0.25*6\,000) = 26\,500$.
Bottom 80% corresponds to a z score of 0.84; $X = 25\,000 + (0.84*6\,000) = 30\,040$.
The highest value cannot be determined since theoretically the corresponding z score would be infinite. (Recall that normal distributions are asymptotic.)

36 The top 15% corresponds to a z score of 1.04; $X = \mu + z\sigma$ or $X = 45 + (1.04*11) = 56.44$ pounds of garbage.
The bottom 28% corresponds to a z score of −0.58; $X = 45 + (-0.58*11) = 38.62$ pounds of garbage.

37 $X = M + zs$ so, $X = 25 + (-1.75*4) = 18$; the z table suggests that 95.99% of the raw scores will be greater than 18. (50 + 45.99 = 95.99%)

38 $X = \mu + z\sigma$ *so,* $X = 99 + (1.33*9) = 111$; the z table suggests that 9.18% of the raw scores will be greater than 111.

39 There are several ways to find this area under the curve. Here is one:
The raw score of 170 corresponds to a z score of 1.33; 9.18% of the area under the curve is beyond a raw score of 170, including area we do not want to include. The raw score of 175 corresponds to a z score of 1.67; there is 4.75% of the area under the curve beyond that point. We can subtract 4.75% from 9.18%, which leaves us with 4.43% between a raw score of 170 and 175.

40 There are several ways to find this area under the curve. Here is one:
The raw score of 0.7 corresponds to a z score of −2.00; 2.28% of the area under the curve is below a raw score of 0.7, including area we do not want to include. The raw score of 0.6 corresponds to a z score of −2.67; there is 0.38% of the area under the curve below that point. We can subtract 0.38% from 2.28%, which leaves us with 1.90% between a raw score of 0.6 and 0.7.

41 Andrew's 54 completed passes from a distribution centered on 44 with a standard deviation of 6 produce a z score of 1.67. Lisa's 48 completed passes from a distribution centered on 38 with a standard deviation of 7 produce a z score of 1.43. Andrew's passing performance was stronger than Lisa's relative to their respective teams.

42 Sarah's 20 minutes deciding what to wear from a distribution centered on 15 with a standard deviation of 4 produces a z score of 1.25. Justine's 90 minutes on social media from a distribution centered on 65 with a standard deviation of 20 produces a z score of 1.25. Relative to their respective activities, the time wasted by each person is the same.

43 $$PR = \frac{42 + (0.5)16}{76} \cdot 100 = 66$$

44

$$X_{0.40} = 13.5 + \frac{(126)(0.40) - 46}{27} \cdot 4 = 14 \ (rounded)$$

$$X_{0.50} = 13.5 + \frac{(126)(0.50) - 46}{27} \cdot 4 = 16 \ (rounded)$$

$$X_{0.65} = 17.5 + \frac{(126)(0.65) - 73}{23} \cdot 4 = 19 \ (rounded)$$

$$X_{0.90} = 25.5 + \frac{(126)(0.90) - 109}{13} \cdot 4 = 27 \ (rounded)$$

45 $\mu = 20.28$; $\sigma^2 = 43.17$; $\sigma = 6.57$

z scores based on the population standard deviation are in table below.

−1.26	−0.80	2.08	0.41	1.78	−1.26	0.26	0.11	−0.19	0.72	−0.95	−1.41	−1.26
−1.41	−1.56	−0.95	−0.80	−1.10	−1.26	−0.65	−0.35	0.11	1.32	1.78	1.62	1.47
0.56	1.47	1.32	1.17	0.87	0.11	−0.19	−0.50	−0.65	−0.80	−1.41	−1.56	−0.50
1.78	1.47	1.32	1.32	1.17	1.02	0.11	−0.95	0.11	−0.35	−0.65	−0.65	−1.41
−0.04	0.41	−0.95	−0.80	−0.50	−1.41	0.11	1.78	−0.04	−0.04	0.72	−0.80	−0.50
−0.95	−0.80	0.41	0.87	1.47	0.56	−0.19	0.41	0.26	0.11	0.56	−0.50	−0.80

46 $M = 11.15$; $s^2 = 30.68$; $s = 5.54$

z scores based on the sample standard deviation are in table below.

−0.21	−1.11	−1.83	1.42	0.33	−0.93	−0.03	0.15	−0.39	0.69	1.06	1.06	−0.93
−1.29	0.87	1.42	−0.57	0.33	−0.03	−0.75	1.24	0.87	−0.75	−0.93	0.87	−1.65
−0.75	−0.75	−0.03	−0.57	−1.29	−0.03	1.24	−0.21	0.51	1.60	0.69	−1.29	1.42
−0.39	−1.47	−0.57	0.87	−1.11	−0.75	−1.83	1.42	1.60	1.24	0.15	−0.39	−1.29
−0.39	−0.03	−1.11	0.69	−1.11	1.06	1.06	−0.39	1.24	−1.83	−0.57	1.24	−0.93
0.87	−0.93	0.15	−0.93	−0.93	1.24	1.42	−0.03	1.24	−0.39	1.42	1.06	−0.03

47 $M = 234.47$; $s^2 = 13\,094.30$; $s = 114.43$

z scores are in table below.

−1.07	−0.52	0.96	−0.97	0.81	1.55	−0.98	−0.15	1.61	−0.96	−0.76	1.54	−1.07
−1.07	−1.11	−0.78	−0.96	−0.45	−0.72	−0.95	−0.41	0.14	−0.05	1.29	1.02	0.75
0.07	0.65	3.45	0.42	0.30	−0.16	−0.36	−0.33	−0.60	−0.71	−1.01	−1.12	−0.54
0.78	−0.91	0.39	1.70	−0.93	0.37	1.63	−1.05	−0.15	−0.44	−1.04	−0.59	1.54
2.50	−0.10	−0.76	−0.69	−0.94	−0.47	0.14	0.076	0.40	0.14	−0.08	−0.43	−0.68
−0.62	−0.71	0.04	0.28	0.60	0.52	−0.34	−0.01	−0.05	3.38	0.07	−0.33	−0.69

48 $\mu = 3.94$; $\sigma^2 = 4.84$; $\sigma = 2.20$

z scores based on the population standard deviation are in table below.

0.12	2.07	−0.24	−0.74	−0.34	−0.79	−0.79	−0.52	1.35	−0.65	−1.15	2.35	−0.79
0.76	2.35	−0.70	0.26	1.07	−1.24	−0.15	2.66	−0.65	−0.47	−0.34	−0.06	−0.38
−0.70	−0.43	−0.47	−0.52	−0.61	−0.47	0.44	−1.02	0.30	1.62	−1.24	1.39	−0.11
−0.34	−0.38	−0.47	−0.47	−0.52	−0.56	−0.65	−0.70	−0.74	2.21	−0.15	1.21	0.26
−0.88	−0.74	−0.24	0.71	1.71	1.89	−0.61	−0.34	−0.88	−0.84	−0.65	0.71	1.26
1.57	−1.11	−0.74	−0.61	−0.38	−0.70	1.80	−0.74	−0.79	−0.47	−0.70	0.35	1.62

Chapter 6

1 **b**

2 **a** 13/52 or ¼ or 0.25 or 25%
b 1/10 or 0.1 or 10%
c 16/32 or ½ or 0.5 or 50%
d 1/50 or 0.02 or 2% (assuming all states are equally likely to be drawn)
e Impossible to determine without knowing more information

3 **a** No, some countries allow dual citizenship.
b No
c Yes
d No
e No
f Yes
g Yes
h Yes
i Uncertain(!)

4 **a** $\frac{40}{100} = 0.40$ or 40%
b $\frac{60}{100} = 0.60$ or 60%
c 0.40 + 0.60 = 1.00 or 100%
d (0.40)(0.40) = 0.16 or 16%
e (0.60)(0.60) = 0.36 or 36%

5 **a** Dependent (getting one number necessarily changes the likelihood of getting another).
b Independent (the numbers selected the previous week do not change the likelihood of the numbers selected the subsequent week).

c Dependent (once one 5 card is selected, the probability of drawing another next decreases).
d Independent (the first drawing does not change the likelihood of getting a 6 on the second drawing).

6 **a** 0.3
b 0.4
c 0.7
d 0.6 (mutually exclusive – so 0.3 + 0.3)
e 0.55 (not mutually exclusive – so 0.3 + 0.35 – 0.1)
f 0.1 (uncertain relationship – so $P(Red|X)\ P(X)$ = 10/35 (0.35))
g 0.29 (conditional – so 0.1/0.35)
h 0.33 (conditional – so 0.1/0.3)
i No
j $P(red)P(X)$ does not equal $P(Red|X)P(X)$.

7 **a** 0.35
b 0.4
c 0.65 (mutually exclusive – so 0.25 + 0.4)
d 0 (mutually exclusive – so 0.35 + 0.25 + 0.4)
e 0.75 (not mutually exclusive – so 0.25 + 0.6 – 0.1)
f 0.65
g 1 (mutually exclusive – so 0.6 + 0.4)
h 0.3 (uncertain relationship – so $P(Yellow|X)\ P(X)$ = 0.5 (0.6))
i 0 (uncertain relationship – so $P(Yellow|Red)\ P(Red)$ = 0 (0.6))
j 0.375 or 0.38 (conditional – so $P(Red\ and\ Y)/P(Y)$ = 0.15/0.4)
k 0.6 (conditional – so $P(Y\ and\ Red)/P(Red)$ = 0.15/0.25)
l No, because $P(X)\ P(Green) \neq P(X|Green)\ P(Green)$; change probabilities to make $P(X) = P(X|Green)$.
m No, because $P(Yellow)\ P(Y) \neq P(Yellow|Y)\ P(Y)$; change probabilities to make $P(Yellow) = P(Yellow|Y)$.

8 **a** 0.8 or 80%
b 0.4 (not mutually exclusive – so 0.2 + 0.3 – 0.1)
c 0.33 (conditional – so 0.1/0.3)
d 0.5 (conditional – so 0.1/0.2)
e Close, but no
f $P(Red)\ P(X) \neq P(Red|X)\ P(X)$; to make them independent we must make $P(Red) = P(Red|X)$.

9 **a** 0.3
b 0.5
c 0.7 (mutually exclusive – so 0.3 + 0.4)
d 0.7 (not mutually exclusive – so 0.3 + 0.5 – 0.1)

e 0.7
f 1 (mutually exclusive – so 0.5 + 0.5)
g 0.25 (uncertain relationship – so $P(Yellow|X)\ P(X) = (0.5)\ (0.5)$)
h 0 (dependency – so $P(Yellow|Red)\ P(Red) = (0)\ (0.3)$)
i 0.3 (conditional – so $P(Yellow\ and\ Y)/P(Y) = (0.15)/(0.5)$)
j 0.38 (conditional – so $P(Y\ and\ Yellow)/P(Yellow) = (0.15)/(0.4)$)
k Yes, because $P(Green)\ P(X) = P(Green|X)\ P(X)$.
l No, because $P(Yellow)\ P(Y) \neq P(Yellow|Y)\ P(Y)$; need to make $P(Yellow) = P(Yellow|Y)$.

10 Bayes' theorem 0.05(0.1)/[0.05(0.1) + 0.025(0.9)] = 0.005/[0.005 + 0.0225] = 0.182 or 0.18 or 18%

11 Bayes' theorem 0.54(0.18)/[0.54(0.18) + 0.11(0.82)] = 0.0972/[0.0972 + 0.0902] = 0.5187 or 0.52 or 52%

12 Bayes' theorem 0.44(0.57)/[0.44(0.57) + 0.36(0.43)] = 0.2508/[0.2508 + 0.1548] = 0.618 or 0.62 or 62%

13 Bayes' theorem 0.25(0.4)/[0.25(0.4) + 0.15(0.6)] = 0.1/[0.1 + 0.09] = 0.526 or 0.53 or 53%

14 0.22, basic probability formula (number of favorable events divided by total number of events; 8/36). There are two ways to get an 11 (5 and 6; 6 and 5) and 6 ways to get a seven (1 and 6; 2 and 5; 3 and 4; 4 and 3; 2 and 5; 1 and 6).

Chapter 7

1 Choose a population of scores. Decide on a sample size, n. Take a random sample of size n. Compute the mean and replace the scores back into the population. Repeat the sampling procedure, always using the same sample size, until all possible samples have been drawn. Plot the relative frequency distribution of the means.

2 The variability of the sampling distribution is determined by the population standard deviation and the size of the samples drawn. The relationship is $\sigma_M = \sigma/\sqrt{n}$. As σ increases, σ_M increases. As σ decreases, σ_M decreases.

3 Point estimation estimates a parameter as a single value. Interval estimation establishes a range of values within which the population parameter is expected to lie.

4 Estimation is an inferential procedure that uses data from a sample to infer the value of a population parameter. Hypothesis testing is a set of inferential procedures that uses data from samples to establish the credibility of a hypothesis about population parameters.

5 It is inversely related. The standard error of the mean decreases as the sample size (n) increases and vice versa. This is because a larger sample is less likely to include extreme scores that would combine to produce an extreme mean. The relationship is $\sigma_M = \sigma/\sqrt{n}$.

6 Single-sample designs use one sample to test a hypothesis about the mean of a population. Two-sample research uses two samples to test a hypothesis about the difference between two population means. These designs are experimental research methods that attempt to identify causal relations among variables. The correlational method does not attempt to exert an influence on a measured response. It cannot identify causal relations among variables; instead, it is aimed at identifying the strength of association between variables.

7 A research hypothesis is a formal statement or expectation about the outcome of a study, often specifying the expectation of a relationship between an independent and dependent variable. A statistical hypothesis is a numerical statement about the outcome of a study. The null and alternative hypotheses are statistical hypotheses.

8 The null hypothesis states that there is no effect of the independent variable on the dependent variable (no relationship between variables). The alternative hypothesis states that there is an experimental effect (a relationship between variables).

9 An example of a single-sample research project should specify an exact numerical value for the null hypothesis. The alternative hypothesis is that the population mean does not equal *that* numerical value. An example of a two-sample research project should specify an expected difference between two conditions representing two different populations (e.g. drug vs. placebo).

10 $\mu_M = 100$; $\sigma_M = 10/\sqrt{9} = 3.33$. Its shape is normal because it comes from a normally distributed population.

11 **a** H_0: $\mu_A = \mu_B$; H_1: $\mu_A \neq \mu_B$
b If the two sample means are very similar.

c Failure to reject the null means that we are uncertain about whether one drug works better than another. (It does not mean that we are certain or even fairly certain that there is no difference between the drugs.)
d If the sample mean for Drug A was markedly larger than the Drug B sample mean.
e If the sample mean for Drug B was markedly larger than the Drug A sample mean.

12 **a** H_0: $\mu = 3$ hours; H_1: $\mu \neq 3$ hours
b Finding a sample mean that was very similar to three hours.
c Finding a sample mean that was much less than three hours.
d Finding a sample mean that was much greater than three hours.

13 There is no relationship between a sample mean value and the standard error. The standard error is influenced by the standard deviation of the population of raw scores and by the sample size, $\sigma_M = \sigma/\sqrt{n}$.

14 Standard error formula is $\sigma_M = \sigma/\sqrt{n}$, so $\sigma_M = 10/\sqrt{20} = 2.24$.

15 Standard error formula is $\sigma_M = \sigma/\sqrt{n}$, so $\sigma_M = 0.5/\sqrt{100} = 0.05$.

16 Well, $\mu = \mu_M$, so $\mu_M = 20$. The rest of the information is not necessary.

17 **a** A sample mean is an unbiased estimate of the population mean, so $M = 17 \approx \mu$. A sample standard deviation is an unbiased estimate of a population standard deviation, so $s = 2 \approx \sigma$.
b Well, $s_M = s/\sqrt{n} = 2/\sqrt{20} = 0.45$.

18 **a** The research hypothesis would be that squirrels that eat the genetically modified nuts would grow to become larger squirrels.
b The null hypothesis would be that $\mu_{modified\ diet} = \mu_{normal\ diet} = 17$ *ounces*.
c The alternative hypothesis would be that $\mu_{modified\ diet} \neq \mu_{normal\ diet} = 17$ *ounces*.
d Finding a sample mean that is significantly larger than 17 ounces.
e Finding a sample mean that is significantly smaller than 17 ounces.
f Finding a sample mean that is very close to 17 ounces.

19 Since there is no way to know which scores were randomly selected, there is no way to give accurate sample means. However, for *n*'s of 5, 10, 15, and 20, the standard errors are 4.80, 3.40, 2.77, and 2.40, respectively. As the sample size increases, the standard error decreases. Moreover, as the

sample size increases, the degree of error between the sample means and the population mean should decrease.

Chapter 8

1 Depending on the particular sample drawn, we might show a treatment effect by chance – due to extreme scores being included in the sample but not because of an experimental treatment. We cannot know for sure that an effect was not due to chance since we are always dealing with probabilities. However, we can control the level of certainty with which we can claim to have an experimental effect by the level at which we set alpha.

2 No answer offered. The research hypothesis should be stated in terms of concepts and relationships – without the use of numbers but with a clear direction (more, less, stronger, weaker, etc.); the statistical hypotheses should formulate the idea into a null and alternative hypothesis (including both directions).

3 **a** H_0: $\mu = 8$; H_1: $\mu \neq 8$
b H_0: $\mu = 12$; H_1: $\mu \neq 12$
c H_0: $\mu = 20$; H_1: $\mu \neq 20$

4 We should use the *t* distribution when the population standard deviation is unknown.

5 The researcher should use the *z* test. The ability to determine the actual standard error should always be preferred to estimating it by using a sample standard deviation.

6 The relative consequences of a Type I vs. a Type II error. If a Type I error is not a particular concern, the researcher may move alpha to .10 (or 10%) to help avoid making a Type II error. If a Type I error is a major concern, the researcher may move alpha to .01 (or 1%) or even less.

7 Type I error

8 Type II error

9 Type II error

10 Type I error

11 We can actually set the precise risk rate for Type I errors (usually .05); we can merely increase or decrease the likelihood of making a Type II error.

12 Cohen's d = *mean difference*/σ = 0.5.

13 Cohen's d = *mean difference*/σ = 3/4 = 0.75.

14 Cohen's d = *mean difference*/σ. So if 0.4 = 12/σ, then σ = 30.

15 Cohen's d = *mean difference*/σ. So if 0.2 = mean difference/20, then the mean difference must = 4. If the known population mean is 100, then the publisher must be claiming that students who use the new textbook will average 104.

16
- **a** Use the z distribution since σ is known.
- **b** H_0: $\mu = 60$; H_1: $\mu \neq 60$
- **c** $z_{crit} = \pm 1.96$
- **d** $z_{obt} = \dfrac{65-60}{5/\sqrt{50}} = \dfrac{65-60}{5/\sqrt{7.07}} = 7.04$
- **e** Reject the null hypothesis.
- **f** Type I
- **g** Yes. Statistical evidence suggests that typing speed is enhanced when using the newly designed keyboard.
- **h** Effect size (Cohen's d) = 5/5 = 1.

17
- **a** Use the z distribution since σ is known.
- **b** H_0: $\mu = 100$; H_1: $\mu \neq 100$
- **c** $z_{crit} = \pm 1.96$
- **d** $z_{obt} = \dfrac{110-100}{15/\sqrt{100}} = 6.67$
- **e** Reject the null hypothesis.
- **f** Type I
- **g** Yes. Statistical evidence suggests that children of parents with a college education have IQ's that are higher than the average IQ.
- **h** Effect size (Cohen's d) = 10/15 = 0.67

18
- **a** Use the t distribution since σ is unknown.
- **b** H_0: $\mu = 90$; H_1: $\mu \neq 90$
- **c** $t_{crit} = \pm 2.021$ ($df = 40$)
- **d** $t_{obt} = \dfrac{110-90}{30/\sqrt{41}} = 4.27$

e Reject the null hypothesis.
f Type I
g Statistical evidence suggests that the administration of this hormone produces golden retrievers that are heavier than the average weight of retrievers.
h Effect size (Cohen's d) = 20/30 = 0.67

19 **a** The single-sample z test; we know sigma.
b H_0: $\mu = 7.5$; H_1: $\mu \neq 7.5$
c Skewness does not concern us in this situation; the sample size is so large that we are robust to the assumption of normality.
d Use the Internet to find $t_{crit} = \pm 1.97$ (sometimes large df's are hard to find, even on the Internet; however, a close look at the table shows that the difference between, say, 199 and 200 df is negligible.)
e $t_{obt} = \dfrac{7.2 - 7.5}{2.4/\sqrt{200}} = \dfrac{-0.3}{0.17} = -1.76$
f No, fail to reject.
g Type II
h There is no statistical evidence to suggest that students at this university sleep different amounts than university students in general.
i Not applicable – the null was not rejected.

20 **a** The t distribution, because we do not know σ.
b H_0: $\mu = 50$; H_1: $\mu \neq 50$
c ± 2.093
d $t_{obt} = \dfrac{M - \mu}{s/\sqrt{n}} = \dfrac{63 - 50}{17/\sqrt{20}} = \dfrac{13}{3.80} = 3.42$
e Yes
f Type I
g Statistical evidence suggests that people who work at home are more satisfied with their job than workers in general.
h The estimate of the effect size (Cohen's d) = *mean difference*/s = 13/17 = 0.76.

21 The random selection of participants may result in sampling error. The procedures of hypothesis testing help determine the likelihood that sampling error accounts for the experimental results (i.e. the difference between the sample mean and the null mean).

22 **a** Use the t distribution since σ is unknown.
b H_0: $\mu = 72.40$; H_1: $\mu \neq 72.40$

c $t_{crit} = \pm 2.306$ ($df = 8$)

d $t_{obt} = \dfrac{77 - 72.4}{3.1/\sqrt{9}} = 4.47$

e Reject the null hypothesis.

f Type I

g Statistical evidence suggests that biological males overestimate their life expectancy.

h The estimate of the effect size (Cohen's d) = *mean difference*/s = 4.6/3.1 = 1.48.

23 **a** $t(7) = 4.52$, $p < .05$

b $t(7) = 3.08$, $p < .05$

c $t_{crit} = \pm 2.365$ ($df = 7$)

d Statistical evidence suggests that the mean height for physically stressed biological males ($M = 69.25$) is greater than the average height for biological males, $t(7) = 4.52$, $p < .05$. Statistical evidence suggests that biological females who have been physically stressed ($M = 61.88$) also are taller than the average biological female, $t(7) = 3.08$, $p < .05$.

24 $t_{obt} = \dfrac{20 - 16}{2.8/\sqrt{8}} = 4.04$

$t_{crit} = \pm 2.365$ ($df = 7$)

Statistical evidence suggests that the mean number of publications among the faculty members of this particular sociology department is higher than the national average, $t(7) = 4.04$, $p < .05$.

25 $LL = 24.50 - 1.31 = 23.19$

$UL = 24.50 + 1.31 = 25.81$

26 $LL = 4.3 - 0.52 = 3.68$

$UL = 4.2 + 0.52 = 4.72$

27 $LL = 56\ 000 - 1\ 087.49 = 54\ 912.51$

$UL = 56\ 000 + 1\ 087.49 = 57\ 087.49$

28 Statistical evidence suggests that students who participated in the program now smoke significantly fewer cigarettes ($M = 10.34$) than the average number of cigarettes consumed by students who smoke, $t(49) = -4.17$, $p < .05$.

29 There is no statistical evidence that the average number of days to process a claim is different from 15, $t(39) = -0.83$, *n.s.*

30 If we can use 24.5 minutes as a population mean for lateness prior to the addition of extra trains, statistical evidence suggests that the addition of extra trains during rush hour reduced the amount of time the train is late, $t(29) = -5.18$, $p < .05$.

Chapter 9

1 **a**

2 Sample; population

3 **c**

4 **d**

5 **d**

6 Sampling error

7 No. However, 1 and 2 are typically used, but researchers are free to use other subscripts as well, for instance, $\mu_{control}$ and μ_{exp} or μ_{drug} and $\mu_{placebo}$ or even letters such as μ_A and μ_B.

8 **b**

9 Representativeness of the two populations would need to be achieved – those drivers who do not have voice recognition on their cell phones as well as those who do. Independent observations would probably not be a problem. The data gathered in terms of number of accidents, time to react to a stimulus, or whatever was being measured would need to be on an interval or ratio scale. Normality might be a concern unless the sample sizes were quite large. An analysis on the sample variances might be done to make sure they are somewhat similar.

10 A pooled standard deviation is simply the square root of the pooled variance. (The pooled variance is the weighted average of the two sample variances.) If the pooled standard deviation is needed, simply take the square root of the pooled variance. If the pooled variance is needed, simply square the pooled standard deviation.

11

	Critical values	Decision
a.	±2.160	Fail to reject
b.	±2.663 (online table)	Reject
c.	±1.746	Reject
d.	±2.064	Fail to reject

12 **a** H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$

b The appropriate inferential test is the independent-samples *t* test.

No siblings	Siblings
$M_1 = 7.33$	$M_2 = 3.17$
$s_1 = 2.16$	$s_2 = 2.14$
$n_1 = 6$	$n_2 = 6$

Using the computational formula,

$$t_{obt} = \frac{7.33 - 3.17}{\sqrt{\left\{\left[(346 - (44)^2/6) + (83 - (19)^2/6)\right]/(6+6-2)\right\}(1/6 + 1/6)}}$$

$t_{obt} = 3.33$

c $t_{crit} = \pm 2.228$ ($df = 10$)

d Reject the null hypothesis.

e First, we need to find pooled standard deviation (square root of the pooled variance). $\sqrt{s_p^2} = \sqrt{\frac{2.16^2(5) + 2.14^2(5)}{6+6-2}} = \sqrt{\frac{23.22 + 22.9}{10}} = \sqrt{4.61} = 2.15$.

The estimate effect size (Cohen's *d*) = *estimated mean difference*/*estimated s* = 4.16/2.15 = 1.93.

f Type I

g Statistical evidence suggests that two-year-olds with siblings have less fear than two-year-olds with no siblings, $t(10) = 3.33$, $p < .05$.

13 **a** H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$

b The appropriate inferential test is the independent-samples *t* test.

$$t_{obt} = \frac{4.2 - 2.2}{\sqrt{[(0.5(9) + 0.7(9))/(10+10-2)](1/10 + 1/10)}}$$

$t_{obt} = 5.71$

c $t_{crit} = \pm 2.101$ ($df = 18$)

d Reject the null hypothesis.

e First, we need to find pooled standard deviation (square root of the pooled variance). $\sqrt{s_p^2} = \sqrt{\frac{0.5(9) + 0.7(9)}{10 + 10 - 2}} = \sqrt{\frac{4.5 + 6.3}{18}} = \sqrt{0.6} = 0.77$. The estimate effect size (Cohen's d) = *estimated mean difference*/*estimated* $s = 2/0.77 = 2.60$.

f Type I

g Statistical evidence suggests that among biological males, a high level of anxiety leads to greater attraction toward biological females than a low level of anxiety, $t(18) = 5.71$, $p < .05$.

14

a $$t_{obt} = \frac{4.2 - 2.2}{\sqrt{[(5.2(9) + 5.4(9))/(10 + 10 - 2)](1/10 + 1/10)}}$$

$t_{obt} = 1.94$

b With $t_{crit} = \pm 2.101$, the t_{obt} of 1.94 does not lead to rejection of the null hypothesis.

Increasing the variability of scores has increased the size of the denominator of the t ratio and reduced the size of t_{obt}.

15

a $$t_{obt} = \frac{4.2 - 2.2}{\sqrt{[(5.2(29) + 5.4(29))/(30 + 30 - 2)](1/30 + 1/30)}}$$

$t_{obt} = 3.51$

b With $t_{crit} = \pm 2.002$ (use an online t table to find t_{crit}), the t_{obt} of 3.51 leads to a rejection of the null hypothesis.

c Increasing the sample size decreases t_{crit} and more importantly increases t_{obt} by shrinking the estimate of the standard error (the denominator).

16 The standard deviation of the sampling distribution of differences between the means is the standard error of the difference.

17 As the sample size increases, df increases, and t_{crit} decreases correspondingly. As the sample size increases, the t distribution approaches the standard normal curve, which means the tails pull in toward zero. Consequently, the values that mark the outermost 5% are closer to 0. This increases the chance for a research situation in which a false null will produce a t_{obs} that will fall in the rejection region.

18 As question 4 shows, increasing the sample size will decrease the size of the denominator of the t ratio by decreasing the estimate of the standard error of the difference and result in a larger t_{obt}. In other words, a difference of a

given amount between means is made to look much more substantial if the error term associated with the inferential test is small. Obviously, large t_{obs} values increase the chances of rejecting the null hypothesis.

19 **a** H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$

b The appropriate inferential test is the independent-samples *t* test.

Males	Females
$M_1 = 15.60$	$M_2 = 9.0$
$s_1 = 4.16$	$s_2 = 3.61$
$n_1 = 5$	$n_2 = 5$

Using the computational formula,

$$t_{obt} = \frac{15.6 - 9.0}{\sqrt{\{[(1286 - 1216.8) + (457 - 405)]/(5 + 5 - 2)\}(1/5 + 1/5)}}$$

$t_{obt} = 2.68$

c $t_{crit} = \pm 2.306$ (with $df = 8$)

d Yes, reject the null.

e First, we need to find pooled standard deviation (square root of the pooled variance). $\sqrt{s_p^2} = \sqrt{\frac{4.16^2(4) + 3.61^2(4)}{5 + 5 - 2}} = \sqrt{\frac{69.22 + 52.13}{8}} = \sqrt{15.17} = 3.89$.

The estimate effect size (Cohen's *d*) = *estimated mean difference*/*estimated s* = 6.6/3.89 = 1.70.

f Type I

g Statistical evidence suggests that biological male college students report more anger reactions than biological female college students, $t(8) = 2.68$, $p < .05$.

20 **a** H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$

b The appropriate inferential test is the independent-samples *t* test.

Teachers	Principals
$M_1 = 41.33$	$M_2 = 32.50$
$s_1 = 5.43$	$s_2 = 6.19$
$n_1 = 6$	$n_2 = 6$

Using the computational formula,

$$t_{obt} = \frac{41.33 - 32.50}{\sqrt{\{[(10398 - 10250.67) + (6529 - 6337.50)]/(6 + 6 - 2)\}(1/6 + 1/6)}}$$

$t_{obt} = 2.60$

c $t_{crit} = \pm 2.228$ ($df = 10$)

d Yes

e First, we need to find pooled standard deviation (square root of the pooled variance). $\sqrt{s_p^2} = \sqrt{\frac{5.43^2(5) + 6.19^2(5)}{6+6-2}} = \sqrt{\frac{147.42 + 191.58}{10}} = \sqrt{33.9} = 5.82$. The estimate effect size (Cohen's d) = *estimated mean difference*/*estimated* s = 8.83/5.82 = 1.52.

f Type I

g Statistical evidence suggests that teachers experience more burnout than principals, $t(10) = 2.60$, $p < .05$.

21 **a** H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$

b The appropriate inferential test is the independent-samples t test.

Lonely	Not Lonely
$M_1 = 5.00$	$M_2 = 7.40$
$s_1 = 1.58$	$s_2 = 1.52$
$n_1 = 5$	$n_2 = 5$

Using the computational formula,

$$t_{obt} = \frac{5 - 7.4}{\sqrt{\{[(135-125)+(283-273.8)]/(5+5-2)\}(1/5+1/5)}}$$

$t_{obt} = -2.45$

c $t_{crit} = 2.306$ (with $df = 8$)

d Reject the null hypothesis.

e First, we need to find pooled standard deviation (square root of the pooled variance). $\sqrt{s_p^2} = \sqrt{\frac{1.58^2(4) + 1.52^2(4)}{5+5-2}} = \sqrt{\frac{9.99 + 9.24}{8}} = \sqrt{2.4} = 1.55$. The estimate effect size (Cohen's d) = *estimated mean difference*/*estimated* s = −2.4/1.55 = −1.54 or just 1.54 (recall that negative effect sizes do not need to be reported as negatives).

f Type I

g Statistical evidence suggests that lonely biological males are rated as less attractive than biological males who are not lonely, $t(8) = -2.45$, $p < .05$.

22 **a** H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$

b The appropriate inferential test is the independent-samples t test.

Buffalo Creek	Kopperston
$M_1 = 43.83$	$M_2 = 37.50$
$s_1 = 4.67$	$s_2 = 1.87$
$n_1 = 6$	$n_2 = 6$

Using the computational formula,

$$t_{obt} = \frac{43.83 - 37.50}{\sqrt{\{[(11637 - 11528.17) + (8455 - 8437.50)]/(6 + 6 - 2)\}(1/6 + 1/6)}}$$

$t_{obt} = 3.06$

c $t_{crit} = \pm 2.228$ ($df = 8$)

d Reject the null hypothesis.

e First, we need to find pooled standard deviation (square root of the pooled variance). $\sqrt{s_p^2} = \sqrt{\frac{4.67^2(5) + 1.87^2(5)}{6 + 6 - 2}} = \sqrt{\frac{109.04 + 17.48}{10}} = \sqrt{12.65} = 3.56$. The estimate effect size (Cohen's d) = *estimated mean difference*/*estimated* $s = 6.33/3.56 = -1.78$.

f Type I

g Statistical evidence suggests that residents of Buffalo Creek experience higher trait anxiety than residents of Kopperston, $t(10) = 3.06$, $p < .05$.

h Recall that one of the assumptions of the t test is that the population variances are equal. In this case, $s_1^2 = 21.81$ and $s_2^2 = 3.50$. We might wonder if perhaps the assumption of homogeneity of variances is violated here. Methods for testing whether two variances are significantly different and are discussed in more advanced statistics books.

23 **a** *Males:*

$$t_{obt} = \frac{23 - 16}{\sqrt{[(61.47(9) + 41.34(14))/(10 + 15 - 2)]/(1/10 + 1/15)}}$$

$t_{obt} = 2.42$

$t_{crit} = \pm 2.069$ (with $df = 23$)

Statistical evidence suggests that first-born biological males are more narcissistic than later-born biological males, $t(23) = 2.42$, $p < .05$.

b *Females:*

$$t_{obt} = \frac{17 - 12}{\sqrt{[(42.51(18) + 43.16(27))/(19 + 28 - 2)]/(1/19 + 1/28)}}$$

$t_{obt} = 2.55$

$t_{crit} = \pm 2.01$ (used Internet source to find complete t table)

Statistical evidence suggests that first-born biological females are more narcissistic than later-born biological females, $t(45) = 2.55$, $p < .05$.

24 Here is the data according to Formula 9.8:

For biological males:

$$LL = (23 - 16) - 2.069(2.89) = 7 - 5.98 = 1.02$$
$$UL = (23 - 16) + 2.069(2.89) = 7 + 5.98 = 12.98$$

For biological females:

$$LL = (17 - 12) - 2.021(1.96) = 5 - 3.96 = 1.04$$
$$UL = (17 - 12) + 2.021(1.96) = 5 + 3.96 = 8.96$$

25 **a** H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$

b The appropriate inferential test is the independent-samples t test:

$$t_{obt} = \frac{41 - 47.2}{\sqrt{\frac{4^2(15-1) + 5^2(14-1)}{15 + 14 - 2}\left(\frac{1}{15} + \frac{1}{14}\right)}}$$
$$= \frac{-6.2}{\sqrt{\frac{224 + 325}{27}(0.14)}}$$
$$= \frac{-6.2}{\sqrt{20.33(0.14)}}$$
$$t = \frac{-6.2}{1.69}$$
$$t_{obt} = -3.68$$

c $t_{crit} = \pm 2.052$ (with $df = 27$)

d Reject the null hypothesis.

e First, we need to find pooled standard deviation (square root of the pooled variance). $\sqrt{s_p^2} = \sqrt{\frac{4^2(14) + 5^2(13)}{15 + 14 - 2}} = \sqrt{\frac{224 + 325}{27}} = \sqrt{20.33} = 4.51$. The estimate effect size (Cohen's d) = *estimated mean difference*/*estimated* s = −6.2/4.51 = −1.37 or just 1.37 (recall that negative effect sizes do not need to be reported as negatives).

f Type I

g Statistical evidence suggests that people with 5 or more negative life experiences in the last five years have higher measures of subjective well-being than participants with 2 or fewer negative experiences, $t(27) = -3.68$, $p < .05$.

26 Here is the data according to Formula 9.8:

$$LL = (41 - 47.2) - 2.052(1.69) = -9.67$$
$$UL = (41 - 47.2) + 2.052(1.69) = -2.73$$

The negative values mean an increase from the mean for the "2 or less group" to the mean for the "5 or more group." (If the order of the means had been switched – subtracting "2 or less" from "5 or more" – the mean difference would be positive.) So, dropping the negative values, it looks as if the subjective well-being score increases somewhere between 2.73 units and 9.67 units for those who have had 5 or more negative life experiences in the past five years (compared with those who have had 2 or less).

27 **a** H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$

b $t_{obt} = \dfrac{17.0 - 13.5}{2} = 1.75$

c $t_{crit} = \pm 2.002$ (used online table)

d No, fail to reject the null.

e There is no statistical evidence suggesting that children who receive training in starting conversations spend a different amount of time interacting with peers than those who do not, $t(58) = 1.75$, *n.s.*

28 **a** H_0: $\mu_1 > \mu_2$; H_1: $\mu_1 \leq \mu_2$ (where 1 is the experimental group and 2 is the control group).

b (Same as Problem 14b) $t_{obt} = 1.75$

c $t_{crit} = 1.672$ (used online table). Now, using a one-tailed test, the null hypothesis would be rejected. *Conclusion*: Statistical evidence suggests that children who receive training in starting conversations spend more time interacting with peers than children who do not, $t(58) = 1.75$, $p < .05$.

d No. There is no compelling reason for using a one-tailed test. Indeed, a finding in the opposite direction (that children who receive training engage in *less* peer interaction) would certainly be theoretically, if not practically, important to know.

29 No answer is provided here; however, it would have to be based on practical as opposed to theoretical grounds.

30 **a** H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$

b The appropriate inferential test is the independent-samples *t* test:

$$t_{obt} = \frac{80 - 69}{\sqrt{\dfrac{106.09(15-1) + 156.25(15-1)}{15 + 15 - 2}\left(\dfrac{1}{15} + \dfrac{1}{15}\right)}}$$

$$= \frac{11}{\sqrt{\frac{1485.26 + 2187.5}{28}(0.13)}}$$

$$= \frac{11}{\sqrt{131.17(0.13)}}$$

$$= \frac{11}{\sqrt{17.49}}$$

$t_{obt} = 2.63$

c $t_{crit} = \pm 2.048$ (with $df = 28$)

d Reject the null hypothesis.

e First, we need to find pooled standard deviation (square root of the pooled variance). $\sqrt{s_p^2} = \sqrt{\frac{106.09(14) + 156.25(14)}{15 + 15 - 2}} = \sqrt{\frac{1485.26 + 2187.5}{28}} = \sqrt{131.17}1.45$. The estimate effect size (Cohen's d) = *estimated mean difference*/*estimated* s = 11/11.45 = 0.96.

f Type I

g Statistical evidence suggests that students who study in one place perform better than students who rotate their study location, $t(28) = 2.63$, $p < .05$.

h A potential confound might be that not only was the variable of single vs. various locations manipulated but also the variable of quiet vs. noisy. The degree of ambient noise present when studying might compete with the number of locations as an explanation for the statistical difference.

31 Representativeness.

32 Homogeneity of variance.

33 $t_{obt} = 2.06$, t_{crit} for 22 degrees of freedom = ±2.07 for a two-tailed test. There is no evidence to suggest that a new t-shirt generates more happiness (subjective well-being) in students, $t(22) = 2.06$, *n.s.* The assumption of interval or ratio data may have been violated. Many researchers feel that Likert scale data is ordinal. Perhaps another way of measuring subjective well-being should be considered – or a different inferential test should be run (see Chapter 18).

34 $t(48) = 1.85$, *n.s.*; $t_{crit} = 2.011$ (used online table) for a two-tailed test. For a one-tailed test, $t_{crit} = 1.677$. Therefore, the null hypothesis would be rejected.

35 Statistical evidence suggests that participants in the low-arousal condition show greater approach behavior than participants in the high-arousal condition, $t(58) = 2.77$, $p < .05$.

36 $t_{obt} = 2.01$, t_{crit} for 58 degrees of freedom = ±2.01. Reject the null hypothesis of no difference. Statistical evidence suggests that highly creative people are more likely to be dishonest on this reporting task than low creative people, $t(58) = 2.01$, $p < .05$.

Chapter 10

1 b

2 a

3 Because the dependent-samples t test uses the difference score between a participant's two responses.

4 Responses will vary. The critical issue is that the conditions and/or materials used are split such that half of the participants experience one first and then the other, while the other half experience them in the opposite order.

5 Responses will vary. The critical issue is that a variable not measured in the actual study needs to be controlled by premeasurement and then matching.

6 b

7 s_D is the symbol for the standard deviation of the difference scores; $s_{\bar{D}}$ is the symbol for the estimate of the standard error of the differences between means. The first is a measure of dispersion for the sample of difference scores found from the actual data gathered; the second is an estimate of dispersion for the theoretical sampling distribution used by the dependent-samples t test. They are not the same, although s_D is used to calculate $s_{\bar{D}}$.

8 Almost all null hypotheses specify no difference between the two conditions; therefore, $\mu_x - \mu_y$ equals 0 and can be dropped in the numerator of Formula 10.3. If the null is configured to result in a difference between μ_x and μ_y, then Formula 10.3 must be used.

9 a

10 **d**

11 **a**

12 **b**

13 **a** H_0: $\mu_{bearded} = \mu_{non\text{-}bearded}$; H_1: $\mu_{bearded} \neq \mu_{non\text{-}bearded}$

b $M_x = 7.63$; $M_y = 5.63$; $\Sigma D = 16$; $\Sigma D^2 = 62$

$$t_{obt} = \frac{7.63 - 5.63}{0.73} = \frac{2.0}{0.73} = 2.74$$

c $df = n_p - 1 = 7$; $t_{crit} = \pm 2.365$

d Yes, reject.

e Mean difference = 2.0; standard deviation of the difference scores = 2.07. The estimate effect size (Cohen's *d*) = *estimated mean difference*/*estimated s* = 2/2.07 = 0.97.

f Type I

g Statistical evidence suggests that bearded men are perceived to be more masculine than nonbearded men, $t(7) = 2.74$, $p < .05$.

h The assumption of interval or ratio data may have been violated. Many researchers feel that Likert scale data is ordinal. Perhaps another way of measuring masculinity should be considered – or a different inferential test should be run (see Chapter 18).

14 Here is the data according to Formula 10.6:

$$LL = (7.62 - 5.62) - 2.365\left(2.07/\sqrt{8}\right) = 2 - 1.73 = 0.27$$

$$UL = (7.62 - 5.62) + 2.365\left(2.07/\sqrt{8}\right) = 2 + 1.73 = 3.73$$

15 **a** H_0: $\mu_{sex} = \mu_{no\ sex}$; H_1: $\mu_{sex} \neq \mu_{no\ sex}$

b $M_x = 5.0$; $M_y = 2.83$; $\Sigma D = 13$; $\Sigma D^2 = 35$

$$t_{obt} = \frac{5.0 - 2.83}{0.48} = \frac{2.17}{0.48} = 4.52$$

c $df = n_p - 1 = 5$; $t_{crit} = \pm 2.571$

d Yes, reject.

e Mean difference = 2.17; standard deviation of the difference scores = 1.17. The estimate effect size (Cohen's *d*) = *estimated mean difference*/*estimated s* = 2.17/1.17 = 1.85.

f Type I

g Statistical evidence suggests that people are more likely to purchase liquor products that use sexual symbolism in their advertising than those not using sexual symbolism, $t(5) = 4.52$, $p < .05$.

16 **a** H_0: $\mu_{Gouda} = \mu_{Swiss}$; H_1: $\mu_{Gouda} \neq \mu_{Swiss}$

b $M_x = 7.0$; $M_y = 5.60$; $\Sigma D = 7$; $\Sigma D^2 = 35$

$$t_{obt} = \frac{7.0 - 5.60}{1.12} = \frac{1.40}{1.12} = 1.25$$

c $df = n_p - 1 = 4$; $t_{crit} = \pm 2.776$

d No, fail to reject.

e No need for an effect size analysis – the null was not rejected.

f Type II

g There is no statistical evidence that college students have a preference between Gouda and Swiss cheese, $t(4) = 1.25$, *n.s.*

h The assumption of interval or ratio data may have been violated. Many researchers feel that Likert scale data is ordinal. Perhaps another way of measuring cheese preference should be considered – or a different inferential test should be run (see Chapter 18).

i The researcher should counterbalance exposure to the cheeses.

17 Here is the data according to Formula 10.6:

$$LL = (7.0 - 5.60) - 2.776\left(\frac{2.51}{\sqrt{5}}\right) = 1.4 - 3.11 = -1.71$$

$$UL = (7.0 - 5.60) + 2.776\left(\frac{2.51}{\sqrt{5}}\right) = 1.4 + 3.11 = 4.51$$

18 The dependent-samples *t* test increases the power of an experiment. That is, the probability of correctly rejecting a false null hypothesis is increased as a result of reducing the variability due to individual differences. This makes the *t* test denominator smaller, thereby making the resulting t_{obt} larger and more likely to fall into a rejection region.

19 **a** H_0: $\mu_x = \mu_y$; H_1: $\mu_x \neq \mu_y$ (X = Pre; Y = Post)

b $M_x = 92.50$; $M_y = 45.0$; $\Sigma D = 190$; $\Sigma D^2 = 12\ 500$

$$t_{obt} = \frac{92.50 - 45.0}{17.015} = 2.79$$

c $df = n_p - 1 = 3$; $t_{crit} = \pm 3.182$

d No, fail to reject.

e No need for an effect size analysis – the null was not rejected.

f Type II

g There is no statistical evidence that the new drug for insomnia decreases the amount of time needed to fall asleep, $t(3) = 2.79$, *n.s.*

20 **a** H_0: $\mu_{story} = \mu_{music}$; H_1: $\mu_{story} \neq \mu_{music}$

b $M_{story} = 8.43$; $M_{music} = 6.57$; $\Sigma D = 13$; $\Sigma D^2 = 85$

$$t_{obt} = \frac{8.43 - 6.57}{1.204} = 1.54$$

c $df = n_p - 1 = 6$; $t_{crit} = \pm 2.447$

d No, fail to reject.

e No need for an effect size analysis – the null was not rejected.

f Type II

g No statistical evidence was found to suggest that a difference exists between being read a story and listening to music as sleep inducers for preschoolers, $t(6) = 1.54$, *n.s.*

21 There are two. The most obvious difference is the assumption regarding independent observations. In repeated-measures designs, obviously not ALL observations are independent of each other – each person is contributing multiple scores. However, it is important that scores WITHIN a given condition are all independent of each other. Secondly, the assumption of normality is slightly different. Instead of saying that both populations of raw scores are normally distributed, in a dependent-samples *t* test situation, the assumption of normality refers to the distribution of difference scores.

22 Statistical evidence suggests that students write papers of higher quality when using the PC computer instead of the Mac, $t(19) = 3.05$, $p < .05$.

23 There is no statistical evidence to suggest that vision is differentially affected by lens color, $t(15) = 0.17$, *n.s.*

24 Statistical evidence suggests that participants experience more back pain when sleeping on a soft mattress compared with a firm mattress, $t(11) = -5.52$, $p < 0.01$.

25 There is statistical evidence to suggest that customers prefer the free music streaming feature more than the free basic TV feature, $t(14) = 2.29$, $p < .05$. There may be a problem with the assumption of interval and/or ratio scaling. Many researchers feel Likert scale data should be best understood as ordinal.

26 There was no statistical evidence found to suggest that seating behavior (same seat vs. switching seats) influences student performance on psychology quizzes, $t(18) = 1.95$, *n.s.*

Chapter 11

1 **d**

2 The lenses of a microscope are also described to vary due to power. Just as low-powered lenses may not see small objects (germs) that high-powered lenses can see, so studies with low power may not be able to detect small treatment effect sizes as well as those studies with high power.

3 **b**

4 Nothing. The Type I error rate is determined by the selected alpha value. Power considerations do not influence the Type I error rate.

5 Power is inversely related to the Type II error rate. As power increases, the Type II error rate decreases.

6 The hypothesized treatment effect size is determined by dividing the difference between the null mean and the hypothesized mean by the standard error. It can be represented mathematically as follows: $\gamma = \frac{\mu_{alt} - \mu_0}{\sigma}$.

7 The size of the treatment effect influences power directly. As it increases, so does power.

8

a $\gamma = \frac{345 - 300}{70} = 0.64$

b $\gamma = \frac{345 - 300}{20} = 2.25$

c $\gamma = \frac{310 - 300}{20} = 0.50$

d $\gamma = \frac{310 - 300}{50} = 0.20$

9 The researcher may attempt to increase the size of the treatment effect by extending the length of sleep deprivation from three hours to something more than three hours. Theoretically, it would be presumed that the relationship between sleep deprivation and cognitive functioning is such that

increased sleep deprivation decreases cognitive functionality. The only other option would be to try to decrease the standard deviation. It is not clear how a researcher could do that.

10 The sample size influences power directly. As it increases, so does power.

11

a $n = \left(\frac{2.8}{0.64}\right)^2 = 19$

b $n = \left(\frac{2.8}{2.25}\right)^2 = 2$

c $n = \left(\frac{2.8}{0.5}\right)^2 = 31$

d $n = \left(\frac{2.8}{0.2}\right)^2 = 196$

12

a	$\gamma = \frac{120-130}{15} = -0.67$	$\delta = -0.67\left(\sqrt{10}\right) = -2.12$	Power = 0.56
b	$\gamma = -0.67$	$\delta = -0.67\left(\sqrt{40}\right) = -4.24$	Power = 0.99
c	$\gamma = \frac{52-50}{10} = 0.20$	$\delta = 0.20\left(\sqrt{15}\right) = 0.77$	Power = 0.13
d	$\gamma = 0.20$	$\delta = 0.20\left(\sqrt{100}\right) = 2.00$	Power = 0.52
e	$\gamma = \frac{30-25}{7} = 0.71$	$\delta = 0.71\left(\sqrt{30}\right) = 3.89$	Power = 0.97

13 The value of δ, for a desired power of 0.80 with $\alpha = 0.05$ and two-tailed test, = 2.8

a $n = \left(\frac{2.8}{0.67}\right)^2 = 17.47$, about 17 participants.

c $n = \left(\frac{2.8}{0.2}\right)^2 = 196$, 196 participants.

e $n = \left(\frac{2.8}{0.71}\right)^2 = 15.55$, about 16 participants.

14 **b** Sampling error. The variability of raw scores for any measure cannot usually be influenced without altering the sampling method (e.g. only accepting participants into the study whose score falls into a prescribed range based on a premeasure).

15 Increasing alpha increases power by increasing the rejection region of the hypothesis test. Decreasing alpha correspondingly decreases power by decreasing the rejection region.

16 Compared with a two-tailed test, a one-tailed test increases alpha for predicted end of the sampling distribution. This, in effect, is like increasing alpha.

17 It could be hypothesized that smokers experience more stress than nonsmokers. Assume that we ran an experiment using 150 participants, one-tailed test, $\alpha = 0.05$, searched for a medium effect size (0.25), and found no significant difference between groups. We could argue that the power of our test was 0.91, a 91% chance of correctly rejecting H_0 if it were false (given the stated effect size). We could further argue that although there might be a small difference between smokers' and nonsmokers' stress, the effect is trivial and not worth instituting a stress-reduction treatment program.

Part 4. Review of *z* Tests, *t* Tests, and Power Analyses

1 A power analysis. Although power analyses can be run after data has been gathered or after a pilot study has been run, it can also be performed before any data has been gathered.

2
- **a** $t_{crit} = \pm 2.093$
- **b** $t_{crit} = \pm 1.746$
- **c** $z_{crit} = \pm 2.33$
- **d** $t_{crit} = \pm 1.895$
- **e** $t_{crit} = \pm 2.626$
- **f** $t_{crit} = \pm 3.335$
- **g** $t_{crit} = \pm 2.776$
- **h** t_{crit} = either 1.86 or −1.86, depending on the predicted direction.
- **i** $z_{crit} = \pm 1.96$
- **j** $t_{crit} = \pm 1.699$
- **k** $t_{crit} = \pm 2.771$
- **l** $t_{crit} = \pm 9.925$

3
- **a** H_0: $\mu = 2.6$; H_1: $\mu \neq 2.6$
- **b** The single-sample t test. There is one sample being compared with a given population mean and σ is not given.
- **c** $t_{obt} = -6.36$
- **d** $df = 17$, so $t_{crit} = \pm 2.11$
- **e** Yes
- **f** The estimate of the effect size (Cohen's d) = *mean difference*/s = 1.08/0.72 = 1.5.
- **g** Not applicable – null was rejected.

h Type I
i There is statistical evidence to suggest that children in this rural area do not play with friends as much as children in general, $t(17) = -6.36$, $p < .05$.

4 **a** H_0: $\mu_{85^\circ} = \mu_{65^\circ}$; H_1: $\mu_{85^\circ} \neq \mu_{65^\circ}$
b The dependent-samples t test. There are two measures from one sample – it is a repeated-measures design.
c $t_{obt} = -2.75$
d $df = 4$, so $t_{crit} = \pm 2.78$
e No
f Not applicable – failing to reject the null.
g $\delta = 1$; Table A.3 says the power is 0.17 or 17%.
h Type II
i There is no statistical evidence to suggest that people experience a different number of dreams depending upon the temperature of the room in which they are sleeping, $t(4) = 2.75$, *n.s.*

5 **a** H_0: $\mu = 110$; H_1: $\mu \neq 110$
b The single-sample z test. There is one sample being compared with a given population mean and σ is given (15).
c $z_{obt} = 5/4.33 = 1.15$
d $z_{crit} = \pm 1.96$
e No
f Not applicable – failing to reject the null.
g $\delta = 1.7$; Table A.3 says the power is 0.40 or 40%.
h Type II
i [Writing up a z test finding is not presented in the textbook, primarily because z tests are rarely found in the scientific literature. However, the following is an appropriate sentence.] There is no statistical evidence to suggest that this class of statistics students is more intelligent than most of the psychologist's previous classes, $z = 1.15$, *n.s.*

6 **a** H_0: $\mu = 50$; H_1: $\mu \neq 50$
b The single-sample t test. There is one sample being compared with a given population mean and σ is not given.
c $t_{obt} = -3.61$
d $df = 14$, so $t_{crit} = \pm 2.15$
e Yes
f The estimate of the effect size (Cohen's d) = *mean difference*/s = 4/4.29 = 0.93.
g Not applicable – null was rejected.
h Type I

i There is statistical evidence to suggest that people who experienced 4 or more moves prior to the age of 12 have lower subjective well-being scores than others, $t(14) = -3.61$, $p < .05$.

7 **a** H_0: $\mu_{Steroid} = \mu_{G.S.}$; H_1: $\mu_{Steroid} \neq \mu_{G.S.}$
b The independent-samples t test. There are two measures from two different and independent samples.
c $t_{obt} = 2.30$
d $df = 14$; $t_{crit} = \pm 2.15$
e Yes
f First, we need to find pooled standard deviation (square root of the pooled variance). $\sqrt{s_p^2} = \sqrt{\frac{1.85^2(7) + 2.27^2(7)}{8+8-2}} = \sqrt{\frac{23.95 + 36.07}{14}} = \sqrt{4.28} = 2.07$. The estimate effect size (Cohen's d) = *estimated mean difference*/*estimated* $s = 2.38/2.07 = 1.14$.
g Not applicable – null was rejected.
h Type I
i There is statistical evidence to suggest that the use of this synthetic anabolic steroid leads to greater weight gain than the growth stimulant, $t(14) = 2.30$, $p < .05$.

8 **a** H_0: $\mu_{Steroid} = \mu_{G.S.}$; H_1: $\mu_{Steroid} \neq \mu_{G.S.}$
b The dependent-samples t test. There are two measures from one sample – it is a repeated-measures design.
c $t_{obt} = 2.73$
d $df = 7$; $t_{crit} = 2.365$
e Yes
f The estimate effect size (Cohen's d) = *estimated mean difference*/*estimated* $s = 2.38/2.45 = 0.97$.
g Not applicable – null was rejected.
h Type I
i There is statistical evidence suggesting that the anabolic steroid leads to more weight gain than the growth stimulant, $t(7) = 2.73$, $p < .05$.

9 **a** H_0: $\mu = 5$; H_1: $\mu \neq 5$
b The single-sample z test. There is one sample being compared with a given population mean and σ is given (30 seconds or 0.5 minutes).
c $z = 0.2/0.18 = 1.11$
d $z_{crit} = \pm 1.96$
e No
f Not applicable – failing to reject the null.
g $\delta = 1.4$; Table A.3 says the power is 0.29 or 29%.

h Type II

i [Writing up a z test finding is not presented in the textbook, primarily because z tests are rarely found in the scientific literature. However, the following is an appropriate sentence.] There is no statistical evidence to suggest that this coach's soccer team is more fit than biological male collegiate student/athletes in general, $z = 1.11$, *n.s.*

10 **a** H_0: $\mu_{angry} = \mu_{control}$; H_1: $\mu_{angry} \neq \mu_{control}$

b The independent-samples t test. There are two measures from two different and independent samples.

c $t_{obt} = 1.57$

d $df = 10$; $t_{crit} = \pm 2.23$

e No

f Not applicable – failing to reject the null.

g $\delta = 1.6$; Table A.3 says the power is 0.36 or 36%.

h Type II

i There is no statistical evidence to suggest that people who are angry make more mistakes while operating a video game car than those who are not angry, $t(10) = 1.57$, *n.s.*

11 This question involves data from three conditions – we have not yet learned what statistical tools can be used to analyze data in this type of research design. See Chapter 12, question 1, in the next part of the text for the answers.

Chapter 12

1 ANOVA

2 Two; independent of

3 An ANOVA avoids the inflated alpha problem that comes with multiple t tests.

4 Treatment variance, if the situation is experimental; primary variance more generally.

5 Error variance or secondary variance

6 MS_{BG}

7 MS_W

8 k

9 One measurement of variance, MS_{BG}, incorporates primary variance as well as error variance. The other measurement of variance, MS_W, only incorporates error variance. By comparing them in ratio form (MS_{BG} over MS_W), primary variance can be evidenced by a larger than 1 result.

10 It is a family of positively skewed distributions that crest at the value of 1. We might even say it has a mode of 1.

11 Independent-samples t test.

12 Normality; homogeneity of variance.

13 ANOVA summary table

14

Source	SS	df	MS	F	p
Between groups	280.3	3	94.43	12.83	< .05
Within groups	247.68	34	7.28		
Total	527.98	37			

Reject the null: F_{obs} of 12.83 exceeds F_{crit} of 2.88.

15

Source	SS	df	MS	F	p
Between groups	5.88	4	1.47	2.59	*n.s.*
Within groups	6.30	11	0.57		
Total	12.18	15			

Fail to reject the null: F_{obs} of 2.59 does not equal or exceed F_{crit} of 3.66.

16 **a** H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: at least two of the means are different.

Aerobics	Circuit	Control	Summary values
$\Sigma X_1 = 243$	$\Sigma X_2 = 273$	$\Sigma X_3 = 313$	$\Sigma X = 829$
$n_1 = 4$	$n_2 = 4$	$n_3 = 4$	$\Sigma X^2 = 58\,115$
$M = 60.75$	$M = 68.25$	$M = 78.25$	$(\Sigma X)^2 = 687\ 241$
			$N = 12,\ \ k = 3$

b $SS_{BG} = \frac{243^2}{4} + \frac{273^2}{4} + \frac{313^2}{4} - \frac{829^2}{12}$
$= 57\,886.75 - 57\,270.08 = \mathbf{616.67}$

c $SS_W = 58\,115 - \frac{243^2}{4} + \frac{273^2}{4} + \frac{313^2}{4} = \mathbf{228.25}$

d $df_{BG} = k - 1 = 3 - 1 = 2$

e $df_W = N - k = 12 - 3 = 9$

f $MS_{BG} = \frac{616.67}{2} = \mathbf{308.34}$

g $MS_W = \frac{228.25}{9} = \mathbf{25.36}$

h $SS_T = 58115 - \frac{829^2}{12} = \mathbf{844.92}$

i $df_T = N - 1 = df_{BG} + df_W = 11$

j $F = \frac{308.34}{25.36} = \mathbf{12.16}$

k $F_{crit}\,(2,9) = 4.26$ (for $\alpha = 0.05$)

l Reject H_0.

m

Source of variation	*SS*	*df*	*MS*	*F*	*p*
Between groups	616.17	2	308.34	12.16	< .05
Within groups (error)	228.25	9	25.36		
Total	844.92	11			

n $\varpi^2 = \frac{616.67 - 2(25.36)}{844.92 + 25.36} = 0.65$

65% of the variance in heart rate is accounted for by the levels of the independent variable.

o (t_{crit} for all t's = 2.201, df = 11, = .05)

Aerobics vs. circuit:

$$t = \frac{68.25 - 60.75}{\sqrt{25.36\left(\frac{1}{4} + \frac{1}{4}\right)}} = 2.11\ n.s.$$

Aerobics vs. control:

$$t = \frac{78.25 - 60.75}{3.56} = 4.92\ (p < 0.05)$$

Circuit vs. control:

$$t = \frac{78.25 - 68.25}{3.56} = 2.81\ (p < 0.05)$$

There is statistical evidence that both the aerobic and circuit training conditions are superior to the control condition, $t(11) = 4.92$, $p < .05$ and

$t(11) = 2.81$, $p < .05$, respectively. There is no statistical evidence of significant difference between aerobic and circuit training, $t(11) - 2.11$, *n.s.*

p Yes. The participants were randomly assigned, and their experience was directed by the researcher. This is an experiment. It is likely that aerobic and circuit methods of training significantly reduce resting heart rate.

17 **a** H_0:$\mu_1 = \mu_2 = \mu_3$; H_1: at least two of the means are different.

West	Midwest	East	Summary values
$\Sigma X_1 = 18$	$\Sigma X_2 = 43$	$\Sigma X_3 = 19$	$\Sigma X = 80$
$n_1 = 5$	$n_2 = 5$	$n_3 = 5$	$\Sigma X^2 = 624$
$M_1 = 3.60$	$M_2 = 8.60$	$M_3 = 3.80$	$(\Sigma X)^2 = 6400$
			$N = 15,\ \ k = 3$

b $SS_{BG} = 506.80 - 426.67 = \mathbf{80.13}$
c $SS_W = 624 - 506.80 = \mathbf{117.20}$
d $df_{BG} = 3 - 1 = 2$
e $df_W = 15 - 3 = 12$
f $MS_{BG} = \dfrac{80.13}{2} = \mathbf{40.07}$
g $MS_W = \dfrac{117.20}{12} = \mathbf{9.77}$
h $SS_T = 624 - 426.67 = \mathbf{197.33}$
i $df_T = 15 - 1 = 2 + 12 = 14$
j $F = \dfrac{40.07}{9.77} = \mathbf{4.10}$
k $F_{crit}(2,12) = 3.88$ (for $\alpha = 0.05$)
l Reject H_0.

m

Source of variation	*SS*	*df*	*MS*	*F*	*p*
Between groups	80.13	2	40.07	4.10	<.05
Within groups(error)	117.20	12	9.77		
Total	197.33	14			

n $\eta^2 = \dfrac{80.13}{197.33} = 0.4061$ *or* 40.61%
o Tukey's HSD for all comparisons equals

$$HSD = q\sqrt{\frac{MS_W}{n}} = 3.77\sqrt{\frac{9.77}{5}} = 3.77(1.40) = 5.27$$

West vs. Midwest: 8.60 – 3.60 = 5 (*n.s.*)
East vs. Midwest: 8.60 – 3.80 = 4.8 (*n.s.*)
West vs. East: 3.60 – 3.80 = –0.2 (*n.s.*)

This is an unusual case where the one-way ANOVA registers significance, but the follow-up test does not find a difference between any pair of groups. Tukey's test is rather conservative.

If Fisher's *LSD* had run, we would have found evidence of two differences:

(t_{crit} for all *t*'s = 2.145 with $df = 14$, $\alpha = 0.05$)

$$t = \frac{8.60 - 3.60}{\sqrt{9.77\left(\frac{1}{5} = \frac{1}{5}\right)}} = 2.53 \ (p < 0.05)$$

East vs. Midwest:

$$t = \frac{8.60 - 3.80}{1.98} = 2.42 \ (p < 0.05)$$

West vs. East:

$$t = \frac{3.60 - 3.80}{1.98} = -0.10 \ (n.s.)$$

p No. This is a correlational design. The independent variable (geographical region) is not manipulated. Although participants are randomly *selected* from each region, participants are not randomly *assigned* to regions to determine the causal effect of regional residence. Therefore, the correct interpretation of these data is that there is an *association* between geographical residence and conservatism, not that one is *causing* the other.

18 **a** H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: at least two of the means are different.

Breathing	Medication	Control	Summary values
$\Sigma X_1 = 75$	$\Sigma X_2 = 70$	$\Sigma X_3 = 65$	$\Sigma X = 210$
$n_1 = 5$	$n_2 = 5$	$n_3 = 5$	$\Sigma X^2 = 3126$
$M_1 = 15$	$M_2 = 14$	$M_3 = 13$	$(\Sigma X)^2 = 44\,100$
			$N = 15$, $k = 3$

b $SS_{BG} = 2950 - 2940 = 10$
c $SS_W = 3126 - 2950 = 176$
d $df_{BG} = 3 - 1 = 2$
e $df_W = 15 - 3 = 12$

f $MS_{BG} = \dfrac{10}{2} = 5$

g $MS_W = \dfrac{176}{12} = 14.67$

h $SS_T = 3126 - 2940 = 186$

i $df_T = 15 - 1 = 2 + 12 = 14$

j $F = \dfrac{5}{14.67} = 0.34$

k $F_{crit}(2,12) = 3.88$ (for $\alpha = 0.05$)

l Fail to reject the H_0.

m

Source of variation	*SS*	*df*	*MS*	*F*	*p*
Between groups	10	2	5	0.34	*n.s.*
Within groups (error)	176	12	14.67		
Total	186	14			

n There is no statistical evidence of a difference among treatment conditions in the alleviation of panic attacks, $F(2,12) = 0.34$, *n.s.*

o Since the F ratio is nonsignificant, ω^2 is superfluous.

p Conducting post hoc comparisons is unwarranted since the F ratio is nonsignificant.

q It would have, if evidence of a difference had been found. The methodology is experimental. However, since no evidence of a difference was found, no causal claims can be made.

19 **a** H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$

Angry	Control	Summary values
$\Sigma X_1 = 54$	$\Sigma X_2 = 41$	$\Sigma X = 95$
$n_1 = 6$	$n_2 = 6$	$\Sigma X^2 = 823$
$M_1 = 9$	$M_2 = 6.83$	$(\Sigma X)^2 = 9025$
		$N = 12$, $k = 2$

b $SS_{BG} = 766.17 - 752.08 = 14.09$

c $SS_W = 823 - 766.17 = 56.83$

d $df_{BG} = 2 - 1 = 1$

e $df_W = 12 - 2 = 10$

f $MS_{BG} = \dfrac{14.09}{1} = 14.09$

g $MS_W = \dfrac{56.83}{10} = 5.68$

h $SS_T = 14.09 + 56.83 = 70.92$
i $df_T = 12 - 1 = 10 + 1 = 11$
j $F = \dfrac{14.09}{5.68} = 2.48$
k $F_{crit}(1,10) = 4.90$ (for $\alpha = 0.05$)
l Fail to reject the H0. (In comparison with Part 4, Problem 10, the conclusion is the same, fail to reject the H_0.)
m

Source of variation	*SS*	*df*	*MS*	*F*	*p*
Between groups	14.09	1	14.09	2.48	*n.s.*
Within groups (error)	56.83	10	5.68		
Total	70.92	11			

n There is no statistical evidence of a difference between conditions in the ability to control the car, $F(1,10) = 2.48$, *n.s.*
o Since the F ratio is nonsignificant, ω^2 is superfluous.
p It would have, if evidence of a difference had been found. The methodology is experimental. However, since no evidence of a difference was found, no causal claims can be made.

20 **a** H_0:$\mu_1 = \mu_2 = \mu_3$; H_1: at least two of the means are different.

New T-shirt	New Shoes	Control	Summary values
$\Sigma X_1 = 32$	$\Sigma X_2 = 41$	$\Sigma X_3 = 23$	$\Sigma X = 96$
$n_1 = 5$	$n_2 = 5$	$n_3 = 5$	$\Sigma X^2 = 664$
$M_1 = 6.4$	$M_2 = 8.2$	$M_3 = 4.6$	$(\Sigma X)^2 = 9216$
			$N = 15,\ \ k = 3$

b $SS_{BG} = 646.8 - 614.4 = \mathbf{32.4}$
c $SS_W = 664 - 646.8 = \mathbf{17.2}$
d $df_{BG} = 3 - 1 = 2$
e $df_W = 15 - 3 = 12$
f $MS_{BG} = \dfrac{32.4}{2} = 16.2$
g $MS_W = \dfrac{17.2}{12} = 1.43$
h $SS_T = 32.4 + 17.2 = 664 - 614.4 = \mathbf{49.6}$
i $df_T = 15 - 1 = 2 + 12 = 14$
j $F = \dfrac{16.2}{1.43} = 11.33$
k $F_{crit}(2,12) = 3.88$ (for $\alpha = 0.05$)

l Reject H_0.

m

Source of variation	*SS*	*df*	*MS*	*F*	*p*
Between groups	32.4	2	16.2	11.33	<.05
Within groups(error)	17.2	12	1.43		
Total	49.6	14			

n $\eta^2 = \frac{32.4}{49.6} = 0.6532$ *or* 65.32%

o Tukey's HSD for all comparisons equals

$$HSD = q\sqrt{\frac{MS_W}{n}} = 3.77\sqrt{\frac{1.43}{5}} = 3.77(0.53) = 2.02$$

Shirt vs. Shoes: 6.4 – 8.20 = −1.8 (*n.s.*)
Shirt vs. Control: 6.40 – 4.60 = 1.8 (*n.s.*)
Shoes vs. Control: 8.2 – 4.60 = 3.6 ($p < .05$)

p Statistical evidence suggests that type of clothing worn influences perceptions of happiness, $F(2,12) = 11.33$, $p < .05$. A post hoc Tukey test found statistical evidence suggesting that those wearing new shoes felt happier than those who were not wearing a new article of clothing at $p < .05$.

q Yes. This is an experimental design.

21 When H_0 is correct, the numerator of the *F* ratio is the result of only error variance (random factors). When H_0 is incorrect, the numerator includes error variance plus the effect due to treatment.

22 Since the *F* distribution is established with the assumption that H_0 is true, most *F* ratios cluster around 1, with the minimum value being 0, and all *F* values positive. Even with H_0 being true, sampling error may sometimes lead to large *F* values, resulting in the distribution being positively skewed.

23 False. Both will lead to the same conclusion about the null hypothesis; power is determined by other factors.

24 Between-group variation can be the result of treatment effect (primary variance), individual differences, and experimental error.

25 Individual differences and experimental error (combined they can be referred to as "secondary variance").

26 Only after the observed F directs the researcher to reject the null hypothesis and when there is an interest in making certain group comparisons.

27 Least significant difference.

28 Honestly significant difference.

29 Is

30

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	58.5	2	29.25	1.30	*n.s.*
Within groups	740.5	33	22.44		
Total	799.0	35			

There is no statistical evidence a difference between groups in systolic blood pressure, $F(2,33) = 1.30$, *n.s.*

31

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	51.91	2	25.96	3.94	< .05
Within groups	276.53	42	6.58		
Total	328.44	44			

There is statistical evidence of a reduction in the number of weekly headaches as a result of treatment, $F(2,42)$ 3.94, $p < .05$. (Our post hoc analyses will probably show the following results.) Both medication and biofeedback significantly reduce headaches in comparison with the control condition. There is no significant difference between the two forms of therapy, however.

32

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	143.22	3	47.74	11.19	< .05
Within groups	273.06	64	4.27		
Total	416.28	67			

There is statistical evidence suggesting that there is at least one difference between majors in terms of the number of shoes brought to university, $F(3,64) = 11.19$, $p < .05$. Tukey's *HSD* found evidence with $p < .05$ that science majors bring fewer shoes than history majors as well as theater majors and psychology majors bring fewer shoes than theater majors. If we chose a more permissive post hoc test, we may have found evidence that psychology majors bring fewer shoes than history majors.

Chapter 13

1 There must be two factors (or independent variables), each factor must have at least two conditions, participants must be only assigned to one combination of levels of the two factors (in other words, not repeatedly measured), and each of the possible combinations of conditions between the factors must have participants (no empty cells).

2 In the first stage of a two-way ANOVA, the total variance is partitioned into between-group and within-group variance. In the second stage, the between-group variance is partitioned into variance due to Factor *A*, variance due to Factor *B*, and variance due to the interaction.

3 A main effect occurs when there is an effect found among the conditions of one factor, independent of the influence of another factor. An interaction occurs when the effect of one factor is altered depending on the value of a second factor – this is new and unique variance not explained by main effects.

4 The following are examples. The labels selected by the student will vary, but the grid design will not.

a 2×2

	In a group	Not in a group
Peg-word System		
Acronyms		

The factor described in rows (mnemonic technique) can be an independent variable.

The factor described in columns (group type) can be an independent variable.

b 3×2

	White Collar	Blue Collar
Group		
Behavioral		
Person Centered		

The factor described in rows (therapy type) can be an independent variable.

c 4×4

	Arts	Humanities	Sciences	Preprofessional
Dorm room				
Library				
Food Commons				
Park Bench				

The factor described in rows (study location) can be an independent variable.

d 3×7

	Sun.	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.
Eating							
Mood							
Dissociative							

5 Only by placing them in the same design will the researcher be able to look for interactions.

6 Two different types (main effects and interaction). Two main effects and one interaction.

7 **a** H_0: $\mu_{A_1} = \mu_{A_2} = \mu_{A_k}$; there is no difference in population means of the levels of Factor A.

H_1: H_0 is false, or at least one of the means of a level of Factor *A* is different from at least one other level.

b H_0: $\mu_{B_1} = \mu_{B_2} = \mu_{B_k}$; there is no difference in population means of the levels of Factor *B*.

H_1: H_0 is false, or at least one of the means of a level of Factor *B* is different from at least one other level.

c H_0: There is no interaction.
H_1: There is an interaction.

8

a $F_A = \dfrac{73.50}{2.12} = 34.67,\ p < .05;\ F_{crit}(2, 45) = 3.23$

$F_B = \dfrac{13.72}{2.12} = 6.47,\ p < 0.05;\ F_{crit}(2, 45) = 3.23$

$F_{A \times B} = \dfrac{3.05}{2.12} = 1.44,\ n.s.;\ F_{crit}(4, 45) = 2.61$

b Critical values: $F_A(1, 30) = 4.17$; $F_B(2, 30) = 3.32$; $F_{A \times B}(2, 30) = 3.32$

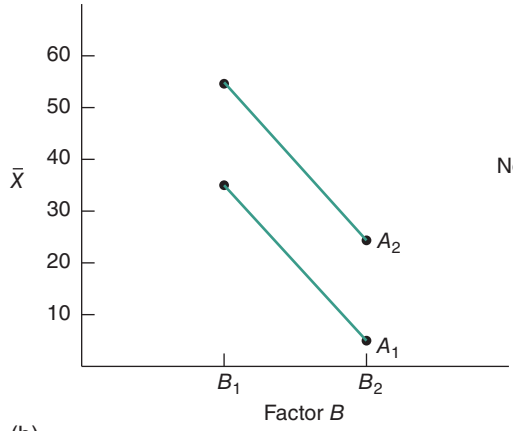
Source	*SS*	*df*	*MS*	*F*	*p*
Factor *A*	3.36	1	3.36	0.90	*n.s.*
Factor *B*	66.67	2	33.34	8.94	<.05
A × *B*	56.89	2	28.45	7.63	<.05
Within groups	111.83	30	3.73		
Total	238.75	35			

c All critical values are based on *df* = 1,16 and equal 4.49.

Source	*SS*	*df*	*MS*	*F*	*p*
Factor *A*	0.45	1	0.45	0.35	*n.s.*
Factor *B*	6.05	1	6.05	4.73	<.05
A × *B*	84.05	1	84.05	65.66	<.05
Within groups	20.48	16	1.28		
Total	111.03	19			

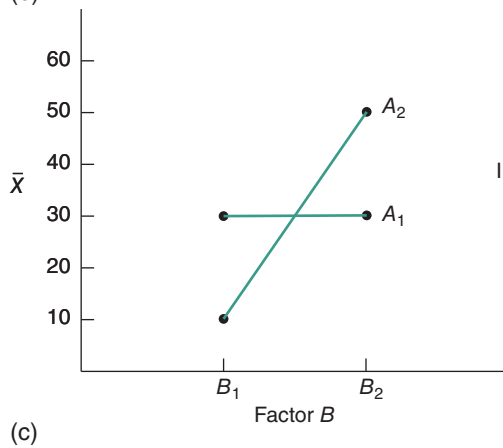
9 Refer to graphs.

(a)



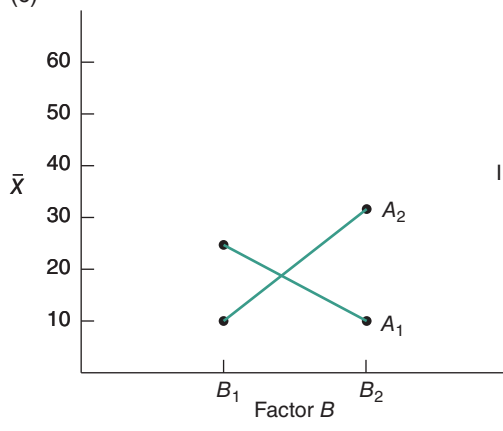
No interaction

(b)



Interaction

(c)



Interaction

10 **a** *df* for F_A are 2, 52.
b *df* for F_B are 2, 52.
c *df* for $F_{A \times B}$ are 4, 52.

11 The size of an *F* reflects the level of certainty that the null of no difference between population means can be rejected. A large treatment effect does lead to large *F* values, but so does a high degree of statistical power. Another statistic is needed to measure effect size alone.

12 These values reflect different ways to represent the ratio of treatment variance associated with the effect in question to the total amount of variance in the study.

13 In a 2 × 2 design, there are only two conditions for each factor. A rejected null for a main effect would have to mean that there is statistical evidence of a difference between the only two-cell means in the study.

14 Fisher's *LSD*. Tukey's *HSD* requires each cell in the design to have the same number of participants.

15 Tukey's *HSD*.

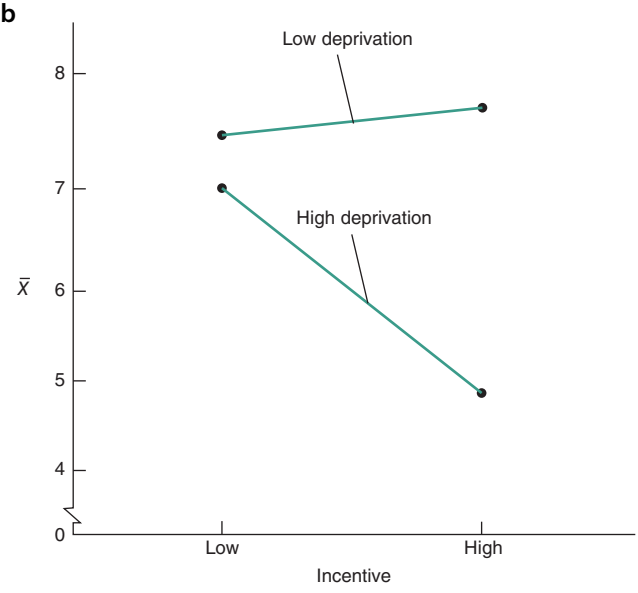
16 **a**

		Factor *B*: Deprivation		
		High Deprivation	*Low Deprivation*	
Factor *A*: Incentive	*Low Incentive*	$M_{A_1B_1} = 7.0$ $\Sigma X_1 = 35$ $\Sigma X_1^2 = 247$ $n_1 = 5$	$M_{A_1B_2} = 7.4$ $\Sigma X_2 = 37$ $\Sigma X_2^2 = 285$ $n_2 = 5$	$M_{A_1} = 7.2$
	High Incentive	$M_{A_2B_1} = 4.8$ $\Sigma X_3 = 24$ $\Sigma X_3^2 = 118$ $n_3 = 5$	$M_{A_2B_2} = 7.6$ $\Sigma X_4 = 38$ $\Sigma X_4^2 = 296$ $n_4 = 5$	$M_{A_2} = 6.2$
		$M_{A_2} = 5.9$	$M_{B_2} = 7.5$	

$M_G = 6.7$; $\Sigma X = 134$; $\Sigma X^2 = 946$; $N = 20$; critical values for all *F*'s (1,16) = 4.49.

Source	SS	df	MS	F	p
Factor *A*	5.0	1	5.0	3.45	*n.s.*
Factor *B*	12.80	1	12.80	8.83	< .05
A × *B*	7.20	1	7.20	4.97	< .05
Within groups	23.20	16	1.45		
Total	48.20	19			

b



c $\varpi_B^2 = \dfrac{11.35}{49.65} = 0.23$ or 23%

$\varpi_{A \times B}^2 = \dfrac{5.75}{49.65} = 0.12$ or 12%

d Main ϖeffect for Factor *B* does not require further analysis – only two conditions.

The significant interaction for a 2 × 2 design could be further analyzed by running Fisher's *LSD* for pairs of cells. For instance, the *t* for the two deprivation conditions for low-incentive registers as nonsignificant, [$t(16) = 0.23$, *n.s.*] but the *t* for the two deprivation conditions for high-incentive registers as significant [$t(16) = 2.59$, $p < .05$].

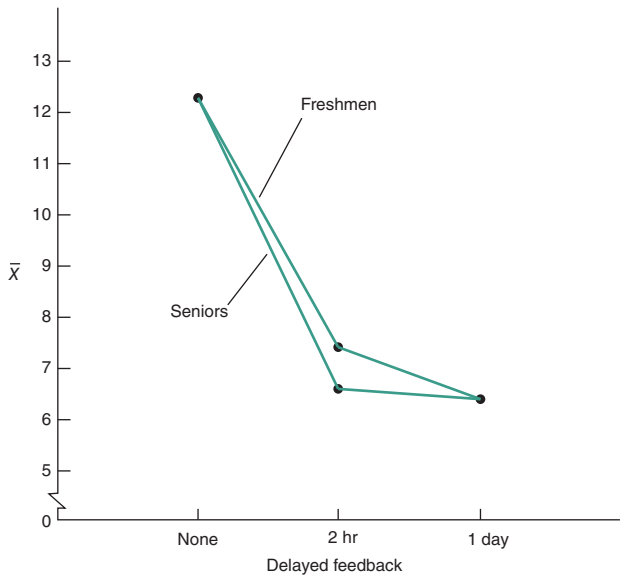
e The interaction needs to be interpreted first. Based on the follow-up *LSD* tests, statistical evidence has been found suggesting that rats with high deprivation and high incentive take less time to traverse a maze than rats in the other conditions. Although Factor *B* (deprivation) is registered as significant, there does not appear to be a main effect for this variable.

Three of the four means are relatively close; only the high-deprivation/high-incentive condition showed change in the observed data.

17 **a** Factor *A* level 1 is Freshmen; level 2 is Seniors; Factor *B* level 1 is No Delay; level 2 is Two-Hour Delay; level 3 is One-Day Delay.
Freshmen/No Delay: $M = 12.20$; $\Sigma X = 61$; $\Sigma X^2 = 759$; $n = 5$
Freshmen/Two-Hour Delay: $M = 7.40$; $\Sigma X = 37$; $\Sigma X^2 = 283$; $n = 5$
Freshmen/One-Day Delay: $M = 6.20$; $\Sigma X = 31$; $\Sigma X^2 = 199$; $n = 5$
Seniors/No Delay: $M = 12.20$; $\Sigma X = 61$; $\Sigma X^2 = 763$; $n = 5$
Seniors/Two-Hour Delay: $M = 6.40$; $\Sigma X = 32$; $\Sigma X^2 = 214$; $n = 5$
Seniors/One-Day Delay: $M = 6.20$; $\Sigma X = 31$; $\Sigma X^2 = 199$; $n = 5$
Freshmen: $M = 8.60$; Seniors: $M = 8.27$
No Delay: $M = 12.20$; Two-Hour Delay: $M = 6.90$; One-Day Delay: $M = 6.20$
$M_G = 8.43$; $\Sigma X_G = 253$; $\Sigma X_G^2 = 2417$; $N = 30$

Source	*SS*	*df*	*MS*	*F*	*p*
Factor *A*	0.84	1	0.84	0.31	*n.s.*
Factor *B*	215.27	2	107.64	39.43	<.05
A × *B*	1.66	2	0.83	0.30	*n.s.*
Within groups	65.60	24	2.73		
Total	283.37	29			

b



c $\eta_B^2 = \frac{SS_B}{SS_B + SS_W} = \frac{215.27}{215.27 + 65.6} = \frac{215.27}{280.87} = 0.77$ or 77%

d No Delay vs. One-Day Delay: $t = \frac{6}{0.74} = 8.11$, $p < .05$

No Delay vs. Two-Hour Delay: $t = \frac{5.3}{0.74} = 7.16$, $p < .05$

Two-Hour Delay vs. One-Day Delay: $t = \frac{0.7}{0.74} = 0.95$, *n.s.*

$t_{crit}\ (24) = 2.064$

Tukey's HSD is $HSD_B = q_B\sqrt{\frac{MS_W}{n_B}} = 3.88\sqrt{\frac{2.73}{10}} = 2.03$. When applied to the three sets of means, the same findings emerge – fail to reject the null between Two-Hour Delay vs. One-Day Delay, and reject the other two.

e Statistical evidence suggests that delayed feedback has an effect on learning [$F(2,24) = 39.43$, $p < .05$], with Two-Hour Delays [$t(24) = 7.16$, $p < .05$] and One-Day Delays [$t(24) = 8.11$, $p < .05$] both showing lower retention than No Delay. No statistical evidence of a difference between Two-Hour and One-Day Delays, $t(24) = 0.95$, *n.s.* There was also no evidence to suggest a difference for educational level, $F(1,24) = 0.31$, *n.s.*, and no evidence of an interaction, $F(2,24) = 0.30$, *n.s.*

18 **a**

Source	*SS*	*df*	*MS*	*F*	*p*
Factor *A*	76.05	1	76.05	89.47	< .05
Factor *B*	31.25	1	31.25	36.76	< .05
A × *B*	4.05	1	4.05	4.76	< .05
Within groups	13.60	16	0.85		
Total	124.95	19			

b $\eta_A^2 = \frac{SS_A}{SS_A + SS_W} = \frac{76.05}{76.05 + 13.60} = 0.85$ *or* 85%

$\eta_B^2 = \frac{SS_B}{SS_B + SS_W} = \frac{31.25}{31.25 + 13.60} = 0.70$ *or* 70%

$\eta_{A\times B}^2 = \frac{SS_{A\times B}}{SS_{A\times B} + SS_W} = \frac{4.05}{4.05 + 13.60} = 0.23$ *or* 23%

c Because it is a 2 × 2 design, no follow-up tests are needed to explore the main effects. However, we could run either follow-up test to help explore the interaction. Let us run Tukey's *HSD*:

$$HSD_{A\times B} = q_{A\times B}\sqrt{\frac{MS_W}{n_{A\times B}}} = 3.65\sqrt{\frac{0.85}{5}} = 1.50$$

Using this value to compare sets of means, we find evidence of the following: medication improves mood and even more so when one is engaged in cognitive therapy. Cognitive therapy improves mood whether one is taking medication or not. And medication improves mood whether or not one is engaged in cognitive therapy.

If using Fisher's *LSD* test:

Since all cells have the same sample size, the numerator for each test will be $\sqrt{0.85(0.4)} = 0.58$, and the t_{crit} will always be ±2.12. The *t*'s from the perspective of cognitive therapy are [(8.6 – 5.2)/0.58 = 5.86] for medication and [(3.8 – 2.2)/0.58 = 2.75] for no medication. The *t*'s from the perspective of medication are [(8.6 – 3.8)/0.58 = 8.28] for those engaged in cognitive therapy and [(5.2 – 2.2)/0.58 = 5.17] for those who are not. All are suggestive of population differences.

By visually looking at a graph of the means, it appears that the interaction is being driven by the cell containing scores from those experiencing both cognitive therapy and medication. This combination of variables seems to produce a new and unique effect, even greater than the main effects found with each variable separately.

19 a

Source	*SS*	*df*	*MS*	*F*	*p*
Factor *A*	0.05	**1**	0.05	**0.03**	***n.s.***
Factor *B*	**18.05**	**1**	**18.05**	**9.86**	**< .05**
A* × *B	**2.45**	**1**	**2.45**	**1.34**	***n.s.***
Within groups	**29.20**	**16**	**1.83**		
Total	**49.75**	**19**			

b $\varpi_B^2 = \dfrac{SS_B - df_b(MS_W)}{SS_T + MS_W} = \dfrac{18.05 - 1(1.83)}{49.75 + 1.83} = \dfrac{16.22}{51.58} = 0.31$ or 31%.

c No statistical evidence found regarding either Factor *A* (biological sex of student) or the interaction. There was statistical evidence found suggesting that Factor *B* (the attractiveness of the professor) influenced student's perceptions of learning, $F(1,16) = 9.86$, $p < .05$, with students believing they can learn more from a physically attractive professor.

d The researchers are assuming that the Likert scale used to measure student's beliefs is measuring this variable on an interval scale. This assumption may be unfounded.

20 **a**

Source	SS	df	MS	F	p
Factor *A*	24.08	1	0.05	9.63	< .05
Factor *B*	6.75	1	18.05	2.70	*n.s.*
A × *B*	2.08	1	2.45	0.83	*n.s.*
Within groups	20.00	8	1.83		
Total	52.91	11			

b $\eta_A^2 = \dfrac{SS_A}{SS_A + SS_W} = \dfrac{24.08}{24.08 + 20.00} = 0.55$ or 55%.

c $M_{A_1} = 7.50, M_{A_2} = 4.67$. Since Factor *A* (Cologne) has only two conditions, there is no need for follow-up tests. Statistical evidence suggests that the wearing of cologne influences students' tendencies to engage that person in their conversation, $F(1,8) = 9.63$, $p < .05$. Students are more likely to engage in conversation longer with a person deemed very attractive ($M = 7.50$) than with a person deemed to be average looking ($M = 4.67$).

21 **a**

Source	SS	df	MS	F	p
Factor *A*	51.04	1	51.04	.94	*n.s.*
Factor *B*	260.04	1	260.04	4.81	<.05
A × *B*	3.38	1	3.38	.06	*n.s.*
Within groups	1082.17	20	54.11		
Total	1396.63	23			

b $M_{B_1} = 28.42$, $M_{B_2} = 21.84$. Since the *F* for Factor *B* allows us to reject the null hypothesis, we can conclude that statistical evidence suggests that the race of the defendant has an effect on sentencing, with Black defendants drawing stiffer sentences.

22 **a**

Source	SS	df	MS	F	p
Factor *A*	220.03	1	220.03	9.02	<.05
Factor *B*	288.17	2	144.09	5.91	<.05
A × *B*	42.72	2	21.36	0.88	*n.s.*
Within groups	731.83	30	24.39		
Total	1282.75	35			

b $M_{A_1} = 14.78$; $M_{A_2} = 19.12$; $M_{B_1} = 21.17$; $M_{B_2} = 16.0$; $M_{B_3} = 14.59$

All critical values are $t_{crit}(30) = \pm 2.042$. Since Factor A is significant and there are only two levels of Factor A, it is not necessary to perform a t test between the two levels.

$$\text{Type A vs. Type B: } t = \frac{5.17}{1.98} = 2.61,\ p < 0.05$$

$$\text{Type A vs. Type X: } t = \frac{6.58}{1.98} = 3.32,\ p < 0.05$$

$$\text{Type B vs. Type X: } t = \frac{1.41}{1.98} = 0.71,\ n.s.$$

c Since the Factor A F leads us to reject the null hypothesis [$F(1,30) = 9.02$, $p < .05$], we can conclude that statistical evidence suggests that incentive affects sales production, with commission ($M = 19.12$) producing more sales than salary ($M = 14.78$). Since the Factor B F also leads us to reject the null hypothesis [$F(1,30) = 5.91$, $p < .05$], we can conclude that statistical evidence suggests that personality type affects sales production. Follow-up analyses suggest that Type A's produce more than Type B's [$t(30) = 2.61$, $p < .05$] or Type X's [$t(30) = 3.32$, $p < .05$], with no evidence of a difference between Type B's and Type X's [$t(30) = 0.71$, *n.s.*]. Unfortunately, for the researcher's hypothesis, there is no statistical evidence suggesting an interaction between incentive and type of personality on sales production.

23 In an independent- or dependent-samples t test, we are testing for a main effect since there is only one independent variable. However, protected t tests can be used in a two-way ANOVA to elucidate the nature of an interaction effect.

24 A one-way ANOVA tests for a main effect since there is only one factor, although that factor can have more than two levels.

25 The presence of a significant interaction qualifies a straightforward interpretation of a main effect. In other words, the interaction may be making it look as if there is a main effect, when in reality the variance is due to the interaction. Higher-order effects sometimes produce illusory lower-order effects.

26 There would be no effect on the main effects, interaction, or MS_W. Adding a constant does not affect the variance; therefore, it has no effect on an ANOVA.

27

Source	SS	df	MS	F	p
Factor *A*	34 003.34	1	34 003.34	100.24	<.05
Factor *B*	31.86	2	15.93	.05	*n.s.*
A × *B*	23.03	2	11.52	.03	*n.s.*
Within groups	38 671	114	339.22		
Total	72 729.20	119			

There is statistical evidence suggesting that the biological sex of a child relates to their ability to delay gratification [$F(1,114) = 100.24$, $p < .05$], with biological females being able to delay longer ($M = 63.03$) than males ($M = 29.37$). There is no statistical evidence of differences between cognitive strategies [$F(2,114) = 0.05$, *n.s.*] and no statistical evidence of an interaction between biological sex and cognitive strategy [$F(2,114) = 0.03$, *n.s.*].

The effect size for Factor *A*, according to η^2, is [34 003.34/(34 003.34 + 38 671)] 0.47 or 47%.

28

Source	SS	df	MS	F	p
Factor *A*	187.26	1	187.26	6.35	<.05
Factor *B*	224.26	1	224.26	7.61	<.05
A × *B*	693.60	1	693.60	23.53	<.05
Within groups	1650.80	56	29.48		
Total	2755.93	59			

Marginal means: $M_{highdrive} = 21.73$; $M_{lowdrive} = 18.2$; $M_{easy} = 18.03$; $M_{difficult} = 21.90$

Cell means: $M_{high/easy} = 16.4$; $M_{high/diff} = 27.07$; $M_{low/easy} = 19.67$; $M_{low/diff} = 16.73$

Follow-up tests: Let us use Tukey's *HSD* to get a value to compare group means from the same factor:

$$HSD_{A\times B} = q_{A\times B}\sqrt{\frac{MS_W}{n_{A\times B}}} = 3.41\sqrt{\frac{29.48}{15}} = 4.78$$

Using this value and then looking at our cell means, we see that one cell is driving all three effects, the "high-drive state/difficult task" condition. There is no evidence for other differences. This simplifies our interpretation. We have statistical evidence to suggest that drive state and task

difficulty combine to create greater arithmetic errors on the given task, $F(1,56) = 23.53$, $p < .05$. Further analysis shows that participants in the high-drive condition make more errors if the task is difficult ($M = 27.07$) compared with those solving easy problems ($M = 16.4$) or those in the low-drive state, regardless of task difficulty ($M_{easy} = 19.67$; $M_{diff} = 16.73$), Tukey's $HSD = 5.26$.

29

Source	SS	df	MS	F	p
Factor *A*	19.01	1	19.01	8.85	< .05
Factor *B*	1.19	2	0.58	0.28	*n.s.*
A × *B*	5.03	2	2.51	1.17	*n.s.*
Within groups	141.75	66	2.15		
Total	166.98	71			

Since the only null rejected is the Factor *A* null, and it only has two conditions, no follow-up tests are needed. The analysis is straightforward; statistical evidence suggests that students who use Facebook while studying ($M = 3.72$) perform more poorly than students who do not ($M = 4.75$), $F(1,66) = 8.85$, $p < .05$. There was no evidence to suggest an effect related to the type of material studied [$F(2,66) = 0.28$, *n.s.*] or an interaction between Facebook use and the type of material studied [$F(2,66) = 1.17$, *n.s*].

We should be cautious about using Likert scale data. Means only make sense when the scale being used has conserved the quantitative space between each integer.

30

Source	SS	df	MS	F	p
Factor *A*	0.56	1	0.56	0.01	*n.s.*
Factor *B*	1105.56	1	1105.56	12.82	< .05
A × *B*	3.06	1	3.06	0.04	*n.s.*
Within groups	1034.75	12	86.23		
Total	2143.93	15			

Since the only rejected null is the Factor *B* null, and it only has two conditions, no follow-up tests are needed. The analysis is straightforward; statistical evidence suggests that students who study with music ($M = 23.25$) performed more poorly on the test than those who did not ($M = 39.88$), $F(1,12) = 12.82$, $p < .05$. There was no evidence to suggest an effect related to the type of material studied [$F(1,12) = 0.01$, *n.s.*] or an interaction between music use and the type of material studied [$F(1,12) = 0.04$, *n.s.*].

Chapter 14

1 Research designs involving the repeated measuring of more than two conditions.

2 In a repeated-measures design, each participant is exposed to each treatment condition. In a between-groups design, each participant receives only one treatment.

3 "Order effects" is the name given to the effect when a difference between treatment conditions occurs due to the order of presentation in a repeated-measures design. When the order of presentation is not being investigated, order effects introduce confounding variance.

4 "Counterbalancing" is a strategy used with repeated-measures designs in which participants differ by the order in which experimental conditions are presented. The purpose of counterbalancing is to prevent confounding of the independent variable with order effects by distributing the carryover effects that come with repeated measuring across all experimental conditions.

5 Repeated-measures designs cannot be used when the variable is not subject to experimental manipulation (e.g. participant variables). It should not be run when carryover effects cannot be controlled. For example, some forms of therapy result in a relatively permanent change in the participant.

6 In a repeated-measures design, between-group variation is made up of treatment effect and experimental error (there is no variation due to individual differences); within-group variation consists of individual differences (between-participant variability) and experimental error.

7 The effect due to individual differences is removed from the error term of a within-groups design; it is mathematically partitioned out.

8 MS_{BG} divided by MS_{error}

9 By using the same participants in every treatment condition, it is impossible for one treatment condition to have more or less of a participant variable than another treatment condition. The variance due to individual differences, quantified due to repeated measuring of individuals, can then be partitioned out, providing a more powerful test.

10 **c** Repeated-measures designs can achieve much more power with few participants, and this is amplified if the variance removed due to individual differences is large.

11 For a one-way ANOVA, $df_{BG} = k - 1 = 2$; $df_W = N - k = 12$. For a repeated-measures ANOVA, $df_{BG} = k - 1 = 2$; $df_{error} = (N - k) - (n - 1) = 28$. F_{crit} (one way) = 3.88; F_{crit} (repeated measures) = 3.34.

12 df_T

13 $df_{BG} = k - 1 = 4$; df_{error} $(N - k) - (n - 1) = 76$. The answers are 4 and 76.

14 9

15

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	52.94	2	26.47	11.51	<.05
Within groups	20	12			
Between participants	1.60	4			
Error	18.40	8	2.30		
Total	72.94	14			

16 **a**

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	80	2	40	16.81	<.05
Within groups	30	12			
Between participants	10.96	4			
Error	19.04	8	2.38		
Total	110	14			

b

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	117.55	4	29.39	18.29	<.05
Within groups	66.82	40			
Between participants	15.4	8			
Error	51.42	32	1.61		
Total	184.37	44			

17 a

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	32.53	2	16.27	14.79	<.05
Within groups	13.20	12			
Between participants	4.40	4			
Error	8.80	8	1.10		
Total	45.73	14			

b $\varpi^2 = \dfrac{32.53 - 2(1.10)}{45.73 + 1.10} = \dfrac{30.33}{46.83} = 0.65$ *or* 65%

$\eta^2 = \dfrac{32.53}{45.73} = 0.71$ *or* 71%

c $HSD = 4.04\sqrt{\dfrac{1.10}{5}} = 1.89$

Mean differences for Incentive *A* vs. *B* = 2, p <.05
Mean differences for Incentive *A* vs. *C* = 3.6, p <.05
Mean differences for Incentive *B* vs. *C* = 1.6, *n.s.*
If we would have run Fisher's *LSD* instead,
t_{crit} (8) = ±2.306

Incentive *A* vs. *B*: $t = \dfrac{7.40 - 5.40}{0.66} = 3.03,\ p < 0.05$

Incentive *A* vs. *C*: $t = \dfrac{7.40 - 3.80}{0.66} = 5.45,\ p < 0.05$

Incentive *B* vs. *C*: $t = \dfrac{5.40 - 3.80}{0.66} = 2.42,\ p < 0.05$

As we can see, there is perfect agreement.

d Statistical evidence suggests that the type of incentive influenced sales, $F(2,8) = 14.79$, $p < .05$. Further analyses found statistical evidence suggesting that Incentive *A* worked better than Incentive *B*, *HSD* = 2, $p < .05$, and Incentive *C*, *HSD* = −3.6, $p < .05$. No evidence of a difference was found between Incentives *B* and *C*, *HSD* = 1.6, *n.s.*

e The problem description does not specify, but it is reasonable to presume that counterbalancing of incentive programs was not used. As a result, carryover effects might be introducing confounding variance.

18 a

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	2754. 17	2	1377.09	47.54	<.05
Within groups	212.75	9			

(*Continued*)

(Continued)

Source	SS	df	MS	F	p
Between participants	38.92	3			
Error	173.83	6	28.97		
Total	2966.92	11			

b $\varpi^2 = \dfrac{2754.17 - (2)28.97}{2966.92 + 28.97} = \dfrac{2696.23}{2995.89} = 0.90$ *or* 90%

$\eta^2 = \dfrac{2754.17}{2966.92} = 0.93$ *or* 93%

$t_{crit}(6) = \pm 2.447$

Positive vs. Negative: $t = \dfrac{52 - 15.75}{3.81} = 9.51,\ p < 0.05$

Positive vs. Control: $t = \dfrac{52 - 27}{3.81} = 6.56,\ p < 0.05$

Negative vs. Control: $t = \dfrac{15.75 - 27}{3.81} = -2.95,\ p < 0.05$

c Statistical evidence suggests that the type of subliminal message influenced problem solving, $F(2,6) = 47.54$, $p < .05$. Further analyses found statistical evidence suggesting that positive messages worked better than negative messages, $t(6) = 9.51$, $p < .05$, and the control group, $t(6) = 6.56$, $p < .05$, and the control group outperformed the negative feedback group, $t(6) = -2.95$, $p < .05$.

d Since participants are repeatedly measured, it is hoped that each set of math problems were deemed equally difficult (and not merely the same list of problems). Furthermore, no statement was made about counterbalancing the order of presentation. If this was not performed, carryover effects may be introducing confounding variance.

19 **a**

Source	SS	df	MS	F	p
Between groups	31.60	2	15.80	33.62	<.05
Within groups	10.80	12			
Between participants	7.06	4			
Error	3.74	8	0.47		
Total	42.40	14			

b $\varpi^2 = \dfrac{31.60 - (2)0.47}{42.40 + 0.47} = \dfrac{30.66}{42.87} = 0.72$ *or* 72%

$\eta^2 = \dfrac{31.60}{42.40} = 0.75$ *or* 75%

c $HSD = 4.04\sqrt{\dfrac{0.47}{5}} = 1.24$

Mean differences for Feta Cheese vs. Caviar = 0.8, *n.s.*
Mean differences for Feta Cheese vs. Popcorn = 3.4, $p < .05$
Mean differences for Caviar vs. Popcorn = 2.6, $p < .05$
If we would have run Fisher's *LSD* instead,
$t_{crit}\ (8) = \pm 2.306$
Feta Cheese vs. Caviar: $t = \frac{1.40 - 2.20}{0.43} = -1.86,\ n.s.$
Feta Cheese vs. Popcorn: $t = \frac{1.40 - 4.80}{0.43} = -7.91,\ p < 0.05$
Caviar vs. Popcorn: $t = \frac{2.20 - 4.80}{0.43} = -6.05,\ p < .05$

There is basic agreement between these two follow-up measures for this exercise.

d Statistical evidence suggests that some hors d'oeuvres go better with wine than others, $F(2,8) = 33.62$, $p < .05$. Further analyses found statistical evidence suggesting that Popcorn went better than Feta Cheese, $HSD = 3.4$, $p < .05$, and caviar, $HSD = -2.6$, $p < .05$. No evidence of a difference was found between Feta Cheese and Caviar, $HSD = 0.8$, *n.s.*

e There may be an issue with using a Likert scale to measure a taste rating, and there was no mention in the problem description that counterbalancing was used. If the order was preserved throughout the study, carryover effects might be introducing confounding variance.

20 **a**

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	32.40	2	16.20	26.27	< .05
Within groups	17.2	12			
Between participants	12.27	4			
Error	4.93	8	0.62		
Total	49.6	14			

b $\varpi^2 = \dfrac{32.40 - (2)0.62}{49.60 + 0.62} = \dfrac{31.16}{50.22} = 0.62$ *or* 62%

$\eta^2 = \dfrac{32.40}{49.60} = 0.65$ *or* 65%

c t_{crit} (8) = ±2.306

Shirt vs. Shoes: $t = \dfrac{6.40 - 8.20}{0.50} = -3.60,\ p < 0.05$

Shirt vs. Control: $t = \dfrac{6.40 - 4.60}{0.50} = 3.60,\ p < 0.05$

Shoes vs. Control: $t = \dfrac{8.20 - 4.60}{0.50} = 7.20,\ p < 0.05$

d Statistical evidence suggests that the type of new clothing influenced perceptions of happiness, $F(2,8) = 16.20$, $p < .05$. Further analyses found statistical evidence suggesting that wearing new shoes generated more happiness than wearing a new shirt, $t(8) = 3.60$, $p < .05$, and the control group, $t(8) = 7.20$, $p < .05$, and the new shirt group seemed happier than the control group, $t(8) = 3.60$, $p < .05$. Notice the difference findings compared with Problem 20 in Chapter 12 – which used the same data but with an independent-groups design. The increased power of a repeated-measures ANOVA found evidence of more differences.

e It appears that the researchers dealt with carryover effect issues through counterbalancing, though having the control condition always go last (as is implied) might be a problem. Furthermore, there is the issue of measuring happiness using a 10-point scale: is this interval data or higher?

21 **a**

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	5114.80	2	2557.40	9.18	< .05
Within groups	2893.60	12			
Between participants	663.73	4			
Error	2229.87	8	278.73		
Total	8008.40	14			

b $\varpi^2 = \dfrac{5114.80 - (2)278.73}{8008.40 + 278.73} = \dfrac{4557.34}{8287.13} = 0.55$ *or* 55%

$\eta^2 = \dfrac{5114.80}{8008.40} = 0.64$ *or* 64%

c t_{crit} (8) = ±2.306

Low vs. Medium: $t = \dfrac{62.80 - 55.80}{10.56} = 0.66,\ n.s.$

Low vs. High: $t = \dfrac{62.80 - 98}{10.56} = -3.33,\ p < 0.05$

Medium vs. High: $t = \dfrac{55.80 - 98}{10.56} = -4.00,\ p < 0.05$

d Statistical evidence suggests that the degree of distraction influenced pain tolerance, $F(2,8) = 9.18, p < .05$. Further analyses found statistical evidence suggesting that high degrees of distraction worked better than both low degrees of distraction, $t(8) = -3.33, p < .05$, and medium degrees of distraction, $t(8) = -4.00, p < .05$. No evidence of a difference was found between low and medium degrees of distraction, $t(8) = 0.66$, *n.s.*

e It is assumed that the 20-minute interval in between the presentation of different conditions eliminated carryover effects. No statement was made in the description about counterbalancing the conditions. Carryover effects due to extreme nature of the dependent variable and/or the order of presentation may be introducing confounding variance.

22 **a**

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	89 334.77	2	44 667.39	16.24	<.05
Within groups	62 742.83	15			
Between participants	35 236.93	5			
Error	27 505.90	10	2 750.59		
Total	152 077.61	17			

b $\varpi^2 = \dfrac{89\ 334.77 - 2(2\ 750.59)}{152\ 077.61 + 2\ 750.59} = \dfrac{83\ 833.59}{154\ 828.20} = 0.54\ or\ 54\%$

$\eta^2 = \dfrac{89\ 334.77}{152\ 077.61} = 0.59\ or\ 59\%$

c $t_{crit}\ (10) = \pm 2.228$

Technique *A* vs. *B*: $t = \dfrac{407.50 - 240.22}{30.28} = 5.52,\ p < 0.05$

Technique *A* vs. *C*: $t = \dfrac{407.50 - 361}{30.28} = 1.54,\ n.s.$

Technique *B* vs. *C*: $t = \dfrac{240.33 - 361}{30.28} = -3.99,\ p < 0.05$

d Statistical evidence suggests that the type of technique influenced reading speed, $F(2,10) = 16.24, p < .05$. Further analyses found statistical evidence suggesting that Technique *A* outpaced Technique *B*, $t(10) = 5.52, p < .05$, and Technique *C* also outpaced Technique *B*, $t(10) = -3.99, p < .05$. No evidence of a difference was found between Techniques *A* and *C*, $t(10) = 1.54$, *n.s.*

e The order of presentation is not addressed by the problem. This could be an issue if carryover effects are experienced.

23 **a**

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	37.00	2	18.50	9.10	< .05
Within groups	299.50	15			
Between participants	279.17	5			
Error	20.33	10	2.03		
Total	336.50	17			

b $\varpi^2 = \dfrac{37.00 - (2)2.03}{336.50 + 2.03} = \dfrac{32.94}{338.53} = 0.10 \; or \; 10\%$

$\eta^2 = \dfrac{37.00}{336.50} = 0.11 \; or \; 11\%$

c $HSD = 3.88\sqrt{\dfrac{2.03}{6}} = 2.26$

Mean differences for Quiet Room vs. Dining Hall = 3.5, p <.05
Mean differences for Quiet Room vs. Various = 2.0, *n.s.*
Mean differences for Dining Hall vs. Various = 1.5, *n.s.*

d Statistical evidence suggests that study location influences student performance, $F(2,10) = 9.10$, $p < .05$. Further analyses showed evidence suggesting that students perform better having studied in a quiet room compared with the dining hall, $HSD = 3.5$, $p < .05$. No evidence of differences between a quiet room and a variety of locations was not found, $HSD = 2.0$, *n.s.*, and no evidence of a difference was found between studying in the dining hall or a variety of locations, $HSD = 1.5$, *n.s.*

24

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	2574.55	2	1287.28	15.60	< .05
Within groups	2933.38	45			
Between participants	457.25	15			
Error	2476.13	30	82.54		
Total	5507.92	47			

Statistical evidence suggests that social desirability is affected by attractiveness, $F(2,30) = 15.60$, p <.05. Follow-up tests provide evidence suggesting that attractive people are considered more socially desirable than average-looking people, $t(30) = 2.73$, p <.05, and unattractive people, $t(30) = 5.59$, p <.05, while average-looking people are considered more socially desirable to unattractive people, $t(30) = 2.86$, p <.05.

25

Source	SS	df	MS	F	p
Between groups	45.64	2	22.82	4.35	<.05
Within groups	228.00	42			
Between participants	80.97	14			
Error	147.03	28	5.25		
Total	273.64	44			

Statistical evidence suggests that success–achievement ratings by biological females of biological males is influenced by visual cues related to concepts of masculinity, $F(2,28) = 15.60$, $p < .05$. Follow-up tests provide evidence suggesting that biological females view as more likely to achieve success and biological males who possess physical characteristics associated with a traditional understanding of masculinity compared with those who possess nontraditional physical characteristics, $t(28) = 2.93$, $p < .05$. Statistical evidence for other effects was not found.

26

Source	SS	df	MS	F	p
Between groups	57.27	2	28.63	10.87	<.05
Within groups	91.40	27			
Between participants	44.00	9			
Error	47.40	18	2.63		
Total	148.67	29			

Statistical evidence suggests that the number of steps that students walk changes as they use a pedometer, $F(2,18) = 10.87$, $p < .05$. Follow-up tests provide statistical evidence suggesting that students walk more after 6 and 12 weeks of using a pedometer compared with the first week, $t(18) = 3.15$, $p < .05$ and $t(18) = 4.52$, $p < .05$, respectively. No statistical evidence was found of a difference between 6 and 12 weeks, $t(18) = 1.36$, *n.s.*

27

Source	SS	df	MS	F	p
Between groups	74.08	2	37.04	2.70	*n.s.*
Within groups	261.88	21			
Between participants	69.96	7			
Error	191.92	14	13.71		
Total	335.96	23			

No statistical evidence was found to suggest that extroverts have different perceptions of satisfaction between past, present, and projected future life experiences, $F(2,12) = 2.70$, *n.s.* Since the overall null hypothesis could not be rejected, no follow-up analyses were run. (As a methodological aside, note that the scale used to measure perceptions of life satisfaction may not be interval or ratio. This is unclear in the problem.)

Part 5. Review of Analyses of Variance

1 c. and f. A one-way ANOVA can be thought of as an extension of an independent-samples *t* test into designs with three or more conditions, and a repeated-measures ANOVA can be thought of as an extension of a dependent-samples *t* test into designs with three or more conditions.

2 No. Although the mathematics are different, a *t* test and its corresponding ANOVA are equally powerful (statistically speaking). Mere convention keeps researchers using *t* tests with two-cell designs (see Section 12.1).

3 A mixed design is a form of a two-way ANOVA but one where one factor is repeatedly measured. So, the two-way ANOVA and repeated-measures ANOVA concepts are needed.

4 Don't be confused by how the data are presented. This is a 2×2 factorial design and requires a two-way ANOVA for analysis. Following is the ANOVA summary table.

Source	*SS*	*df*	*MS*	*F*	*p*
Factor *A*	85.56	1	85.56	20.17	<.05
Factor *B*	390.06	1	390.06	44.69	<.05
A × *B*	52.56	1	52.56	6.02	<.05
Within groups	104.75	12	8.73		
Total	632.93	15			

All three *F*'s direct the researcher to reject the corresponding null hypothesis. Of course, the main effects need to be carefully examined after the interaction is interpreted to see if they are illusory or not. A superficial examination of the cell and marginal means suggests that the drug imipramine outperforms vitamin B_1 (main effect) but especially so when the drug is combined with "talk" therapy (interaction). The other main effect (therapy style) seems to be an artifact of the interaction.

5 This is a repeated-measures design. It can be analyzed using either a dependent-samples *t* test or a repeated-measures ANOVA – each having the same degree of statistical power to avoid a Type II error. Conventionally, the *t* test is used in two-cell designs. The analysis found statistical evidence that participants stared longer at the image with a red background ($M = 3.31$) compared with the image with the off-white background ($M = 2.94$), $t(6) = 3.29$, $p < .05$.

6 This altered design from Problem 5 now requires us to run a repeated-measures ANOVA. There are three conditions now necessitating the ANOVA. The summary table is shown below.

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	1.65	2	0.83	8.23	< .05
Within groups	25.78	18			
Between participants	24.57	6			
Error	1.21	12	0.10		
Total	190.78	20			

The observed *F* directs the researcher to reject the null of no difference between the three conditions. Other statistical tools (e.g. *HSD* or *LSD*) can be used to evaluate mean differences between specific conditions. Fisher's *LSD* run with SPSS found statistical evidence suggesting that both of the bold colors, red ($M = 3.54$) and green ($M = 3.24$), were stared at longer than the off-white ($M = 2.86$). This analysis supports the researchers' contention that the difference in staring time may be due not to the color red, but simply due to the brightness of the color compared with the off-white.

7 This is a between-groups single-factor design. The appropriate statistical tool would be the one-way ANOVA. The summary table is shown below.

Source of variation	*SS*	*df*	*MS*	*F*	*p*
Between groups	52.10	2	26.05	6.78	<.05
Within groups(error)	69.14	18	3.84		
Total	121.24	20			

This *F* directs the researcher to reject the null of no difference between the three conditions. Other analytical tools (e.g. *HSD* or *LSD*) can be used to evaluate mean differences between specific conditions. Fisher's *LSD* run with SPSS found statistical evidence suggesting that the story condition (M = 8.43) takes longer for children to fall asleep than does the milk condition (M = 4.57).

8 The appropriate analysis for this design is a one-way ANOVA. There is only one factor, and participants are assigned to one condition. We may recall that a similar problem was presented at the end of Chapter 14 (repeated-measures ANOVA). The design has been changed, but the data is the same. The summary table is shown below.

Source of variation	*SS*	*df*	*MS*	*F*	*p*
Between groups	37.00	2	18.50	0.93	*n.s.*
Within groups(error)	299.50	15	19.97		
Total	336.50	17			

The design change created tremendous change. In the Chapter 14 problem, the *F* was rejected, and subsequent analysis found evidence of other differences. In the between-groups version, the *F* cannot be rejected. A careful examination shows us why; the within-groups *SS* is huge compared with the between-groups *SS*. Furthermore, most of this within-groups error is associated with individual differences. In the repeated-measures ANOVA, we see that over 279 units of this error is partitioned out, leaving us with only about 20 units of SS_{error}. But, in a one-way design, this error cannot be removed. No evidence of differences between conditions was found, $F(2,15) = 0.93$, *n.s.*

9 Do not be confused by how the data are presented. Reorganize the data to see that this is a between-groups design, with only two groups. It can be analyzed using either an independent-samples *t* test or a one-way ANOVA – each having the same degree of statistical power to avoid a Type II error. Conventionally, the *t* test is used in two-cell designs. The analysis did not find statistical evidence that stated gender was related to preparation

time for the dance, $t(13) = -2.04$, *n.s.* It is true that the obtained t value was very close to the rejection criteria. However, the needed probability of 5% was not met. It is possible that in failing to reject the null hypothesis, we will be making a Type II error. The increased power of more participants may have resulted in a different outcome.

10 This is a repeated-measures design, and there are more than two conditions. The proper statistical tool for analysis would be a repeated-measures ANOVA. The summary table is shown below.

Source	*SS*	*df*	*MS*	*F*	*p*
Between groups	21.72	2	10.86	3.53	*n.s.*
Within groups	66.66	18			
Between participants	29.71	6			
Error	36.95	12	3.08		
Total	88.38	20			

The observed F does not direct the researcher to reject the null hypothesis of no difference between the three conditions. The observed F is large, but not quite large enough. Perhaps there is not enough power to detect a genuine effect (despite the fact that the design allows for the powerful repeated-measures ANOVA). Stated more properly, no statistical evidence was found to suggest that the outdoor temperature influences the creativity in writing poems, $F(2,12) = 3.53$, *n.s.*

11 There are two factors involved in this design: first-mover and biological sex. Each is also a between-groups factor. The clear analytical choice is the two-way ANOVA. The summary table is shown below.

Source	*SS*	*df*	*MS*	*F*	*p*
Factor *A* (First Mover)	1.89	1	1.89	0.67	*n.s.*
Factor *B* (Bio Sex)	14.06	1	14.06	5.01	< .05
$A \times B$	0.00	1	0.00	0.00	*n.s.*
Within groups	33.66	12	2.81		
Total	49.61	15			

The only observed F that directs us to reject a null hypothesis is the one for Factor *B* (Biological Sex). Looking at the marginal means, the analysis has provided us with statistical evidence suggesting that biological male sellers (M = \$5.34) fare better than biological female sellers (M = \$3.47) during negotiations of this type, $F(1,15) = 5.01$, $p < .05$.

Chapter 15

1 In a correlational design, a study is conducted without exerting control over the phenomenon under investigation; data is gathered as it presents itself to the researcher. In an experimental design, some procedural variables are held constant, and others are purposely manipulated by the experimenter (e.g. drug dosages). By manipulating an independent variable in an experimental design, we are more likely to be able to show a causal relationship in the effect of the independent variable on the dependent variable.

2 **a** Causation is always determined by the methodological design, not the statistical analysis.

3 "Bivariate distribution" is the term used to describe a data set in which there are pairs of scores, each pair belonging to a given participant. Height and shoe size, for instance, are bivariate data when these measures come from the same sample of individuals.

4 No answer provided. Answers will vary. Positive correlations are marked by higher scores on one variable being associated with higher scores on a second variable.

5 No answer provided. Negative correlations are marked by higher scores on one variable being associated with lower scores on a second variable.

6 **b** Two of them are not legitimate correlations (correlations cannot exceed the absolute value of 1). Of the remaining two, –0.69 is stronger than 0.58.

7 A scatter plot of a +1 or –1 correlation would be data lined up perfectly. The slope would have to be either positive or negative, but the angle of the slope would be irrelevant.

8 An estimate of the magnitude of the correlation, direction of correlation (i.e. positive or negative), linearity, and presence of outliers. The angle of the slope is not informative.

9 $\rho = \dfrac{0.13}{4} = 0.03$

10 $\rho = \dfrac{-3.11}{4} = -0.78$

11

			r_{crit} at $\alpha = 0.05$	Reject H_0?	r_{crit} at $\alpha = 0.01$	Reject H_0?
a.	$r = 0.39$	$df = 100$	±0.19	Y	±0.25	Y
b.	$r = -0.47$	$df = 21$	±0.41	Y	±0.53	N
c.	$r = -0.09$	$df = 11$	±0.55	N	±0.68	N
d.	$r = 0.44$	$df = 6$	±0.71	N	±0.83	N
e.	$r = -0.62$	$df = 12$	±0.53	Y	±0.66	N
f.	$r = 0.93$	$df = 29$	±0.36	Y	±0.46	Y

12

a The scatter plot will have data points going from the lower left to the upper right. They will be gathered in an oval, not a straight line.

b The scatter plot will have data points going from the upper left to the lower right. They will be gathered in a tight oval, close to a straight diagonal line.

c The scatter plot will have data points going from the lower left to the upper right. They will be gathered in a slight oval, something close to a circle.

d The scatter plot will have data points that will gather into the shape of either a "U," upside down "U," "C," or backward "C." All of these reflect curvilinear relationships between two variables.

13 These are just rough estimates, designed to see if we understand the basic way in which scatter plots reflect correlations.

a $r = 0$

b $r = +0.50$

c $r = -0.50$

d $r = +0.90$

14

	r_{crit} 5%	r_{crit} 1%	r_{crit} 5%	r_{crit} 1%
a.	0.388	0.496	Reject	Reject
b.	0.388	0.496	Reject	Fail to reject
c.	0.532	0.661	Reject	Reject
d.	0.444	0.561	Reject	Reject
e.	0.355	0.456	Reject	Fail to reject
f.	0.666	0.798	Reject	Reject
g.	0.514	0.641	Reject	Fail to reject
h.	0.195	0.254	Reject	Fail to reject

(Critical values for h are approximate, based on $df = 100$.)

15 **a**

$\Sigma X = 24$	$\Sigma Y = 21$	$\Sigma XY = 125$
$(\Sigma X)^2 = 576$	$(\Sigma Y)^2 = 441$	$n_p = 4$
$\Sigma X^2 = 150$	$\Sigma Y^2 = 113$	

$$r_{obt} = \frac{-4}{16.25} = -0.25$$

b H_0: $\rho = 0$; H_1: $\rho \neq 0$
c r_{crit} (2) = ±0.950
d Do not reject the null hypothesis.
e $r^2 = (-0.25)^2 = 0.0625$ or 6.25%. (However, r^2 would not be reported since r is nonsignificant.)

16 **a**

$\Sigma X = 32$	$\Sigma Y = 25$	$\Sigma XY = 213$
$(\Sigma X)^2 = 1024$	$(\Sigma Y)^2 = 625$	$n_p = 4$
$\Sigma X^2 = 266$	$\Sigma Y^2 = 175$	

$$r_{obt} = \frac{52}{54.78} = 0.95$$

b H_0: $\rho = 0$; H_1: $\rho \neq 0$
c r_{crit} (2) = ±0.950
d Reject the null hypothesis when r is equal to or greater than r_{crit}.
e $r^2 = (0.95)^2 = 0.90$, or 90%

17 **a**

$\Sigma X = 43$	$\Sigma Y = 41$	$\Sigma XY = 453$
$(\Sigma X)^2 = 1849$	$(\Sigma Y)^2 = 1681$	$n_p = 4$
$\Sigma X^2 = 471$	$\Sigma Y^2 = 449$	

$$r_{obt} = \frac{49}{63.44} = 0.77$$

b H_0: $\rho = 0$; H_1: $\rho \neq 0$
c $r_{crit}(2) = \pm 0.950$
d Do not reject the null hypothesis.
e $r^2 = 0.593$ or 59.3%. (r^2 would not be reported since r is nonsignificant)

18 **a**

$\Sigma X = 20$	$\Sigma Y = 24$	$\Sigma XY = 110$
$(\Sigma X)^2 = 400$	$(\Sigma Y)^2 = 576$	$n_p = 5$
$\Sigma X^2 = 106$	$\Sigma Y^2 = 124$	

$$r_{obt} = \frac{70}{75.63} = 0.93$$

b H_0: $\rho = 0$; H_1: $\rho \neq 0$
c r_{crit} (3) = ±0.878
d Reject the null hypothesis.
e r^2 = 0.865 or 86.5%

19 Answers will vary. One example would be to measure a company's sales personnel in terms of their degree of extroversion (personality variable) and their effectiveness at selling (observed behavior). This is a correlational design (no manipulation of an independent variable) and involves one personality variable and one observed behavior.

20 Answers will vary. One example would be to manipulate dosages of a given drug to various individuals (e.g. 0, 5, 10, 20 ml) and then measure their heart rate or breathing (observed behavior). This is an experimental design (manipulation of an independent variable) and involves a medicinal variable and an observed behavior.

21 Causal interpretations are prohibited in Exercise 18 because the data is correlationally gathered. There is no manipulation. Any number of variables might explain the relationship between sales effectiveness and extroversion. Causal interpretation is allowed in Exercise 19 because the data is experimentally gathered. There is manipulation of a variable – which participant gets what level of the drug. If a correlation is found, we are justified in claiming the drug that is at least partially responsible for the change in observed behavior.

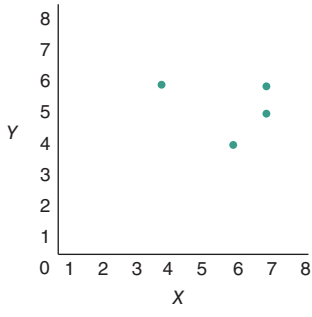
22 **a**

$\Sigma X = 18$	$\Sigma Y = 30$	$\Sigma XY = 96$
$(\Sigma X)^2 = 324$	$(\Sigma Y)^2 = 900$	$n_p = 5$
$\Sigma X^2 = 80$	$\Sigma Y^2 = 190$	

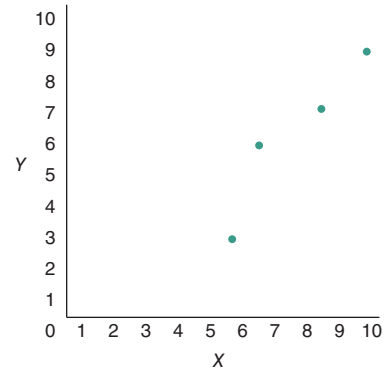
$$r_{obt} = \frac{-60}{61.64} = -0.97$$

b H_0: $\rho = 0$; H_1: $\rho \neq 0$
c r_{crit} (3) = ±0.878
d Reject the null hypothesis.
e r^2 = 0.941 or 94.10%

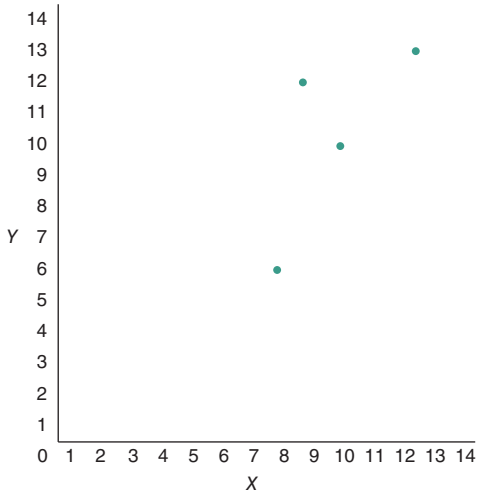
23 Refer to graph.



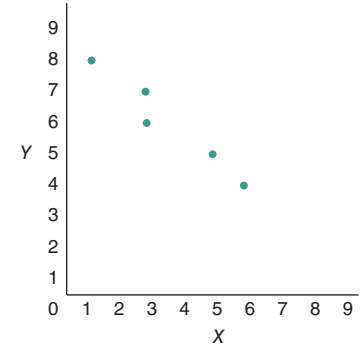
(Problem 15)



(Problem 16)



(Problem 17)



(Problem 22)

24 $\rho = \dfrac{64}{75} = 0.85$

25 A restricted range of scores may underestimate the size of the population correlation.

26 The use of extreme groups may increase the size of the correlation.

27 r will underestimate ρ.

28 **a**

$\Sigma X = 56$	$\Sigma Y = 503$	$\Sigma XY = 3887$
$(\Sigma X)^2 = 3136$	$(\Sigma Y)^2 = 253\,009$	$n_p = 7$
$\Sigma X^2 = 542$	$\Sigma Y^2 = 36\,365$	

$$r_{obt} = \frac{-959}{1008.60} = -0.95$$

b H_0: $\rho = 0$; H_1: $\rho \neq 0$
c r_{crit} (5) = ±0.754
d $r^2 = 0.9025$ or 90.25%
e Reject the null hypothesis.

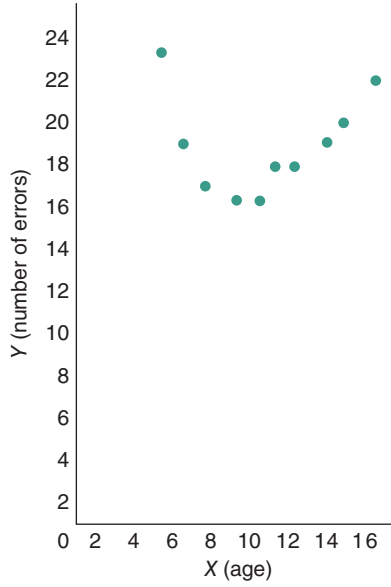
29 **a**

$\Sigma X = 18$	$\Sigma Y = 621$	$\Sigma XY = 2022$
$(\Sigma X)^2 = 324$	$(\Sigma Y)^2 = 385\,641$	$n_p = 6$
$\Sigma X^2 = 94$	$\Sigma Y^2 = 64\,979$	

$$r_{obt} = \frac{954}{1007.93} = 0.95$$

b H_0: $\rho = 0$; H_1: $\rho \neq 0$
c r_{crit} (4) = ±0.811
d $r^2 = 0.9024$ or 90.25%
e Reject the null hypothesis.

30 Refer to graph.



The variables appear to be related, but not in a linear fashion. The most reasonable course of action is to not run a Pearson r (it will most likely underestimate the relationship between these variables), but to rather run a curvilinear measure of association, something not covered in this text.

31 $r(11) = 0.87$, $p < .05$. Statistical evidence suggests that greater amounts of smoking correlate with a greater number of sick days taken by employees, $r^2 = 0.752$ or 75.2%.

32 $r(10) = 0.05$, *n.s.* There is no statistical evidence to suggest that a relationship exists between intelligence, as measured in this study, and word processing speed. No need to generate an r^2.

33 Drilling: $r(24) = 0.65$, $p < .05$
Rubber Dam: $r(24) = 0.46$, $p < .05$

34 $r(21) = 0.413$, $p < .05$

Chapter 16

1 The mean. It is the most frequent score in a normal distribution, and the error associated with it will likely be smaller than other guesses.

2 Regression analysis is important for making predictions to solve practical problems and to build and test theories.

3 Just as in Chapter 15, bivariate data is typically not experimentally gathered. Data gathered from correlational designs do not imply causation. If it is experimentally gathered, causal interpretations are implied.

4 The regression line is fitted to the scatter plot in such a way that the sum of the squared errors (Σe^2) is minimized.

5 Y intercept

6 $\Sigma(Y - Y_p)$ results in a summed error of 0.

7 s_e is the estimate of the average standard deviation for the set of conditional distributions associated with a given population of bivariate data.

8 A small standard error of the estimate means there is very little error associated with the prediction; a large standard error of the estimate means there is large error associated with the prediction.

9 If $r = 0$, then $b = 0$.

10 The slope is a ratio of the rise over the run for bivariate data placed on a scatter plot. It tells us the degree and direction of change in Y for each unit change in X. It does not tell us anything about the strength of the relationship between the two variables. (Of course when the null hypothesis $b = 0$ cannot be rejected, it does communicate that the relationship is either nonexistent or too weak to detect.)

11 Y_p is always the mean of the conditional distribution associated with a particular X value. It is merely a predicted value, but it is unbiased, and the prediction error associated with it is as small as possible.

12 When the null hypothesis that $b = 0$ cannot be rejected or when the null hypothesis that $r = 0$ cannot be rejected. A regression analysis should not be run if a relationship between the two variables cannot be established.

13 The assumption of normality for the conditional distributions is important because Y_p is the mean of each conditional distribution. For skewed distributions there is concern that the mean does not represent centrality well (see Chapter 3).

14 The assumption of homoscedasticity allows us to assume s_e is the same for every value of X. If each X were associated with conditional distributions of differing variances, there would have to be a different standard error of the estimate for each X score.

15 Multiple regression uses multiple predictor variables; each one is able to add to the predictive power.

16 Overgeneralizing to populations different from the populations used to establish the regression equation might lead to predictions that are no better than chance or systematically biased.

17 **a** $b = -0.22$
b $b = 0.26$
c $b = 0.55$

18 With a negative correlation, it is reasonable to conclude that higher gas prices will be associated with fewer miles driven for vacation and vice versa. However, caution is needed; we must not convince ourselves that we understand the causal relationship between these two factors.

19 **a** 2.28
b 3.24
c 3.88

20 **a** $b = 1.81$
b $Y_p = 8.83 + 1.81(X - 4.50)$
c 1.22

21 **a** $b = -0.84$
b $Y_p = 8.00 - 0.84(X - 8.67)$
c 3.39
d $Y_p = 8.00 - 0.84(1 - 8.67) = 14.44$

22

$M_X = 12.80$	$M_Y = 40.40$	
$\Sigma X = 64$	$\Sigma Y = 202$	$\Sigma XY = 2985$
$(\Sigma X)^2 = 4096$	$(\Sigma Y)^2 = 40\,804$	$n_p = 5$
$\Sigma X^2 = 990$	$\Sigma Y^2 = 9\,142$	

a $b = \dfrac{1997}{854} = 2.34$

b $Y_p = 40.40 + 2.34(0 - 12.80) = 10.45$

c $s_e = \sqrt{\left[\dfrac{1}{5(3)}\right]\left[(5)9142 - 40\,804 - \left(\dfrac{[5(2985) - (64)(202)]^2}{(5)990 - 4096}\right)\right]} = 3.97$

d $Y_p = 40.40 + 2.34(16 - 12.80) = 47.89$ seconds

e 47.89 ($\pm 1 s_e$) = 43.92 to 51.86

23

$M_X = 12$	$M_Y = 23$	
$\Sigma X = 72$	$\Sigma Y = 138$	$\Sigma XY = 1739$
$(\Sigma X)^2 = 5184$	$(\Sigma Y)^2 = 19\,044$	$n_p = 6$
$\Sigma X^2 = 886$	$\Sigma Y^2 = 3\,560$	

a $b = \dfrac{10\,434 - 9\,936}{5\,316 - 5\,184} = 3.77$

b $Y_p = 23 + 3.77(0 - 12) = -22.24$

c $s_e = \sqrt{\left[\frac{1}{6(4)}\right]\left[(6)3560 - 19\ 044 - \left(\frac{[6(1739) - (72)(138)]^2}{6(886) - 5184}\right)\right]} = 4.27$

d $Y_p = 23 + 3.77(10 - 12) = 15.46$; when converted to \$1000's = \$15 460

e 15.46 ($\pm 1.96 s_e$) = 7.091 to 23.829; when converted to \$1000's = \$7 091 to \$23 829

24

$M_X = 3.38$	$M_Y = 3.43$	
$\Sigma X = 16.88$	$\Sigma Y = 17.17$	$\Sigma XY = 58.09$
$(\Sigma X)^2 = 284.93$	$(\Sigma Y)^2 = 294.81$	$n_p = 5$
$\Sigma X^2 = 57.67$	$\Sigma Y^2 = 59.01$	

a $b = \frac{290.45 - 289.83}{288.35 - 284.93} = 0.18$

b $Y_p = 3.43 + 0.18(3.00 - 3.38) = 3.36$

Yes, since we would predict the student to achieve a GPA of 3.36 (3.00 minimum required).

c $s_e = \sqrt{\left[\frac{1}{5(3)}\right]\left[(5)59.01 - 294.81 - \left(\frac{[5(58.09) - (16.88)(17.17)]^2}{5(57.67) - 284.93}\right)\right]} = 0.092$

$Y_p = 3.43 + 0.18(3.67 - 3.38) = 3.48$

$3.48 \pm 1se$ = 3.388 to 3.572

25

a $b = 0.22$

b $Y_p = 2.31 + 0.22(X - 14)$

c $Y_p = 2.31 + 0.22(0 - 14) = -0.77$

d $s_e = 0.96$

e $Y_p = 2.31 + 0.22(15 - 14) = 2.53$

f $Y_p = 2.31 + 0.22(10 - 14) = 1.43$ $\quad 1.43 \pm 1s_e = 0.47$ to 2.39

26 $F(1,10) = 0.029$, *n.s.* The null hypothesis that $b = 0$ cannot be rejected, $r(10) = 0.05$, *n.s.* A regression analysis is not recommended.

27

a $b = 0.59$

b $Y_p = 46.33 + 0.59(X - 104.5)$

c $Y_p = 46.33 + 0.59(0 - 104.5) = -15.33$

d $s_e = 4.99$

e $Y_p = 46.33 + 0.59(100 - 104.5) = 43.68$

f $43.68 \pm 2s_e = 33.7$ to 53.66

28

a $b = -1.81$

b $Y_p = 5.00 - 1.81\ (X - 2.09)$

c $s_e = 0.79$
d $Y_p = 5.00 - 1.81\ (3 - 2.09) = 3.35$
e $3.35 \pm 2s_e = 1.77$ to 6.58
f Be aware that the subjective well-being scale may not be interval or ratio.

29 Drilling: $Y_p = 5.77 + 0.64(X - 5.15)$
$s_e = 1.98$

$$Y_p = 5.77 + 0.64(7 - 5.15) = 6.95$$

Rubber Dam: $Y_p = 5.42 + 0.47(X - 5.15)$
$s_e = 2.44$

$$Y_p = 5.42 + 0.47(7 - 5.15) = 6.29$$

30 $Y_p = 58.35 + 0.35(54 - 56.96) = 57.31$
$s_e = 16.65$
$Y_p = 57.31 \pm 16.65 = 40.66$ to 73.96

Part 6. Review of Linear Correlation and Linear Regression

1 **a**

2 In a word, prediction. A regression analysis allows us to leverage a known relationship between two variables for predictive use when a given value is unknown.

3 **a** Bivariate information regarding the performance of people who play both tennis and ping pong.
b Specific information regarding Sarah's tennis-playing ability – so it can be used in a regression analysis to predict future performance in ping pong.

4 There is not enough information given to make a prediction. (There is no bivariate data or even any descriptive words used.)

5 As the size of the correlation increases, prediction error decreases.

6 A correlation. Each variable can be measured, and the amount of shared variance between them can be represented indirectly with an r or directly with an r^2.

7 A one-way ANOVA. Each political category would be a condition; charitable donations would be the measured variable. Even though it is not an experiment, this design seems most appropriate.

8 Regression. Regression is the only statistical tool that will help in generating a predicted value for an unknown.

9 A two-way ANOVA. There are two variables that can be classified as categories (biological sex and political attitude). These will be the two dimensions. Even though it is not an experimental design, a two-way ANOVA would be best suited for this situation.

10 **a** −0.87
b $Y_p = 5.30 - 1.32(10 - 4.2) = -2.36$
c $t(9) = -0.59$, *n.s.* No statistical evidence of a difference was found.

11 **a** This one is tricky. If we decide to see these groups as being independent of each other (setting aside that each set of scores is a sibling pair), we would run an independent-samples t test to test the null that $\mu_{males} = \mu_{females}$. The result is $t(18) = 1.57$, *n.s.* The null of equal population means cannot be rejected. If, however, we see each pair of scores as a dependent set, then we can run a dependent-samples t test. The result is $t(9) = 2.68$, $p < .05$. The null of equal population means can be rejected.
b The null $\rho = 0$ can be rejected, $r(10) = +0.66$, $p < .05$.
c $Y_p = 2.90 + 0.77(2 - 3.72) = 1.58$ hours or about 1 hour and 35 minutes. Just enough for most movies!

12 **a** Yes, there is evidence of a relationship. The null of $\rho = 0$ can be rejected, $r(5) = -0.77$, $p < .05$.
b $r^2 = 0.59$. This is a measure of the shared variance. It is quite high, nearly 60%.
c $Y_p = 11.57 - 1.88(5 - 1.86) = 5.67$ or 5 and 2/3 of a month. Anyone familiar with developmental milestones in children realizes that this would be highly unusual. At this rate, a child born with 8 or 9 older siblings might be predicted to walk on day 1. Perhaps the relationship between these two variables is not linear, at least for higher values of X (sibling number).

13 **a** Yes, the null hypothesis that $\rho = 0$ can be rejected, $r(11) = -0.61$, $p < .05$. There is statistical evidence suggesting that lower retirement percentages and higher memory scores are linked.
b A regression analysis can be used to make predictions.

c $Y_p = 9.02 - 0.046(20 - 67.85) = 11.22$ (since the slope was so slight, three decimal places were used for the calculations); $Y_p = 9.02 - 0.046(40 - 67.85) = 10.30$. So a drop of nearly 1 full memory point would be expected.

d No. This data was gathered correlationally; no manipulation of a causal variable was employed. One could imagine retirement causing memory decline, memory decline causing early retirement, or some other variable or set of variables causing both of these variables to covary.

14 Assuming the standard deviation of the population is not known, a single-sample *t* test should be used.

Chapter 17

1 Population assumptions like normality and homogeneity of variance.

2 The expected frequencies of each cell are determined based on the total number of observed frequencies; therefore, the total number of expected frequencies will equal the total number of observed frequencies. The expected frequencies are a set of numbers that suggest what the observed frequencies should be in the available categories if the null is correct.

3 The chi-square goodness-of-fit test is used on categorical data spread across on dimension or factor. The chi-square test for independence is used on categorical data spread across two dimensions or factors.

4 Sample size is very important for determining the degrees of freedom for most inferential tests, but for the chi-square tests, the degrees of freedom are determined solely by the number of categories available.

5 Although the chi-square test is bidirectional, the measure of poor fit is squared – this makes each cell difference between the observed and expected frequency a positive value. Follow-up standardized residual analysis, however, can be used to show which way the cell is deviating from what is expected.

6 d

7 c Both are measures of effect size.

8 a Both are follow-up tests to better explain an effect.

9

f_o	f_e	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$	R
70	79.60	92.16	1.16	−1.08
160	159.20	0.64	0.004	0.06
168	159.20	77.44	0.49	0.70
30	20.40	92.16	4.52	2.13
40	40.80	0.64	0.016	−0.13
32	40.80	77.44	1.90	−1.38
			χ^2 = **8.09**	
			χ^2_{crit} = 5.99	

There are more unfavorable responses to Treatment I than would be expected by chance, $\chi^2(2, N = 500) = 8.09$; we have statistical evidence to reject the null hypothesis of no relationship. The cell, Unfavorable Response to Treatment I, makes a significant contribution to the significant χ^2.

Cramér's $V = 0.13$

10 Using the contingency table formula $\chi^2 = \dfrac{110(255-1520)^2}{(55)(55)(53)(57)} = 19.26$

$\chi^2_{crit} = 3.84 \qquad \chi^2(1, N = 110) = 19.26, \quad p < 0.05$

There is statistical evidence of a difference but opposite to the hypothesis. It seems biological females were more likely than biological males to keep their pencil.

Cramér's $V = 0.42$

11
a $df = 1$
b $df = 6$
c $df = 12$
d $df = 2$

12

	Design	χ^2_{obt}	df	χ^2_{crit}	Reject H_0?
a.	2 × 2	4.5	1	3.84	Yes
b.	3 × 3	9.0	4	9.49	No
c.	1 × 5	17.22	4	9.49	Yes
d.	2 × 4	5.55	3	7.82	No

13 **a** H_0: There is no relationship between type of bumper sticker and being stopped by the police.
H_1: The variables are related (not independent).

b and c

	Stop Brutality Sticker	Smile Sticker	
Stopped	18	5	23
Not Stopped	7	20	27
	25	25	$N = 50$

f_o	f_e	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
18	11.50	42.25	3.67
5	11.50	42.25	3.67
7	13.50	42.25	3.13
20	13.50	42.25	3.13
			$\chi^2 = 13.60$

$\chi^2_{crit} = 3.84$

d Statistical evidence suggests that drivers displaying "Stop Brutality" stickers are stopped more often than drivers displaying "Smile" stickers, $\chi^2(1, N = 50) = 13.60$, $p < .05$.

e Cramér's $V = 0.52$

14 **a** (f_o)

30	50	20	20	120
10	30	40	20	100
40	80	60	40	$N = 220$

(f_e)

21.82	43.64	32.73	21.82
18.18	36.36	27.27	18.18

b (f_o)

7	7	14
5	11	16
12	18	$N = 30$

(f_e)

5.60	8.40
6.40	9.60

15 **a** (f_o)

27	17	13	57
25	13	45	83
52	30	58	$N = 140$

(f_e)

21.17	12.21	23.61
30.83	17.79	34.39

b (f_o)

10	24	34
62	36	98
72	60	$N = 132$

(f_e)

18.55	15.45
53.45	44.55

16 2. Provided the two completed cells are not in the same row or column.

17 3; 5.

18

Spring	Summer	Fall	Winter	
160	190	170	130	
162.5	162.5	162.5	162.5	
f_o	f_e	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$	$(f_o - f_e)/\sqrt{f_e}$
160	162.5	6.25	0.04	−0.49
190	162.5	756.25	4.65	2.16
170	162.5	56.25	0.35	0.59
130	162.5	1056.25	6.50	−2.55
			$\chi^2 = 11.54$	

There is statistical evidence to reject the null hypothesis of no differences, $\chi^2 = (3, N = 650)\ 11.54, p < .05$. Follow-up analyses suggest that more people buy RV's in the summer than would be expected by the null and fewer people buy them in the winter than would be expected by the null. The advice to dealers depends on other issues. If there is a need to even out sales throughout the year, the dealers would be advised to advertise more heavily in the winter. If they want to take advantage of the peak buying season, they might be advised to advertise more heavily in the summer.

19

	Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
f_o	56	29	17	22	25	15	33
f_e	28.14	28.14	28.14	28.14	28.14	28.14	28.14

f_o	f_e	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
56	28.14	776.18	27.58
29	28.14	0.74	0.03
17	28.14	124.10	4.41
22	28.14	37.70	1.34
25	28.14	9.86	0.35
15	28.14	172.66	6.14
33	28.14	23.62	0.84
			$\chi^2 = 40.69$

$\chi^2_{crit} = 12.59$

$\chi^2(6, N = 197) = 40.69, p < .05$; statistical evidence exists to reject the null hypothesis.

20

	Accident	No Accident	
Rain	29	35	64
No Rain	31	48	79
	60	83	143

$$\chi^2 = \frac{143(1\,392 - 1\,085)^2}{(64)(79)(60)(83)} = \frac{13\,477\,607}{25\,178\,880} = 0.54$$

$\chi^2_{crit} = 3.84$; $\chi^2(1, N = 143) = 0.54$, *n.s.*

21

f_o	f_e	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$	R
2	8.30	39.69	4.78	−2.18∗
20	26.08	36.97	1.42	−1.19
27	14.62	153.26	10.48	3.24∗
19	12.70	39.69	3.13	1.77
46	39.92	36.97	0.93	0.96
10	22.38	153.26	6.85	−2.62∗
			$\chi^2 = 27.59$	
			$\chi^2_{crit} = 5.99$	

Statistical evidence suggests that there are fewer easy births among primiparous mothers, fewer difficult births among multiparous mothers, and more difficult births among primiparous mothers than would be expected by chance, $\chi^2(2, N = 124) = 27.59$, $p < .05$. Those R values with asterisks reflect cells that make a significant contribution to the χ^2 value.

22

	Correct ID	Incorrect ID	
Primiparous	8	19	27
Multiparous	34	9	43
	42	28	$N = 70$

$$\chi^2 = \frac{70(72 - 646)^2}{(27)(43)(42)(28)} = 16.89$$

Yes, $\chi^2_{crit} = 3.84$; $\chi^2(1, N = 70) = 16.89$, $p < 0.05$.

23 Because we are summing across all cells, χ^2 will increase as the number of categories increases. Therefore, χ^2_{crit} needs to become corresponding larger as well.

24 Yes, assuming each participant is only counted in one category, $\chi^2 = 613.65$, $p < .05$; the null hypothesis is rejected. A residual analysis found evidence suggesting that more people than would be expected get tattoos for familial reasons and less people than would be expected get tattoos for conformity or personal expression reasons.

25 Younger mothers give birth to more physically immature and fewer physically mature babies than would be expected by chance. Older mothers give birth to more physical mature and fewer physically immature babies than would be expected by chance, $\chi^2(1, N = 114) = 26.67$, $p < .05$. Cramér's $V = 0.48$.

26 There is an association between diabetes and prolonged healing, with diabetics showing longer healing times, $\chi^2(1, N = 810) = 137.08$, $p < .05$.

27 There is no differential effect due to treatment, $\chi^2(1, N = 172) = 0.75$, *n.s.*

28 $f_o =$

Part of the country	Tennis	Golf	
Northeast	7	9	16
Southeast	6	18	24
Southwest	20	25	45
Midwest	15	20	35
	48	72	120

$f_e =$

Part of the country	Tennis	Golf	
Northeast	6.4	9.6	16
Southeast	9.6	14.4	24
Southwest	18	27	45
Midwest	14	21	35
	48	72	120

There is no statistical evidence of a relationship between the part of the country and the sport most enjoyed (considering only golf and tennis), $\chi^2(3, N = 120) = 2.84$, *n.s.*

29 The key to answering this question is to realize that the results can be converted into numbers representing categories. The chi-square analysis found statistical evidence of a relationship between education level and gun rights position, $\chi^2(2, N = 660) = 11.24$, $p < .05$. Follow-up residual analysis suggests that people with a graduate degree are more in favor of gun control than a hypothesis of no relationship would predict. Cramér's V measure of the effect size is 0.41.

Chapter 18

1 The raw data are organized from lowest to highest (or highest to lowest), and ranks are applied orderly. When ties occur, the next two ranked positions are averaged, and each old value is replaced with the averaged rank. The same process is applied when three or more ties occur with the same scaled value. Take the corresponding number of ranked positions needed, find the average of those ranked positions, and assign each scaled value the same new averaged ranked value.

2 Outlier data, which in an interval or ratio scale lies far away from the rest of the values in the distribution, is brought to within just one value of the rest of the data once it is ranked. This is sometimes seen as a solution for researchers who have an outlying data point.

3 **a**

4 **c**

5 The Mann–Whitney U test.

6 The parametric test should be selected every time (assuming all assumptions have been met). Nonparametric tests, because of the precision loss associated with ranked data compared with interval or ratio data, are less statistically powerful. In other words, the Type II error rate will be larger with a nonparametric test.

7

X	R_X	Y	R_Y
3	2	7	4
2	1	2	1.5
4	3.5	4	3
9	6	12	6
8	5	8	5
4	3.5	2	1.5

8

Score	Rank	Condition
2	2	*X*
2	2	*Y*
2	2	*Y*
3	4	*X*
4	6	*X*
4	6	*X*
4	6	*Y*
7	8	*Y*
8	9.5	*X*
8	9.5	*Y*
9	11	*X*
12	12	*Y*

9

X	Y	D	Rank
3	7	–4	–6
2	2	0	2
4	4	0	–2
9	12	–3	–5
8	8	0	– (Discard)
4	2	2	4

10

X	R_X	Y	R_Y
77	3	45	6.5
54	8	45	6.5
96	1.5	83	3
12	10	37	8
73	5	93	1
76	4	14	10
56	7	52	5
96	1.5	85	2
68	6	62	4
15	9	19	9

11

X	R_X	Y	R_Y	D	D^2
77	3	45	6.5	−3.5	12.25
54	8	45	6.5	1.5	2.25
96	1.5	83	3	−1.5	2.25
12	10	37	8	2	4
73	5	93	1	4	16
76	4	14	10	−6	36
56	7	52	5	2	4
96	1.5	85	2	−0.5	0.25
68	6	62	4	2	4
15	9	19	9	0	0
					$\Sigma D^2 = 81$

$$r_s = 1 - \frac{6\Sigma D^2}{n_p\left(n_p - 1\right)} = 1 - \frac{6(81)}{10(100-1)} = 1 - \frac{486}{990} = 0.51$$

No, the null hypothesis cannot be rejected. The Spearman critical value for $\alpha = 0.05$ two-tailed test is 0.648.

No, it would not have mattered if the ranking had been reversed; the same value would have been produced. If one variable was ranked highest to lowest and the other ranked lowest to highest, the formula would have generated a –0.51.This would have been confusing for interpretation purposes.

12

X	R_X	Y	R_Y
11	2	19	9
14	5	10	1
16	7.5	15	5.5
11	2	15	5.5
15	6	16	7
16	7.5	11	2
18	10	14	4
11	2	19	9
12	4	19	9
17	9	12	3

13

X	R_X	Y	R_Y	D	D^2
11	2	19	9	–7	49
14	5	10	1	4	16
16	7.5	15	5.5	2	4
11	2	15	5.5	–3.5	12.25
15	6	16	7	–1	1
16	7.5	11	2	5.5	30.25
18	10	14	4	6	36
11	2	19	9	–7	49
12	4	19	9	–5	25
17	9	12	3	–6	36
					$\Sigma D^2 = 258.5$

$$r_s = 1 - \frac{6\Sigma D^2}{n_p(n_p - 1)} = 1 - \frac{6(258.5)}{10(100 - 1)} = 1 - \frac{1551}{990} = -0.57$$

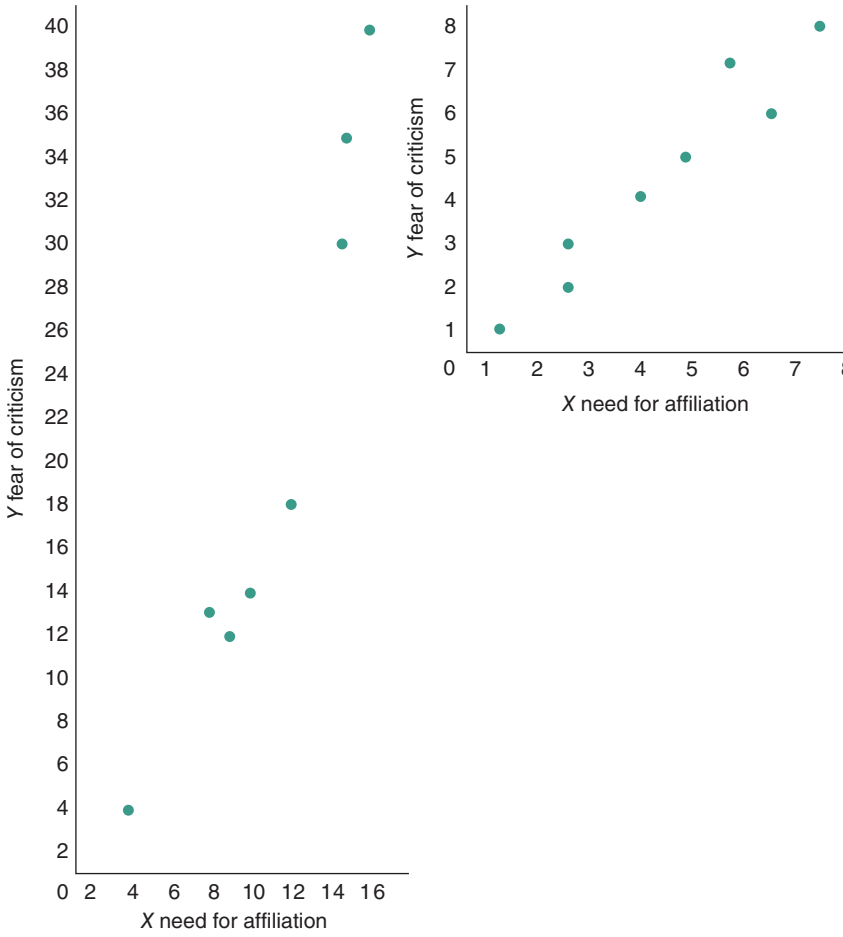
No, the null hypothesis cannot be rejected. The Spearman critical value for $\alpha = 0.05$ two-tailed test is 0.648. This may be surprising; the numbers seem to be significantly negatively correlated. However, the relationship is not strong enough to count as evidence against the null hypothesis. This exercise may give us a sense of just how little statistical power is found in nonparametric tests.

14 Here is the worked-out formula for the point-biserial correlation:

$$r_{pb} = \frac{M_{Y_1} - M_{Y_0}}{s_y}\sqrt{\frac{n_1 n_0}{n(n-1)}} = \frac{8.5 - 6.11}{3.47}\sqrt{\frac{9(8)}{17(16)}}$$

$$= \frac{2.39}{3.47}\sqrt{\frac{72}{272}} = 0.69(0.51) = 0.36$$

15 Scatter plots for a. and c.



b

Affiliation		Criticism			
Score	**Rank**	**Score**	**Rank**	***D***	***D*²**
16	1	40	1	0	0
14	2.5	35	2	0.5	0.25
14	2.5	30	3	–0.5	0.25
12	4	18	4	0	0
10	5	14	5	0	0
8	7	13	6	1	1
9	6	12	7	–1	1
4	8	4	8	0	0
				$\Sigma D = 0$	$\Sigma D^2 = 2.5$

c Refer to scatter plot above.

d $r_s = 1 - \frac{6(2.5)}{8(8^2-1)} = 0.97$

e H_0: $\rho_s = 0$; H_1: $\rho_s \neq 0$

f $r_{s_{crit}} = 0.738$; $r_s(8) = 0.97$, $p < 0.05$.

16

Attractiveness	Popularity		
Rank	**Rank**	***D***	***D*²**
1	1	0	0
2	3	–1	1
5	2	+3	9
3	4	–1	1
4	5	–1	1
7	7	0	0
9	6	+3	9
6	8	–2	4
8	9	–1	1
10	10	0	0
		$\Sigma D = 0$	$\Sigma D^2 = 26$

a $r_s = 1 - \frac{6(26)}{10(10^2-1)} = 0.84$

b H_0: $\rho_s = 0$; H_1: $\rho_s \neq 0$

c $r_{s_{crit}} = 0.65$; $r_s(10) = 0.84$, $p < 0.05$.

d There is evidence of a significant, positive correlation between physical attractiveness and popularity, $r_s(10) = 0.84$, $p < .05$.

17

Democrat	Republican
$M_{Y_1} = 2.71$	$M_{Y_0} = 7.86$
$n_1 = 7$	$n_0 = 7$

$s_y = 3.34$; $N = 14$

a $r_{pb} = \dfrac{2.71 - 7.86}{3.34}\sqrt{\dfrac{(7)(7)}{(14)(13)}} = -0.80$

b H_0: $\rho = 0$; H1: $\rho \neq 0$

c $r_{crit} = 0.532$; $r_{pb}(12) = -0.80$, $p < .05$

d There is statistical evidence of a negative correlation between political affiliation and attitudes toward military intervention in Central America, with Republicans favoring more aggressive intervention, $r_{pb}(12) = -.80$, $p < .05$.

e $r_{pb}^2 = (-0.80)^2 = 0.64$ or 64%

18

Correct	Incorrect
36	16
39	14
22	26
30	9
	7
	11

$\Sigma Y_0 = 127$; $\Sigma Y_1 = 83$; $M_{Y_0} = 31.75$; $M_{Y_1} = 13.83$; $n_0 = 4$; $n_1 = 6$; $s_y = 11.40$; $N = 10$

a $r_{pb} = \dfrac{13.83 - 31.75}{11.40}\sqrt{\dfrac{(6)(4)}{(10)(9)}} = -0.82$

b H_0: $\rho = 0$; H_1: $\rho \neq 0$

c $r_{crit} = 0.632$; $r_{pb}(8) = -0.82$, $p < .05$

d There is statistical evidence of a negative correlation between the answer to the critical question and the total test score on the test, with an incorrect answer associated with lower total test scores, $r_{pb}(8) = -.82$, $p < .05$. [Note: In interpreting this correlation, we should not be

thinking, "Of course missing a question will lead to a lower overall score, this is a trivial finding." Since the difference between M_{Y_0} and M_{Y_1} is not one point, clearly something else is going on here. Most likely, those students who answer the critical question correctly are more likely to answer other questions correctly (note the much higher mean (31.75) for the correct group.)]

19

a $r_s = 1 - \frac{6(53.5)}{9(9^2 - 1)} = 0.55$

b H_0: $\rho_s = 0$; H_1: $\rho_s \neq 0$

c $r_{s_{crit}} = 0.70$; $r_s(9) = 0.55$, *n.s.*

d There is no evidence found suggesting that posture and body shape are related, $r_s(9) = 55$, *n.s.*

20

a $r_s = 1 - \frac{6(62)}{12(12^2 - 1)} = 0.78$

b H_0: $\rho = 0$; H_1: $\rho \neq 0$

c $r_{s_{crit}} = 0.587$; $r_s(12) = 0.78$, $p < .05$

d There is statistical evidence of a positive correlation between performance on clinical and written exams, $r_s(12) = 0.78$, $p < .05$.

21

a $r_{pb} = \frac{13.65 - 11.47}{1.89}\sqrt{\frac{(10)(10)}{(20)(19)}} = 0.59$

b H_0: $\rho = 0$; H_1: $\rho \neq 0$

c $r_{crit} = 0.444$; $r_{pb}(18) = 0.59$, $p < .05$

d There is statistical evidence of a positive correlation between age a child first walks and the presence or absence of an older sibling, with earlier age of walking associated with the presence of an older sibling, $r_{pb}(18) = 0.59$, $p < .05$.

22

a $r_{pb} = \frac{6.6 - 5.0}{3.08}\sqrt{\frac{(5)(5)}{(10)(9)}} = 0.28$

b H_0: $\rho = 0$; H_1: $\rho \neq 0$

c $r_{crit} = 0.632$; $r_{pb}(8) = 0.28$, *n.s.*

d No statistical evidence was found to suggest that expressed gender and attitudes about state-mandated paid maternity leave are related, $r_{pb}(8) = 0.28$, *n.s.*

23

a $r_s = 1 - \frac{6(44)}{8(8^2 - 1)} = 0.48$

b H_0: $\rho_s = 0$; H_1: $\rho_s \neq 0$

c $r_{crit} = 0.738$; $r_s(8) = 0.48$, *n.s.*

d Since r_s is nonsignificant, there would be little point in reporting r_s^2. However, for the sample, it appears that about 23% of the variance is shared between the two measures.

24 **a** H_0: The population distribution of *A* (behavioral therapy) is the same as the population distribution of *B* (psychoanalysis).
H_1: The population distribution of *A* is not the same as the population distribution of *B*.

a Because the data is in ranks, and the null hypothesis is not stated in terms of a relationship, but rather in terms of differences, the Mann–Whitney *U* test should be used instead of a point-biserial correlation.

Rank:	1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12
Condition:	*B*,	*B*,	*B*,	*A*,	*B*,	*B*,	*A*,	*A*,	*A*,	*B*,	*A*,	*A*

ΣR_A	ΣR_B
4	1
7	2
8	3
9	5
11	6
12	10
51	27

$$n_A = 6 \quad n_B = 6$$

$$U_A = (6)(6) + \frac{6(6+1)}{2} - 51 = 6$$

$$U_B = (6)(6) + \frac{6(6+1)}{2} - 27 = 30$$

$$U = 6$$

a $U_{crit} = 5$ (with $df = 6,6$) $U = 6$, *n.s.*

b There was no statistical evidence found of a difference in interviewing skills based on training track, $U(6,6) = 6$, *n.s.*

25 When using the Mann–Whitney *U* test when one of the samples is greater than 20, transform the *U* value to a *z* value, and use critical *z* values as cutoffs. This can be done because a sample size greater than 20 yields a sampling distribution of *U* that approximates a normal distribution.

26 With three scores having difference scores of 0, discard one and assign one of the remaining two ranks to the Positive group and the other to the Negative group.

27 Both the smaller ΣR and U will equal 0.

28

Rank:	1,	2,	3,	4,	5,	6,	7,	8,	9,	10,
Condition:	*A*,	*B*,	*A*,	*A*,	*A*,	*B*,	*B*,	*B*,	*A*,	*B*,

ΣR_A	ΣR_B
1	2
3	6
4	7
5	8
9	10
22	33
$n_A = 5$	$n_B = 5$

$$U_A = (5)(5) + \frac{5(5+1)}{2} - 22 = 18$$

$$U_B = (5)(5) + \frac{5(5+1)}{2} - 33 = 7$$

$$U = 7$$

$$z_U = \frac{7 - (5)(5)/2}{\sqrt{[(5)(5)][(5+5+1)]/12}} = -1.15$$

$$z_{crit} = 1.96$$

Since $z_U < z_{crit}$, do not reject the null hypothesis.

29

	Music			
Participant	**New Age**	**Hip Hop**	**Difference**	**Rank**
P_1	2	6	–4	–4
P_2	1	6	–5	–6
P_3	3	2	1	1
P_4	4	8	–4	–4

(*Continued*)

(Continued)

	Music			
Participant	New Age	Hip Hop	Difference	Rank
P_5	3	6	−3	−2
P_6	1	5	−4	−4

T = smallest ΣRanks = 1

$T_{crit} = 0$ (with $df = 6$); $T = 1$, *n.s.* There is no statistical evidence to suggest that type of music has an effect on how fast students eat, $T(6) = 1$, *n.s.* This problem seems to be another example of the power weakness associated with nonparametric tests.

30

$$U_A = (6)(6) + \frac{6(6+1)}{2} - 24.5 = 32.50$$

$$U_B = (6)(6) + \frac{6(6+1)}{2} - 53.5 = 3.5$$

$$U_{crit} = 5 \ (with \ df \ = 6, 6)$$

$$U = 3.5; \ p < 0.05.$$

31
a Mann–Whitney *U* test
b Wilcoxon signed-ranks test
c Spearman rank correlation coefficient
d Point-biserial correlation coefficient
e Spearman rank correlation coefficient

32

$$z_{obt} = \frac{14 - 55(55+1)/4}{\sqrt{[(55)(55+1)][(2(55)+1)]/24}} = -6.33$$

$Z_{crit} = 1.96$. Therefore, we have statistical evidence to reject the null hypothesis.

33

New T-shirt	Control
8	9
10	6
19	22
7	11

(*Continued*)

(Continued)

New T-shirt	Control
1	23
4	24
5	12
3	14
13	2
20	15
21	18
16	17

Rank:	1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12,
Condition:	*A*,	*B*,	*A*,	*A*,	*A*,	*B*,	*A*,	*A*,	*B*,	*A*,	*B*,	*B*,
Rank:	13,	14,	15,	16,	17,	18,	19,	20,	21,	22,	23,	24
Condition:	*A*,	*B*,	*B*,	*A*,	*B*,	*B*,	*A*,	*A*,	*A*,	*B*,	*B*,	*B*

$$\Sigma R_A = 1 + 3 + 4 + 5 + 7 + 8 + 10 + 13 + 16 + 19 + 20 + 21 = 127$$
$$\Sigma R_B = 2 + 6 + 9 + 11 + 12 + 14 + 15 + 17 + 18 + 22 + 23 + 24 = 173$$

$$U_A = (12)(12) + \frac{12(12+1)}{2} - 127 = 95$$

$$U_B = (12)(12) + \frac{12(12+1)}{2} - 173 = 49$$

$$U = 49$$

U_{crit} = 37; fail to reject the null hypothesis. No statistical evidence has been found to reject the null hypothesis that suggests that there is no difference in happiness between those wearing new t-shirts and those who are not.

34 **a** H_0: The population distribution of *A* (propranolol) is the same as the population distribution of *B* (diuretic).

H_1: The population distribution of *A* (propranolol) is not the same as the population distribution of *B* (diuretic).

b $T = 8.5$, $p < .05$

c $T_{crit} = 25$ (with $df = 15$)

d There is statistical evidence of a difference in effectiveness of propranolol and diuretic for lowering systolic blood pressure, with propranolol being more effective, $T(15) = 8.5$, $p < .05$.

35 $U_{crit} = 64$ (with $df = 15,15$)
$U_A = 165$, $U_B = 60$
Since $U_B < U_{crit}$ reject the null hypothesis. There is statistical evidence to suggest that propranolol is more effective than a diuretic.

36 $z_{obt} = -2.21$; $p < .05$. We have found statistical evidence to reject the null hypothesis. It appears that the drug does a better job of helping alcoholics avoid drinking than the vitamin.

Part 7. Review

1 A parametric hypothesis test relies not only on methodological assumptions but also on certain statistical assumptions associated with the type of scaled data gathered (interval or ratio) and the features of the populations from which the data was gathered – normal distributions and homogeneity of variances across the populations. Parametric tests make use of the mean and other mean-based statistics like the variance and standard deviation. Nonparametric tests still rely on methodological assumptions, but do not make any statistical assumptions related to the type of scaled data gathered or the features of the population of data. Nonparametric statistics can be applied to frequency count data as well as ranked data (ordinal scale).

2 Parametric tests are to be preferred simply because they have more power. If the null hypothesis is wrong, parametric tests are more likely to generate statistical evidence to reject the null. This typically makes nonparametric tests a "Plan B" option, something to be resorted to if needed.

3
a 5
b 2
c 8
d 3
e 4
f 9
g 1
h 6

4 Please note, this is not a *t* test or Mann–Whitney *U* test – it is a correlation. Because one variable is dichotomous, the point-biserial correlation needs to be run. A computerized analysis is found, $r(12) = 0.51$, *n.s.* It is close to the critical value of ±0.53, but did not reach it. So the null must not be rejected.

5 The Mann–Whitney U test is the proper analytical tool to use. Analysis by computer generated a $U = 10$. This is not smaller than U_{crit}, which is 5. The null hypothesis cannot be rejected. This was also the case for the same data presented in the review for Part 4, question 10.

6 Use the chi-square test for independence for this problem. Yes, the null can be rejected. The $\chi^2 = 30.81$, much greater than the $\chi^2_{crit} = 15.51$ for 8 (4 × 2) degrees of freedom. Further analysis shows more married people than would be expected if the two dimensions were independent own dogs and fewer divorced or widowed people than would be expected if the two dimensions were independent own dogs.

7 The Wilcoxon is the proper test to run. A computerized analysis produced a $z_{obt} = -1.86$, *n.s.* The null hypothesis cannot be rejected.

8 The proper test for this situation would be the chi-square goodness-of-fit test. The obtained chi-square is 6.73; the critical chi-square for 4 degrees of freedom is 9.49. We cannot reject the null hypothesis of no differences between book types.

9 The proper test depends on how the researcher interprets the scale of measurement for stress. A conservative approach would suggest that this scale is ordinal and so the test of choice should be the Spearman. When run, the analysis generates an $r = 0.50$, *n.s.* ($r_{crit} = \pm 0.70$). A more permissive approach would treat the stress measure as an interval scale and a Pearson r could be run. In this case the $r = 0.68$, $p < .05$ ($r_{crit} = \pm 0.67$). The fact that the more standard measure allows for a rejection, while the less powerful nonparametric does not, highlights the cost that is incurred when using nonparametrics.

10 Chi-square goodness of fit. It is categorical data and there is only one factor.

11 If the number of burps is being counted, the proper test should be the independent-samples t test. There are two independent groups, and the dependent variable is being measured on a ratio scale. If, however, there were reasons to believe that a statistical assumption would be violated, the Mann–Whitney U test would be the preferred analytical tool.

12 The proper test would be a chi-square goodness-of-fit test. Students are either biological males or females, and the assumed null would be that there should be equal numbers of both in the senior class of this major.

13 **a** Answers will vary. One possible answer would be placing students into a preferred social media category while also categorizing students based on major area of study or some other categorical variable (e.g. personality type). This would be a two categorical variables study – social media category and major of study.

b Answers will vary. One possible answer would be to place students into one of the two different types of social media platforms and then measure the amount of content they provide within a given time period.

c Answers will vary. A Wilcoxon test would be run on any repeated-measures design where the data was suspected not to meet the statistical assumptions of the dependent-samples *t* test. The answer to "b" could be modified to make it a repeated-measures design, and the data measured (input) could be ranked instead of measured as a continuous variable.

14 **a** Answers will vary. One possible answer would be to create a National League and American League category and simply count the times the fantasy league was won by a team with a roster predominately composed of players from one league or the other.

b Answers will vary. One possible answer would be to create a bivariate database with a "0" or "1" to represent a roster of players from predominately one league or the other and the other variable being the number of times that team has won the league. (This would require fantasy league owners to keep the same drafting strategy across years.)

c Answers will vary. One possible answer would be to create a bivariate database with a ranked representation of where each team finished in a given year as the other variable and the percentage of National League players on the roster that year as the other. Because the finish position is ranked, a Spearman would need to be used.

15 Given the manner in which these variables are being measured, a chi-square test for independence is the analytical tool to use.

16 In this situation, it appears that there is a continuous measure (literacy) as well as a dichotomous measure (marital status). A point-biserial correlation would be the tool of choice.

Appendix C

Basic Data Entry for Microsoft® Excel and SPSS®

Microsoft® Excel

Following are some guidelines for how to enter data into Excel:

1) Open a new file (or an existing file if adding additional data).
2) Activate the cell in which we intend to input data.
3) Input the data by highlighting the cell where data is to be placed and simply typing in the information. There are three types of information that might be imputed: text, numeric, or formulas. The term "data" typically refers to numeric information, but sometimes words can be considered data, especially if a variable is categorical.
4) Hitting the return key will feature the cell directly below the cell currently being featured. Hitting the tab key will feature the cell immediately to the right of the cell currently being featured. The mouse can also be used to move the cursor over the cell that needs to be featured. Simply click the mouse and the cell will become activated.
5) Data is typically organized such that each participant is assigned one row and each variable is assigned on column. Column names are typically added across the top row to help identify the data. Participant numbers are often not inputted due to the already existing numbering system down the left-hand column. However, sometimes it is important to create a specific column for participant identification purposes. This is usually the leftmost column.
6) If the data is repetitive, patterned, or serial, shortcuts can be used to input the data. For instance, we can place the first set of values or pattern, and while it is highlighted, we can grab the lower right-hand corner of the grid and drag the box downward.
7) Information can be edited while the cell is activated through standard word processing editing procedures, or, if the data has already been registered, a

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

cell can be edited by activating it again using the mouse and then going up to the data entry cell at the top of the spreadsheet and making the edits.

8) To highlight a set of data, simply place the cursor over one corner of a contiguous set of data, and then drag the cursor in the direction needed to cover the cells needed.
9) There are cutting and pasting functions available. In fact, there are numerous data management tools available to the user. These are identified across the top of the file. The most useful ones for inputting data can be found under "Home," "Formulas," and "Data."
10) Under "Data" look to see if there is a "Data Analysis" option. If there is not, we will need to add it. Please see your Excel User's Manual to determine how to install the "Data Analysis ToolPak." This feature is required to run many of the statistical tests discussed in the textbook.
11) Excel is a very flexible and sophisticated data organization tool. This is but a brief introduction into data entry and organization. Please consult the user's manual or any number of online tutorials to understand further the features of this very helpful program.

SPSS®

Following are some guidelines for how to enter data into SPSS:

1) The first important thing to notice about SPSS is the two different views that can be reached by clicking boxes in the lower left of the screen: the "Data View" and the "Variable View."
2) The "Variable View" should be accessed first (even though the program typically opens with the "Data View" activated). Here is where variables are named. They are also identified by Type (Numeric is the default value), and many other specifications can be made. For instance, we can note what various categorical values mean in the "Values" column, and we can identify what type of scaling was used in for the measure under the "Measure" column. The default is "Scale," which means interval or ratio. "Ordinal" or "Nominal" can be selected.
3) Toggling to the "Data View" allows us to input data. Here the procedure is very similar to any other data base program. We can use the return key to move down a column and the tab key to move to the right along a row. The mouse can also be used to move around the grid. Data can be imputed into a cell once it is highlighted.
4) Typically, data from the same participant is placed along the same row. This is important for any repeated-measures or bivariate data analysis. In fact, data along the same row will be assumed to come from either the same participant or participants who were matched ahead of time. For this reason,

data from independent groups will need to be identified through a second variable, oftentimes labeled as "Condition." In this way, scores independently gathered across two or more conditions will populate one column, and nominal values such as "0," "1," "2," and so on will populate another column.

5) SPSS, unlike Excel, is a program expressly constructed for the statistical analysis of data. As a result, it is quite sophisticated. To understand better this power and flexibility of this program, one is encouraged to purchase a tutorial resource and/or to use tutorial videos found online.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Addison, W. E. (1989). Beardedness as a factor in perceived masculinity. *Perceptual and Motor Skills*, *68*, 921–922.

American Psychological Association (2009). *Publication manual of the American Psychological Association*. Washington, DC: American Psychological Association.

Anderson, C. A., & Bushman, B. J. (2002). Human aggression. *Annual Review of Psychology*, *53*, 27–51.

Aronson, E., & Mills, J. (1959). The effects of severity of initiation on liking for the group. *Journal of Abnormal and Social Psychology*, *59*, 177–181.

Baddeley, A. D., & Longman, D. J. A. (1978). The influence of length and frequency of training sessions on the rate of learning to type. *Ergonomics*, *21*(8), 627–635.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437.

Baker, M. (2015, August 27). Over half of psychology studies fail reproducibility test. *Nature*, Retrieved from www.nature.com/news/over-half-of-psychology-studies-fail-reproducibility-test-1.18248.

Balague-Dahlberg, G. (1986). Transfer of biofeedback training of heart rate decrease (Master's thesis). University of Illinois at Chicago.

Bandura, A. (1973). *Aggression: A social learning analysis*. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1983). Psychological mechanisms of aggression. In R. G. Green, & E. Donnerstein (Eds.), *Aggression: Theoretical and empirical views* (pp. 1–40). New York: Academic Press.

Barber, T. X., & Hahn, K. W. Jr. (1962). Physiological and subjective responses to pain-producing stimulation under hypnotically suggested and waking-imagined analgesia. *Journal of Abnormal and Social Psychology*, *59*, 177–181.

Baron, R. S., Logan, H., & Kao, C. F. (1990). Some variables affecting dentists' assessment of patients' distress. *Health Psychology*, *9*(2), 143–153.

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

Baslet, G. (2011). Psychogenetic non-epileptic seizures: A model of their pathogenic mechanism. *Seizure*, *20*, 1–13.

Batson, C. D., Eklund, J. H., Chermok, V. L., Hoyt, J. L., & Ortiz, B. G. (2007). An additional antecedent of empathic concern: valuing the welfare of the person in need. *Journal of Personality and Social Psychology*, *93*, 65–74.

Bellhouse, D. R. (2001). The Reverend Thomas Bayes FRS: A biography to celebrate the tercentenary of his birth (Unpublished manuscript). Retrieved from http://www2.isye.gatech.edu/~brani/isyebayes/bank/bayesbiog.pdf

Benson, P. L., Karabenick, S. A., & Lerner, R. M. (1976). Pretty pleases: The effects of physical attractiveness, race, and sex on receiving help. *Journal of Experimental Social Psychology*, *12*, 409–415.

Berkowitz, L., & Daniels, L. (1964). Affecting the salience of the social responsibility norm: Effect of past help on response to dependency relationships. *Journal of Abnormal and Social Psychology*, *67*, 275–281.

Berkowitz, L., & LePage, A. (1967). Weapons as aggression eliciting stimuli. *Journal of Personality and Social Psychology*, *7*(2), 202–207.

Biaggio, M. K. (1989). Sex differences in behavioral reactions to provocation of anger. *Psychological Reports*, *64*, 23–26.

Borenstein, M., & Cohen, J. (1988). *Statistical power analysis: A computer program*. Hillsdale, NJ: Lawrence Erlbaum.

Bourianoff, G. G., & Stubis, E. S. (1988). Stress management with headaches. In M. L. Russell (Ed.), *Stress management for chronic disease*. New York: Pergamon.

Box, J. F. (1978). *R. A. Fisher: The life of a scientist*. New York: Wiley.

Brady, J. V. (1958). Ulcers in executive monkeys. *Scientific American*, *199*, 95–100.

Brower, M., & Price, B. (2001). Neuropsychiatry of frontal lobe dysfunction in violent and criminal behavior: A critical review. *Journal of Neurology, Neurosurgery, and Psychiatry*, *71*(6), 720–726.

Burke, R. J., & Greenglass, E. R. (1989). It may be lonely at the top but it's less stressful: Psychological burnout in public schools. *Psychological Reports*, *64*, 615–623.

Buss, D. M. (1985). Human mate selection. *American Scientist*, *73*, 47–51.

Busseri, M. A., Choma, B. L., & Sandova, S. W. (2009). "As good as it gets" or "the best is yet to come"? How optimists and pessimists view their past, present, and anticipated future life satisfaction. *Personality and Individual Differences*, *47*, 352–356.

Buttery, T. J., & White, W. F. (1978). Student teachers' affective behavior and selected biorhythm patterns. *Perceptual and Motor Skills*, *46*, 1033–1034.

Byers, W. (2011). *The blind spot: Science and the crisis of uncertainty*. Princeton, NJ: Princeton University Press.

Campbell, D. T., & Stanley, J. (1963). Experimental and quasi-experimental designs for research on Teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.

Campbell, R. B. (2001). John Graunt, John Arbutnott, and the human sex ratio. *Human Biology*, *73*(4), 605–610.

Carey, K. (2016, December 29). A peek inside the strange world of fake academia. *The New York Times*, *A3*, Retrieved from http://www.nytimes.com/2016/12/29/upshot/fake-academe-looking-much-like-the-real-thing.html?_r=1.

Carlsmith, J. M., & Gross, A. E. (1969). Some effects of guilt on compliance. *Journal of Personality and Social Psychology*, *11*(3), 232–239.

Carmer, S. G., & Swanson, M. R. (1973). An evaluation of ten multiple comparisons procedures by Monte Carlo methods. *Journal of the American Statistical Association*, *68*, 66–74.

Carrie, C. M. (1981). Reproductive symptoms: Interrelations and determinants. *Psychology of Women Quarterly*, *6*(2), 174–186.

Carson, R. (1962). *Silent Spring*. Boston: Houghton Mifflin.

Caspi, A., Elder, G. H., & Bem, D. J. (1987). Moving against the world: Life course patterns of explosive children. *Developmental Psychology*, *23*, 308–313.

Chronicle of Higher Education (1990, March). Computer Notes, 36, p. 28.

Cochran, W. G. (1976). Early development of techniques in comparative experimentation. In D. B. Owen (Ed.), *On the history of statistics and probability*. New York: Marcel Dekker.

Cohen, B. (2013). *Explaining psychological statistics* (4th ed.). Hoboken, NJ: Wiley and Sons.

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

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312.

Combs, D. J. Y., Powell, C. A. J., Schurtz, D. R., & Smith, R. H. (2009). Politics, schadenfreude, and ingroup identification: the sometimes happy thing about a poor economy and death. *Journal of Experimental Social Psychology*, *45*, 635–646.

Confidence in Institutions: Trends in Americans' Attitudes toward Government, Media, and Business (2016). In The Associated Press–NORC Center for Public Affairs Research. Retrieved from http://www.apnorc.org/projects/Pages/HTML%20Reports/confidence-in-institutions-trends-in-americans-attitudes-toward-government-media-and-business0310-2333.aspx

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Corder, G. W., & Foreman, D. I. (2014). *Nonparametric statistics: A step-by-step approach* (2nd ed.). Hoboken, NJ: Wiley.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, *37*(5), 553–558.

Croasman, J. T., & Ostrom, L. (2011). Using Likert-type scales in the social sciences. *Journal of Adult Education*, *40*(1), 19–22.

Cronbach, L. J. (1967). The two disciplines of scientific psychology. In D. N. Jackson, & S. Messick (Eds.), *Problems in human assessment*. New York: McGraw Hill.

Cushny, A. R., & Peebles, A. R. (1904). The action of optical isomers. II. Hyocines. *Journal of Physiology*, *32*, 501–510.

Darley, J. M., & Batson, C. D. (1973). From Jerusalem to Jericho: a study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, *27*, 100–108.

Day, R. H., & Kasperczyk, R. T. (1984). The Morinaga misalignment effect with circular stimulus elements. *Bulletin of the Psychonomic Society*, *22*, 193–196.

De Moivre, A. (1738/1959). On the law of normal probability. In D. E. Smith (Ed.), *A source book in mathematics* (Vol. 2) (pp. 566–575). New York: Dover.

Delzell, D. A. P., & Poliak, C. D. (2013). Karl Pearson and eugenics: Personal opinions and scientific rigor. *Science and Engineering Ethics*, *19*, 1057–1070.

Devlin, H. (2013, 12 December). Academic benefits of music 'a myth'. *The Times of London*, Retrieved from http://www.thetimes.co.uk/tto/science/article3946476.ece.

Dion, K. K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, *24*, 285–290.

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in-and-out-of synchrony. *Cognitive Psychology*, *11*, 478–484.

Donnerstein, E. (1980). Aggressive-erotica and violence against women. *Journal of Personality and Social Psychology*, *39*, 269–277.

Draper, N. R., & Smith, H. (1966). *Applied regression analysis*. New York: Wiley.

Dunnington, G. W. (1955). *Carl Friedrich Gauss: Titan of science*. New York: The Free Press.

Dutton, D. G., & Aron, A. P. (1974). Some evidence of heightened sexual attraction under conditions of high anxiety. *Journal of Personality and Social Psychology*, *30*, 510–517.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but…: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, *110*, 109–128.

Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology*. (H. A. Ruger & C. E. Bussenius, Trans.). New York, NY: Columbia University.

Elliott, F. (1978). Neurological aspects of antisocial behavior. In W. H. Reid (Ed.), *The psychopath: A comprehensive study of antisocial disorders and behaviors* (pp. 146–189). New York: Bruner/Mazel.

Faculty History Project (2011). Retrieved from: http://um2017.org/faculty-history/faculty/rensis-likert/memorial

Feldman, B. G., & Paul, N. G. (1976). Identity of emotional triggers in epilepsy. *Journal of Nervous and Mental Disease*, *162*, 345–352.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fisher, R. A. (1939). Student. *Annals of Eugenics*, *9*, 1–9.

Fisher, R. A., & MacKenzie, W. A. (1923). Studies in crop variation. II: The manurial response of different potato varieties. *Journal of Agricultural Science*, *13*, 311–320.

Frank, M. L., & Lester, D. (1988). Geophysical variables and behavior: II. Temporal variation of suicide in teens and young adults. *Perceptual and Motor Skills*, *67*, 168–170.

Freedman, D. H. (2010, November). Lies, damned lies, and medical science. *The Atlantic*, *306*(11). https://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/?single_page=true.

Freud, S. (1922). *Beyond the pleasure principle*. (C. J. M. Hubback, Trans.). London, Vienna: International Psycho-Analytical.

Galton, F. (1869). *Hereditary genius: Its laws and consequences*. New York: Appleton.

Galton, F. (1885). The measure of fidget. *Nature*, *32*, 174–175.

Galton, F. (1908). *Memories of my life* (2nd ed.). London: Methuen.

Gerson, J., Plagnol, A., & Corr, P. J. (2016). Subjective well-being and social media use: Do personality traits moderate the impact of social comparison on Facebook? *Computers in Human Behavior*, *63*, 813–822.

Gino, F., & Ariely, D. (2012). The dark side of creativity: Original thinkers can be more dishonest. *Journal of Personality and Social Psychology*, *102*(3), 445–449.

Glass, G. V., & Stanley, J. C. (1970). *Statistical methods in education and psychology*. Englewood Cliffs: Prentice Hall.

Gliozzi, M. (2008). Cardono, Girolamo. *Complete dictionary of scientific biography*. Retrieved from http://www.encyclopedia.com/people/philosophy-and-religion/other-religious-beliefs-biographies/girolamo-cardano

Goranson, R. E., & Berkowitz, L. (1966). Reciprocity and responsibility reactions to prior help. *Journal of Personality and Social Psychology*, *3*, 227–232.

Gorenstein, E. E. (1982). Frontal lobe functions in psychopaths. *Journal of Abnormal Psychology*, *91*, 368–379.

Gosset, W. S. (1970). *Letters from W S. Gosset to R. A. Fisher, 1915–1936*. Issued for private circulation. Dublin: Arthur Guinness.

Grace-Martin, K. (2008). Can Likert Scale data ever be continuous? [The Analysis Factor]. Retrieved from: http://www.theanalysisfactor.com/can-likert-scale-data-ever-be-continuous/

Gravetter, F. J., & Wallnau, L. B. (2017). *Statistics for the behavioral sciences* (10th ed.). Canada: Cengage Learning.

Grimm, L., & Kanfer, F. H. (1976). The tolerance of aversive stimulation. *Behavior Therapy*, *7*, 593–601.

Grossarth-Maticek, R., Eysenck, H. J., & Vetter, H. (1988). Antismoking attitudes and general prejudice: An empirical study. *Perceptual and Motor Skills*, *66*, 927–931.

Grosskurth, P. (1980). *Havelock Ellis: A biography*. New York: Knopf.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics*, *29*, 205–220.

Hald, A. (2003). *A history of probability and statistics and their applications before 1750*. Hoboken, NJ: Wiley.

Hald, G. M., Malamuth, N. M., & Yuen, C. (2009). Pornography and attitudes supporting violence against women: revisiting the relationship in non-experimental studies. *Aggressive Behavior*, *36*, 14–20. doi:10.1002/ab.20328.

Hall, J. W. (1972). A comparison of Halpin and Croft's organizational climates and Likert and Likert's organizational systems. *Administrative Science Quarterly*, *17* (4), 586–590.

Halpin, B. (1978). Effects of arousal level on olfactory sensitivity. *Perceptual and Motor Skills*, *46*, 1095–1102.

Hare, R. D. (1984). Performance of psychopaths on cognitive tasks related to frontal lobe function. *Journal of Abnormal Psychology*, *93*(2), 133–140.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: Wiley.

Huesmann, L. R., & Eron, L. D. (1984). Cognitive processes and the persistence of aggressive behavior. *Aggressive Behavior*, *10*, 243–251.

Hupka, R. B., & Eshett, C. (1988). Cognitive organization of emotion: Differences between labels and descriptors of emotion in jealousy situations. *Perceptual and Motor Skills*, *66*, 935–949.

Ioannidis, J. P. A. (2005, August 1). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124 https://doi.org/10.1371/journal.pmed.0020124.

Isen, A. M. (1970). Success, failure, attention, and reaction to others: The warm glow of success. *Journal of Personality and Social Psychology*, *15*, 294–301.

Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: cookies and kindness. *Journal of Personality and Social Psychology*, *21*, 384–388.

Jackson, E. M., & Howton, A. (2008). Increasing walking college students using a pedometer intervention: Differences according to body mass index. *Journal of American College Health*, *57*(2), 159–164.

Jastrow, R. (1992). *God as the astronomers*. New York: W. W. Norton and Company, Inc.

Jones, I. (2015). *Research methods for sports studies* (3rd ed.). New York: Routledge.
Jourbert, C. E. (1989). Birth order and narcissism. *Psychological Reports*, *64*, 721–722.
Juon, H. S., Doherty, E. E., & Ensminger, M. E. (2006). Childhood behavior and adult criminality: Cluster analysis in prospective study of African Americans. *Journal of Quantitative Criminology*, *38*, 553–563.
Kaitz, M., Roken, A. M., & Eidelman, A. I. (1988). Infants' face-recognition by primiparous and multiparous women. *Perceptual and Motor Skills*, *67*, 495–502.
Kanfer, F. H., & Grimm, L. G. (1980). Managing clinical change: A process model of therapy. *Behavior Modification*, *4*, 419–444.
Katz, V. J. (1993). *A history of mathematics*. HarperCollins College Publishers.
Kaufman, A. S. (1979). *Intelligent testing with the WISC-R*. New York: Wiley.
Keppel, G. (1982). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice-Hall.
Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs: Analysis of variance and multiple regression/correlation approaches*. New York: Freeman.
Kiess, H. O. (1989). *Statistical concepts for the behavioral sciences*. Boston: Allyn and Bacon.
Kirk, R. E. (Ed.) (1972). *Statistical issues: A reader for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole.
Kirk, R. E. (1989). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Belmont, CA: Brooks/Cole.
Kirschner, P. A., & Karpinski, A. C. (2010). Facebook and academic performance. *Computers in Human Behavior*, *26*, 1237–1245.
Kish, L. (1990). A CHOICES Profile: Rensis Likert: Social scientist and entrepreneur. *Choices*, *5*(4), 36–38.
Kraus, N., Hornickel, J., Strait, D., Slater, J., & Thompson, E. (2014). Engagement in community music classes sparks neuroplasticity and language development in children from disadvantaged backgrounds. *Frontiers in Psychology*. Retrieved from journal.frontiersin.org. doi:10.3389/fpsyg.2014.01403/full.
Krauss, I. K. (1980). Between- and within-group comparisons in aging research. In L. W. Poon (Ed.), *Aging in the 1980's psychological issues* (pp. 542–551). Washington, DC: American Psychological Association.
Kühl, S. (1994). *The Nazi connection: Eugenics, American Racism, and German National Socialism*. New York: Oxford University Press.
Landauer, T. K., & Whiting, J. W. M. (1964). Infantile stimulation and adult stature of human males. *American Anthropologist*, *66*, 1007–1028.
Lashley, K. S. (1915). The acquisition of skill in archery. *Carnegie Institutions Publications*, *211*, 107–128.
Latane, B., & Darley, J. (1970). *The unresponsive bystander: Why doesn't he help*. New York: Appleton-Century-Crofts.

Leek, J. T., & Jager, L. R. (2017). Is most published research really false? *Annual Review of Statistics and its Applications*, *4*, 109–122.

Lehman, M., Moelats, D., Varewyck, M., Stynes, F., van Noorden, L., & Martens, J.-P. (2013). Activating and relaxing music entrains the speed of beat synchronized walking. *PLoS One*, *7*, e67932 https://doi.org/10.1371/journal.pone.0067932.

Lemay, E. P. Jr., Clark, M. S., & Greenberg, A. (2010). What is beautiful is good because what is beautiful is desired: Physical attractiveness stereotyping as projection of interpersonal goals. *Personality and Social Psychology Bulletin*, *36*, 339–353.

Leonhardt, D. (2000, July 28). John Tukey, 85, Statistician; Coined the Word 'Software'. Retrieved from http://www.nytimes.com/2000/07/28/us/john-tukey-85-statistician-coined-the-word-software.html

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants employ selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*, *109*(5), 1431–1436.

Liberman, B. (1971). *Contemporary problems in statistics: A book of readings for the behavioral sciences*. Oxford: London.

Lifton, R. J. (2000). *The Nazi Doctors: Medical killing and the psychology of genocide*. New York: Basic Books.

Lombardi, C. M., & Hurlbert, S. H. (2009). Misprescription and misuse of one-tailed tests. *Austral Ecology*, *34*, 447–468.

Loschelder, D. D., Swaab, R. I., Trötschel, R., & Galinsky, A. D. (2014). The first-move disadvantage. *Psychological Science*, *25*. http://journals.sagepub.com/doi/pdf/10.1177/ 0956797613520168

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542.

Malamuth, N. M., Heim, M., & Feshback, S. (1980). Sexual responsiveness of college students to rape depictions: Inhibitory and disinhibitory effects. *Journal of Personality and Social Psychology*, *38*, 399–408.

Mammarella, N., Russo, R., & Avons, S. E. (2002). Spacing effects in cued-memory tasks for unfamiliar faces and nonwords. *Memory & Cognition*, *30*, 1238–1251.

Marczyk, G., DeMatteo, D., & Festinger, D. (2005). *Essentials of research design and methodology*. Hoboken, NJ: Wiley and Sons.

McCullagh, P. (2003). John Wilder Tukey. *Biological Memoirs of Fellows of the Royal Society*, *49*, 537–555.

McGrath, A. (2015). *The big question: Why we can't stop talking about science, faith, and god*. New York: St. Martin's Press.

Mehr, S. A., Schachner, A., Katz, R. C., & Spelke, E. S. (2013). Two randomized trials provide no consistent evidence for nonmusical cognitive benefits of brief preschool music enrichment. *PLoS One*, *8*, e82007. doi:10.1371/journal.pone.0082007.

Miller, L. (1990). *Relations among cognitions and behaviors of aggressive children and their mothers* (Doctoral dissertation). University of Illinois at Chicago.
Mirels, H. L., & Dean, J. B. (2006). Right-wing authoritarianism, attitude salience, and beliefs about matters of fact. *Political Psychology*, *27*(6), 839–866.
Moore, D. A., & Tenney, E. R. (2012). Time pressure, performance, and productivity. In M. A. Neale, & E. A. Mannix (Eds.), *Looking back, moving forward: A review of group and team-based research (research on managing groups and teams, volume 15)* (pp. 305–326)Emerald Group Publishing Limited.
Nichols, A. L., & Maner, J. K. (2008). The good subject effect: Investigating participant demand characteristics. *Journal of General Psychology*, *135*, 151–165.
Obrist, P. A. (1962). Some autonomic correlates of serial learning. *Journal of Verbal Learning and Verbal Behavior*, *1*, 100–104.
Oishi, S., & Schimmack, U. (2010). Residential mobility, well-being, and mortality. *Journal of Personality and Social Psychology*, *98*(6), 980–994.
Olweus, D. (1979). The stability of aggressive reaction patterns in human males: A review. *Psychological Bulletin*, *85*, 852–875.
Pearson, K. (1896). Mathematical contributions to the theory of evolutionIII. Regression, heredity and panmixia. *Philosophical Translations of the Royal Society of London, Series A*, *187*, 253–318.
Pearson, K. (1903). On the inheritance of the mental and moral characters in man, and its comparison with the inheritance of the physical characters. *Journal of the Anthropological Institute*, *33*, 179–237.
Pew Research Center (n.d.). Questionnaire design. Retrieved from http://www.pewresearch.org/ methodology/u-s-survey-research/questionnaire-design/
Pfungst, O. (1911). *Clever Hans (The horse of Mr. Von Osten): A contribution to experimental animal and human psychology*. New York: Holt and Co.
Pham, M. T., Hung, I. W., & Gorn, G. J. (2011). Relaxation increases monetary valuations. *Journal of Marketing Research*, *48*(5), 814–826.
Plous, S. (1993). *The psychology of judgment and decision making*. New York: McGraw-Hill.
Porter, R. W., Brady, J. V., Conrad, D., Mason. J. W., Galambos, R., & McKrioch, D. (1958). Some experimental observations on gastrointestinal lesions in behaviorally conditioned monkeys. *Psychosomatic Medicine*, *20*, 379–394.
Puskarich, C. A. (1988). The effects of progressive muscle relaxation on seizure frequency in adults with epileptic seizures (Doctoral dissertation). University of Illinois at Chicago.
Reeves, J. (2015). The secularization of chance: Toward understanding the impact of the probability revolution on Christian belief in divine providence. *Zygon*, *50*(3), 604–620.
Riddle, O. (1947). *Biographical memoir of Charles Benedict Davenport 1866–1944*. National Academy of Sciences, 25.
Robbins, R. A. (1988). Objective and subjective factors in estimating life expectancy (Unpublished manuscript). The Pennsylvania State University at Harrisburg, Division of Behavioral Science and Education.

Rohwedder, S., & Willis, R. J. (2010). Mental retirement. *Journal of Economic Perspectives*, *24*, 119–138. doi:10.1257/jep.24.1.119.

Romano, S. T., & Bordieri, J. E. (1989). Physical attractiveness stereotypes and students' perceptions of college professors. *Psychological Reports*, *4*, 1099–1102.

Rosenhan, D. L., & White, G. M. (1967). Observation and rehearsal as determinants of prosocial behavior. *Journal of Personality and Social Psychology*, *5*, 424–431.

Rosenthal, R., & Fode, K. (1963). The effect of experimenter bias on performance of the albino rat. *Behavioral Science*, *8*, 183–189.

Rowatt, W. C., Powers, C., Targhetta, V., Comer, J., Kennedy, S., & Labouff, J. (2006). Development and initial validation of an implicit measure of humility relative to arrogance. *The Journal of Positive Psychology*, *1*(4), 198–211.

Rubin, M., Paolini, S., & Crisp, R. (2010). A processing fluency explanation of bias against migrants. *Journal of Experimental Social Psychology*, *46*(1), 21–28.

Russell, B. (1936/1997). *Religion and science*. Oxford: Oxford University Press.

Ruth, W. J., Mosatche, H. S., & Kramer, A. (1989). Freudian sexual symbolism: Theoretical considerations and an empirical test in advertising. *Psychological Reports*, *64*, 1131–1139.

Ryan, T. P. (2008). *Modern regression methods* (2nd ed.). Hoboken, NJ: Wiley and Sons.

Sande, G. (2001, July). John Wilder Tukey. *Physics Today*, *54*(7), 80 http://physicstoday.scitation.org/doi/10.1063/1.1397408?journalCode=pto.

Schalling, D. (1978). Psychopathy-related personality variables and the psychophysiology of socialization. In R. D. Hare, & D. Schalling (Eds.), *Psychopathic behavior: Approaches to research* (pp. 85–106). Chichester, England: Wiley.

Sedlmeier, P., & Gigerenzer, G. (1989). Do statistical studies of power have an effect on the power of studies. *Psychological Bulletin*, *105*(2), 309–316.

Seery, M. D., Holman, E. A., & Silver, R. C. (2010). Whatever does not kill us: Cumulative lifetime adversity, vulnerability, and resilience. *Journal of Personality and Social Psychology*, *99*, 1025–1041.

Segal-Caspi, L., Roccas, S., & Sagiv, L. (2012). Don't just a book by its cover, revisited: Perceived and reported traits and values of attractive women. *Psychological Science*, *23*, 1112–1116.

Semmel, B. (1958). Karl Pearson: Socialist and Darwinist. *British Journal of Sociology*, *9*(2), 111–125.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Boston: Houghton Mifflin Co.

Shavelson, R. J. (1988). *Statistical reasoning for the behavioral sciences* (2nd ed.). Boston: Allyn and Bacon.

Shelton, T. O., & Mahoney, M. J. (1978). The content and effect of "psyching-up" strategies in weight lifters. *Cognitive Therapy and Research*, *2*, 275–284.

Siegal, S., & Castellan, N. J. (1988). *Nonpammetric statistics for the behavioral sciences* (2nd ed.). New York: McGraw Hill.

Simpson, T. (1755). A letter to the right honourable George Earl of Macclesfield, President of the Royal Society, on the advantage of taking the mean of a number of observations, in practical astronomy. *Philosophical Translations of the Royal Society of London*, *46*, 82–93.

Simpson-Housley, P., & DeMan, A. (1989). Flood experience and posttraumatic trait anxiety in Appalachia. *Psychological Reports*, *64*, 896–898.

Sines, J. O. (1959). Selective breeding for development of stomach lesions following stress in the rat. *Journal of Comparative and Physiological Psychology*, *52*, 615–617.

Sines, J. O., Cleeland, C., & Adkins, J. (1963). The behavior of normal and stomach lesion susceptible rats in several learning situations. *Journal of Genetic Psychology*, *102*, 91–94.

Slaby, R. G., & Roedell, W. C. (1982). The development and regulation of aggression in young children. In J. Wordell (Ed.), *Psychological development in the elementary years*. New York: Academic Press.

Standage, K. (1972). Treatment of epilepsy by the reciprocal inhibition of anxiety. *Guy's Hospital Reports*, *121*, 217–221.

Standage, K. F., & Fenton, G. W. (1975). Psychiatric symptom profile of patients with epilepsy: a controlled investigation. *Psychological Medicine*, *5*, 152–160.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Stouffer, S. A. (1958). Karl PearsonAn appreciation on the l00th Anniversary of his birth. *Journal of the American Statistical Association*, *53*, 23–27.

Student (1908). The probable error of a mean. *Biometrika*, *6*, 1–25.

Symonds, C. (1970). *Some observations on the facilitation or arrest of epileptic seizures*. Oxford: London.

Tankard, J. W. (1984). *The statistical pioneers*. Cambridge, MA: Schenkman.

Temkin, N. R., & Davis, G. R. (1984). Stress as a risk factor for seizures among adults with epilepsy. *Epilepsia*, *25*, 450–456.

Tucker, J. (2016, March 9). Does social science have a replication crisis? *Washington Post*., Retrieved from https://www.washingtonpost.com/news/monkey-cage/wp/2016/03/09/does-social-science-have-a-replication-crisis/?utm_term=.1c5d71247aa0.

UNESCO (1952). The race concept: Results of an Inquiry, p. 27.

Valins, S., & Ray, A. A. (1967). Effects of cognitive desensitization on avoidance behavior. *Journal of Personality and Social Psychology*, *7*, 345–350.

Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes's Theorem and the additivity principle. *Memory & Cognition*, *30*(2), 171–178.

Wainer, H., & Thissen, D. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*, 213–217.

Walker, H. M. (1934). Abraham De Moivre. *Scripta Mathematica*, *2*, 316–333.
Walker, H. M. (1940). Degrees of freedom. *Journal of Educational Psychology*, *55*, 253–269.
Walker, H. M. (1968). Pearson, K. In D. L. Sills (Ed.), *International encyclopedia of the social sciences* (Vol. II) (pp. 496–503). New York: Macmillan and The Free Press.
Weisberg, H. I. (2014). *Willful ignorance: The mismeasure of uncertainty*. Hoboken, NJ: Wiley and Sons.
Welkowitz, J., Ewen, R. B., & Cohen, J. (1988). *Introductory statistics for the behavioral sciences*. New York: Harcourt Brace Jovanovich.
Wiggins, J. (1973). *Personality and prediction: Principles of personality assessment*. Menlo Park: Addison-Wesley.
Wilson, W. A. (2016, May). Scientific regress. *First Things*. Retrieved from https://www.firstthings.com/ article/2016/05/scientific-regress
Woods, A. M., & Rusin, M. J. (1988). Stress management and the elderly. In M. L. Russell (Ed.), *Stress management for chronic disease* (pp. 49–62). New York: Pergamon.
Yarhouse, M. A., Dean, J. B., Stratton, S. P., & Lastoria, M. (2018). *Listening to sexual minorities: A* study of faith and sexual identity on Christian College Campuses. Downers Grove, Illinois: Intervarsity Press.
Yule, G. U. (1895). On the correlation of total pauperism with proportion of out-relief. I. All ages. *Economic Journal*, *5*, 603–611.
Zakahi, W. R., & Duran, R. L. (1988). Physical attractiveness as a contributing factor to loneliness: An exploratory study. *Psychological Reports*, *63*, 747–751.
Zullow, H. M. (1984). The interaction of rumination and explanatory style in depression (Master's thesis). University of Pennsylvania.
Zullow, H. M., & Seligman, M. E. P. (1990). Pessimistic rumination predicts defeat of presidential candidates. *Psychological Inquiry*, *1*(1), 52–61.

Glossary

Abscissa The horizontal (X) axis of a graph.

Addition rule Used to determine the probability of occurrence of one or more of many possible events.

Alpha inflation The result of conducting multiple t tests such that the probability of a type I error increases with the number of t tests.

Alpha level The value set by the researcher that specifies the probability of making type I error (rejecting a true null hypothesis).

Alternative hypothesis The opposite of the null hypothesis. The hypothesis that is automatically accepted when the null hypothesis is rejected. Accepting the alternative hypothesis means that the results of a study are probably not due to chance; there is probably an effect, difference, or relationship between variables.

Analysis of variance (ANOVA) A statistical test designed to determine if there are population differences among several sample means.

A posteriori approach Determining the probability of an event empirically by gathering data and then dividing the number of times A has occurred by the total number of data points gathered.

A priori approach See Classical approach to probability.

A priori tests A category of tests, selected prior to the rejection of the null hypothesis in an overall analysis, that are used to locate differences between pairs of means.

Asymptotic When a line on a graph continually approaches but never reaches the X axis.

Balancing A means of controlling an extraneous variable by representing it equally across all conditions.

Bar graph A frequency distribution for categorical (nominal) data.

Bayes' theorem The formula developed by Thomas Bayes that allows one to find the probability of B given A if one knows the probability of A given B as well as the probability of B and the probability of A given not B.

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

Between-group variation A measure of the variation among group means in a study with two or more groups.

Between-participants design See *Independent-samples design.*

Biased sample A nonrandomly selected sample such that each member of a population does *not* have an equal chance of being selected.

Bimodal distribution A distribution with two modes.

Bivariate distribution A distribution of two variables in which scores are paired. A correlation is based on a bivariate distribution.

Categorical data Nominal data arranged by categories. For example, marital status: single, married, divorced, widowed. See *Nominal scale.*

Cell The section of a design identifying a single group amidst multiple factors. For instance, a 2 × 2 factorial design has four cells: $\text{Cell}_{1,1}$; $\text{Cell}_{1,2}$; $\text{Cell}_{2,1}$; and $\text{Cell}_{2,2}$.

Central limit theorem The mathematical theorem that states that the sampling distribution of means approaches a normal curve as the sample size increases. The mean of the sampling distribution is equal to the mean of the population of raw scores. The standard deviation of the sampling distribution is equal to $\sigma/\sqrt{n}$.

Central tendency A statistic that indicates wherein the distribution scores tend to bunch. The mean, median, and mode are common measures of central tendency.

Chi-square distribution A theoretical sampling distribution of chi-square values. The shape of the chi-square distribution changes as a function of the degrees of freedom.

Chi-square test One of two significance tests (goodness of fit; test for independence) using frequency count data.

Class interval Groups of equal-sized ranges as determined by the researcher based on how much information loss one is willing to sacrifice in exchange for tabular simplicity.

Classical approach (or a priori approach) Logically determining the probability of an event by dividing the number of ways the event can occur by the total number of possible outcomes.

Coefficient of determination A squared correlation coefficient. It measures the amount of variation of Y scores accounted for by the variation of X scores. A measure of common or shared variance.

Cohen's *d* One of the simplest, direct, and most often used measures of effect size.

Computational (raw-score) formulas A formula that uses the raw scores of the distribution. Although a computational formula obscures the conceptual basis of the formula, it is used for ease in hand calculations.

Conditional distribution In regression, the spread of Y scores for a given X score.

Conditional probability An expression of likelihood given that another particular event has occurred.

Confidence interval A range of values within which a researcher can state with a certain degree of confidence that a population parameter will fall.

Confidence limits The upper and lower values of a confidence interval.

Confounding variable An uncontrolled extraneous variable that systematically varies with an independent variable. A confounding variable can offer a plausible alternative explanation for the results of the study.

Contingency (frequency or cross-tabulation) table A table that categorizes observations as frequency counts along a two-factor grid.

Continuous variable A variable that theoretically has an infinite number of points between any two numbers.

Control group In experimental research, a condition marked by the absence of the independent variable.

Correlated (dependent) samples A research design in which the scores of one experimental condition are not independent of scores of another experimental condition, also called a *dependent-samples design or repeated-measures design*. Another example of a correlated samples design is a matched-participants design in which the selection of one sample determines who will be selected for the other sample(s).

Correlated-samples *t* test See *Dependent-samples t test*.

Correlation (or prediction or association) A description of the degree to which two or more variables relate to one another. This is one of the goals of the researcher.

Correlation coefficient A measure of the degree of association between two variables. A correlation coefficient can range from −1 to +1. The higher the absolute value of the correlation coefficient, the stronger the relationship between the two variables.

Correlational designs Studies that do not control and manipulate variables. Correlational research examines the covariation among variables.

Counterbalancing A strategy used with repeated-measures designs in which participants differ by the order in which experimental conditions are presented. The purpose of counterbalancing is to prevent confounding of the independent variable with order effects by distributing carryover effects that come with repeated measuring across all experimental conditions.

Cramér's *V* A measure of association for nominal variables. In this text, it is used as a measure of effect size for both the chi-square goodness-of-fit test and the chi-square test for independence.

Critical region The area under a curve that has the values that lead to rejection of the null hypothesis.

Critical values The values that separate the rejection regions from those regions of the null distribution that would not lead to a rejection of the null hypothesis.

Cumulative frequency distribution A frequency distribution, which includes a column that indicates the total frequency of scores up to and including a given class interval.

Curvilinear relationship A nonlinear relationship between X and Y. For example, when lower scores on X are associated with lower scores on Y, medium X scores are associated with medium Y scores, but higher X scores are associated with lower Y scores.

Definitional formulas Formulas that emphasize the conceptual basis of the statistic. Although it may be much more involving if used for hand calculations, it clearly reflects the essence of the statistic.

Degrees of freedom The number of values that are free to vary with certain restrictions placed on all values.

Dependent-samples *t* test A significance test used with dependent sampling in which scores across conditions are paired. See *Repeated-measures design*.

Dependent variable A measured variable that is expected to be a consequence of the independent variable. In experimental research, the dependent variable is causally determined, in part, by the independent variable. In correlational research in which a regression equation is used to predict the value of one variable given the value of another variable, the dependent variable is also called the *predicted variable* or *criterion variable*.

Description Defining, identifying, classifying, categorizing, and organizing a topic of interest. This is the initial goal of the researcher.

Descriptive statistics Statistical techniques designed to describe and summarize data in an abbreviated form.

Deviation score (x) (or error score) The difference between a score and the mean of the distribution.

Dichotomous variable A variable that takes only two contrary values (e.g. depressed = 0, not depressed = 1).

Discontinuous (or discrete) variable A variable that can take on only a finite number of values and for which there are no meaningful intermediate values.

Distribution A list of scores arranged in order of magnitude.

Distribution-free tests In inferential statistics, tests that do not make assumptions about characteristics of the population distribution.

Dummy coding Assigning arbitrary numbers to two designate groups of observations (e.g. citizen = 0 and alien = 1).

Error (or secondary) variance A measure of the variation between group means or among scores within a group due to uncontrolled, random factors (including individual differences between participants) in the experiment or study.

Eta-squared (η^2) A statistic that estimates the size of an effect; a ratio of primary variance over total variance.

Expected frequencies The number of observations expected to occur when the null hypothesis is true.

Experiment A reserved term for a particular research design involving (i) a high degree of control over the independent variable, (ii) careful measurement of the dependent variable, and (iii) complete control of all other variables.

Experimental error A measure of variance either between group means or within groups that is due to unsystematic influences on individual scores. The contributors to experimental error can include unreliable measures, inconsistent administration, and disturbances in the research environment.

Experimental group A group marked by the presence of an independent variable; distinguished from a "control group," which is marked by the absence of an independent variable.

External validity The extent to which experimental findings can be generalized to different populations, settings, treatment variables, and measurement variables.

Extraneous variable Any variable found in an experimental context that is not either being manipulated or carefully measured.

Factor A variable that has at least two conditions associated with it; in an experimental context, an "independent variable." A two-way ANOVA has two factors or independent variables.

Factorial ANOVA The factorial ANOVA analyzes data from a research study that has two or more independent variables.

Factorial (or complex) design A research design that has two or more factors (or "independent variables" if used in an experimental context).

***F* distribution** A theoretical sampling distribution of *F* values. The shape of the *F* distribution is a function of the degrees of freedom. *F* distributions are positively skewed, with most *F* values bunched around 1.

Fisher's *LSD* (Least Significant Difference) A multiple comparison test used to locate the source of significance following a significant *F* test. Also called a *protected t test*.

Frequency The number of times each score occurs in a distribution.

Frequency count Data in the form of "how many" rather than "how much." A chi-square analysis uses frequency counts.

Frequency polygon A graph with measured scores on the *X* axis and frequencies on the *Y* axis. Each point above an *X* score represents the frequency of occurrence of that score.

Goodness-of-fit test A chi-square test that examines the degree to which observed data from a set of categories arranged along a single factor coincide with theoretical or previously empirical-derived expectations.

Grand mean The mean of all the scores in an experiment.

Grouped frequency distribution A table possessing equal-sized class intervals with an adjacent column noting the frequency of occurrence for values corresponding to each class interval.

Heteroscedasticity In regression analysis, a violation of homoscedasticity wherein the variances of the conditional distributions are unequal.

Higher order A relative judgment of types of effects. In an analysis involving multiple effect types, the effect involving the most number of factors or independent variables is the highest order effect and so on.

Histogram A graph of vertical bars with shared borders in which the height of each bar corresponds to the frequency of scores for a given class interval.

Holding constant A means of controlling an extraneous variable by treating it as a constant.

Homogeneity of variance Variances of equivalent values in two or more populations. Parametric significance tests assume this equivalence.

Homoscedasticity An assumption in regression analysis in which all conditional distributions have equal variances.

Hypothesis A prediction, emerging from a theory, about what data is expected to be found given a particular situation.

Hypothesis testing A method for testing claims made about population parameters. Hypothesis tests are also called *significance tests*.

Independent events Events that are unrelated. The occurrence of one event does not affect the occurrence of another event.

Independent observations The methodological assumption that each score within a *sample* is independent of all other scores.

Independent-samples design A research design in which the scores of one condition are unrelated to (independent of) the scores in any other condition.

Independent-samples *t* test A significance test used to compare the sample means of two samples. See *Independent-samples design*.

Independent variable The manipulated variable believed to be the "cause" part of a cause–effect relationship. In regression, the predictor variable.

Individual differences A measure of variance between group means or between individual scores within a group that is due to the unique skills, abilities, and tendencies of individual participants.

Inferential statistics Statistical techniques using sample data that allow researchers to make inferences about the characteristics of the population from which the sample came.

Interaction In a factorial design, when the effect of one factor is altered depending on the value of a second factor. This is new and unique variance not explained by main effects.

Internal validity The degree to which an experiment can allow the investigator to make a cause-and-effect statement of the relationship between the independent and dependent variable. The presence of confounding variables decreases the internal validity of an experiment.

Interquartile range (IQR) The difference between the first and third quartiles.

Interval estimate See *Confidence interval*.

Interval scale A scale of measurement in which the quantitative distance between intervals is held constant across the breadth of the scale. The zero point, however, is arbitrary.

Kurtosis The peakedness or flatness of a distribution curve.

Least squares criterion The criterion used in the least squares method to establish a regression line. See *Least squares method*.

Least squares method The method of fitting a regression line to a scatter plot such that the sum of the squared errors is at a minimum.

Leptokurtic A distribution curve that is relatively narrow and possessing an accentuated peak.

Level of significance (alpha level) The probability value (e.g. .05) at which the null hypothesis is rejected.

Linear relationship A relationship between two variables in which as the value of one variable changes, the value of a second variable changes by a constant.

Main effect An effect found among the conditions of one factor, independent of the influence of another factor.

Manipulation The controlled presentation of an independent variable.

Mann–Whitney *U* test A nonparametric hypothesis test used to compare two-independent samples. It is the nonparametric counterpart to the independent-samples *t* test.

Matched-participants design See *Matched-samples design*.

Matched-samples design A research design in which participants are paired on a theoretically important participant variable and then randomly split and assigned to different groups. This technique is designed to control potent participant variables (e.g. IQ, age, ethnicity, etc.).

Mean The sum of scores divided by the number of scores.

Mean deviation A measure of dispersion or variability in the distribution of scores. The mean deviation is the sum of the absolute values of deviation scores divided by the number of scores.

Mean square (MS) In ANOVA, a sum of squares divided by its degree of freedom.

Mean square between (MS_{BG}) A weighted measure of variation between group means.

Mean square within (MS_W) A weighted average of within-group variances of two or more samples, also called the pooled variance.

Measurement The assignment of numbers to attributes, objects, or events according to a set of predetermined rules.

Measures of central tendency See *Central tendency*.

Measures of variability (or dispersion) Numerical measures that reflect the degree to which scores of a distribution are spread out.

Median The midpoint of a distribution in which 50% of the scores fall below the midpoint.

Mesokurtic A curve that has a degree of peakedness that is intermediate between leptokurtic and platykurtic curves.

Midpoint The balance point of an interval.

Mode The score in a distribution that occurs most frequently.

Multiple comparisons Tests of differences between means performed after an ANOVA. Multiple comparisons are performed to locate the source of significance found by an *F* test.

Multiple regression A statistical technique that uses two or more variables to predict a criterion variable.

Multiple regression equation A regression equation in which more than one predictor variable is used to predict a criterion variable.

Multiplication rule A rule used to determine the probability of the joint occurrence of two or more events.

Mutually exclusive events When the occurrence of one event precludes the occurrence of another event.

Negatively skewed distribution A skewed distribution in which the elongated tail points toward the smaller or negative numbers.

Nominal scale A measurement scale that conveys no quantitative information, but rather merely distinguishes one attribute, object, or event from others.

Nonparametric tests Statistical tests that do not make assumptions about population parameters and do not require interval or ratio data.

Normal distribution (normal curve) A symmetrical bell-shaped curved line escalating gradually at first and then more aggressively, inflecting at some point and then tapering to a peak.

Normality The statistical assumption that the population from which the sample is taken is normally distributed.

Null hypothesis Symbolized as H_0, the null hypothesis is a statement about some population characteristic. It usually states no effect, no difference, or no relationship between variables.

Observation Making careful and systematic measurements of events occurring in the world by using either one of the five senses or scientific tools and instruments.

Observed frequencies The actual frequency counts recorded in a set of categories.

Omega-squared (ω^2) A statistic that estimates the size of an effect; an adjusted ratio of primary variance over total variance.

One-tailed (or directional) test A statistical test in which the rejection region lies only in one tail of the sampling distribution.

One-way ANOVA A statistical test (*F* test) used to compare several sample means to determine if one or more have come from populations with different means. Used only with research designs with one independent variable or (if nonexperimental) one factor.

Operational definition A description of the concrete measurement of a concept for the purposes of a given research project.

Order effects Differences between treatment conditions due to the order of presentation in a repeated-measures design. Order effects should be removed by altering the order of treatments among participants.

Ordinal scale A scale of measurement registering the relative position between attributes, objects, or events.

Ordinate The vertical (Y) axis of a graph.

Paired-samples *t* test See *Dependent-samples t test.*

Parameter A numerical population value. Parameters are usually inferred from sample statistics.

Parametric tests Significance tests that require interval or ratio scaled data and that make assumptions about the characteristics of populations.

Participant variable A characteristic of a participant that is fixed at the time of the experiment.

Pearson product-moment correlation coefficient A measure of association between two variables created by Karl Pearson. It is a very powerful and frequently used measure of association.

Percentile The value in a distribution below which falls a certain percentage of scores.

Percentile rank A number assigned to a score that indicates the percentage of scores found below that score.

Platykurtic A distribution curve that is relatively broad and possessing a muted peak.

Planned comparisons See *A priori tests.*

Point-biserial correlation A correlational analysis used when one variable is continuous and the other variable is dichotomous, symbolized as r_{pb}.

Point estimation Using a sample statistic to infer the value of a population parameter.

Pooled variance A weighted average of variances from two samples. See *Mean square within.*

Population Every member of a given group.

Positively skewed distribution A skewed distribution in which the elongated tail points toward the larger positive numbers.

Post hoc (or posteriori) tests A category of tests, usually selected prior to the rejection of the null hypothesis in an overall analysis, which are then used to locate differences between pairs of means.

Power The probability of an inferential test to reject correctly a null hypothesis.

Probabilistic dependence When knowledge of the occurrence of one event changes the probability of occurrence of a second event.

Probabilistic independence When knowledge of the occurrence of one event has no effect on determining the probability of occurrence of a second event.

Probability A measure of the likelihood that an event will occur. Probability values range from 0 to 1.

Protected *t* test See *Fisher's LSD.*

Qualitative independent variable An independent variable that changes by kind or type. For example, different kinds of drugs.

Quantitative independent variable An independent variable that changes by an amount. For example, different doses of a drug.

Quartile One-fourth of a distribution of scores.

Quasi-experiment A research design that has many of the same characteristics of an experiment, but one in which participants are not randomly assigned to conditions.

Random assignment The placing of participants into study conditions such that each participant is equally likely to be assigned to a given condition.

Random factors Unsystematic sources of variation. Individual differences and experimental error are random factors. These factors generate error variance, also called secondary variance.

Randomization The assignment of participants to treatment conditions so that each participant is just as likely to be assigned to one or another condition. Randomization is intended to spread participant variables equally across treatment conditions to eliminate participant variables as confounds.

Random sample A sample of scores taken from a population in such a way that each score of the population has an equal chance of being included in the sample.

Random sampling A means of selecting participants (or observations) for a study in such a way that each participant (or observations) in the population has an equal chance of being included in the sample.

Range A measure of dispersion in a distribution. The highest score of a distribution minus the lowest score.

Ratio scale A scale of measurement that has all the characteristics of an interval scale plus a "true" zero point.

Raw score A quantitative score obtained in a study. Also called an *original score*.

Real limits The upper and lower boundaries of an interval of measurement. The upper real limit of the number is one-half the unit of measurement above the number; the lower real limit of a number is one-half the unit of measurement below the number.

Regression A set of statistical procedures applied to correlated bivariate data that allow a researcher to use the value of one variable to predict the value of another variable.

Regression equation An equation used to predict a Y value given a specific X value.

Regression line (linear) A mathematically generated straight line fitted to a scatter plot by the least squares criterion. The regression line includes all predicted values of Y for all X scores.

Related-samples *t* test See *Dependent-samples t test*.

Repeated-measures design An experimental design (usually) in which participants are exposed to all conditions of the study.

Representativeness The methodological assumption that the participants comprising the sample represent the population in question.

Research hypothesis A statement (prediction) about the expected outcome of a study; usually derived from theory or previous research findings.

Robust statistic A statistic (e.g. *t* test) that is resistant to violations of a particular assumption (usually the assumption of normality).

Sample A subset of participants (or observations) drawn from a population.

Sample space Specification of the probabilistic situation.

Sampling The process of selecting participants (or observations) into a study.

Sampling distributions Theoretical distributions of a statistic based on all possible random samples of size n taken from the same population.

Sampling error The difference between a sample statistic and a population parameter (e.g. $M - \mu$).

Scatter plot A graphic representation of a bivariate distribution.

Scientific hypothesis See *Research hypothesis*.

Semi-interquartile range (SIQR) The interquartile range divided by 2.

Shared variance The amount of variance in one variable that can be explained or accounted for by variance in a second variable.

Simple frequency distribution Scores arranged from highest to lowest, with the frequency of occurrence of each score indicated in a column beside the scores.

Skewed distribution An asymmetrical distribution in which scores tend to bunch at either the left or the right end of the curve. Skewed distributions may be either positive or negative.

Slope The angle of a straight line. A descriptive feature of a regression line.

Spearman rank correlation A correlation coefficient for two sets of ranked data, symbolized as r_s.

Standard deviation A measure of dispersion or variability of a distribution of scores. The standard deviation is the square root of the variance. In a normal distribution the mean plus and minus one standard deviation marks approximately the middle 68% of the scores.

Standard error of the difference The standard deviation of a sampling distribution of differences between means.

Standard error of the estimate In regression, a measure of prediction error. The average standard deviation of the conditional distributions.

Standard error of the mean The standard deviation of a sampling distribution of means.

Standard normal curve A graph of the standard normal distribution.

Standard normal distribution A distribution of z scores with a mean of 0 and a standard deviation of 1. Derived from a raw score distribution that is normally distributed.

Standard score A raw score expressed in standard deviation units. A z score is an example of a standard score.

Statistic A numerical value of a sample.

Statistical hypothesis A numerical statement of the potential outcome of a study. Statistical hypotheses usually come in mutually exclusive and collectively exhaustive pairs: the null and alternative hypothesis.

Statistically significant A conclusion that a test statistic is unlikely to have occurred by chance. In a well-controlled experiment, a statistically significant finding indicates that the independent variable has had an effect on the dependent variable.

Sum of squares (*SS*) The sum of the squared deviations from the mean. This concept is a component of numerous statistical formulas.

***t* Distribution** A family, based on different sample sizes, of various bell-shaped distributions of *t* values. Also called the *Student t distribution*. The *t* distribution has a mean of 0.

***t* Test** Any number of inferential tests based on the *t* distribution.

Test for independence A chi-square test that examines the degree to which observed data from a set of categories arranged along two factors coincide with theoretical or previously empirical-derived expectations.

Test statistic In hypothesis testing, the obtained value that is compared with the critical value to determine statistical significance.

Theory An attempt to explain and organize collections of data regarding a topic of interest (or "phenomenon") by referring to general principles and relationships that are independent of the topic to be explained.

Treatment In experimental research, an analogous term to "independent variable" – most often used when its presence may result in the change of a participant's behavior.

Treatment variance (or primary variance) In an ANOVA, the amount of variance among sample means due to the action of the independent variable or if nonexperimental, group membership.

Truncated range When one end of a distribution of sample scores is arbitrarily cut off. In a correlational analysis, if one variable has a truncated range, the correlation of the sample will tend to underestimate the population correlation.

Tukey's *HSD* (Honestly Significant Difference) A conservative multiple comparison test used to locate the source of significance in designs with more than two conditions.

Two-tailed (or nondirectional) test A statistical test in which the critical region is divided equally between the two tails of the sampling distribution.

Two-way ANOVA A statistical test designed to determine if there are population differences within a research design containing two factors (or independent variables).

Type I error Rejecting a true null hypothesis.

Type II error Failing to reject a false null hypothesis.

Understanding The ability to make some cause-and-effect statement regarding a topic of interest and other variables. This is the highest goal of the researcher.

Unimodal distribution A distribution with one mode.

Univariate distribution A frequency distribution based on one variable.

***U* value** The test statistic used in the Mann–Whitney *U* test.

Variance A measure of dispersion or variability of a distribution of scores. The variance is the average squared deviation score. Because of the squaring, it is not stated in the original units of the measured variable, unlike the standard deviation. However, the variance is a statistic commonly used in formulas for hypothesis testing.

Weighted mean The mean of two or more individual means in which the individual means are weighted according to their respective sample sizes.

Wilcoxon signed-ranks test A nonparametric hypothesis test used to compare two dependent samples. It is the nonparametric counterpart to the dependent-samples *t* test.

Within-group variation A measure of the variation of scores within a group.

Within-participants design See *Repeated-measures design*.

***Y* intercept** The point on the *Y* axis at which a prediction line intersects.

***z* score** A transformed raw score that indicates the number of standard deviation units the raw score is from the mean of the distribution.

***z* test** A significance test used to decide if a sample mean comes from a population with a specified mean. This test may also be used to compare two sample means. In either case, the *z* test requires knowledge of the population standard deviation.

68-95-99.7 rule If a data set is normally distributed, this rule is used to understand how standard deviations roughly approximate how scores are distributed. Roughly 68% of scores fall within ±1 standard deviation of the mean; roughly 95% of scores fall within ±2 standard deviations of the mean; and roughly 99.7% of scores will fall within ±3 standard deviations of the mean.

List of Selected Formulas

Population/ sample mean	$\mu = \frac{\Sigma X}{N} \quad M = \frac{\Sigma X}{n}$	(3.1a, 3.1b)
Weighted mean	$M = \frac{n_1(M_1) + n_2(M_2)\cdots + \cdots n_n(M_n)}{n_1 + n_2\cdots + \cdots n_n}$	(3.2)

Definitional formulas

Population variance	$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$	(4.6)
Sample variance	$s^2 = \frac{\Sigma(X-M)^2}{n-1}$	(4.7)

Computational formulas

Population variance	$\sigma^2 = \frac{\Sigma X^2 - \left[(\Sigma X)^2/N\right]}{N}$	(4.12)
Sample variance	$s^2 = \frac{\Sigma X^2 - \left[(\Sigma X)^2/n\right]}{n-1}$	(4.13)

Any formula for the standard deviation is the square root of the variance formula

Population/ sample formulas for *z*	$z = \frac{X-\mu}{\sigma} \quad z = \frac{X-M}{s}$	(5.3a, 5.3b)
Population/ sample formulas for *X* given *z*	$X = \mu + z\sigma \quad X = M + zs$	(5.4a, 5.4b)

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

Probability of favorable event

$$P = \frac{(number\ of\ \ favorable\ events)}{(total\ number\ of\ \ events)} \tag{6.1}$$

Addition rule formula for two events

$$P(A\ \ or\ \ B) = P(A) + P(B) - P(A\ \ and\ \ B) \tag{6.4}$$

Multiplication rule formula for two events

$$P(A\ \ and\ \ B) = P(A|\,B)P(B) \tag{6.6}$$

Conditional probability formula

$$P(A|B) = \frac{P(A\ and\ B)}{P(B)} \tag{6.7}$$

Bayes' theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|notB)P(notB)} \tag{6.8}$$

Standard error of the mean

$$\sigma_M = \frac{\sigma}{\sqrt{n}} \tag{7.2}$$

Estimated standard error of the mean

$$s_M = \frac{s}{\sqrt{n}} \tag{7.3}$$

Single-sample *z* and *t* test

$$z_{obt} = \frac{M-\mu}{\sigma_M} \quad t_{obt} = \frac{M-\mu}{s_M} \quad df = n - 1 \tag{8.1, 8.3}$$

Cohen's *d* for single-sample *z* and *t* test

$$d = \frac{M-\mu}{\sigma} \quad d = \frac{M-\mu}{s} \tag{8.2, 8.4}$$

Definitional formula for $s_{M_1-M_2}$

$$s_{M_1-M_2} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{9.2}$$

The pooled variance

$$s_p^2 = s_1^2(n_1-1) + \frac{s_2^2(n_2-1)}{n_1+n_2-2} \tag{9.3}$$

Variance formula for $s_{M_1-M_2}$

$$s_{M_1-M_2} = \sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{9.4}$$

Computational formula for $s_{M_1-M_2}$

$$s_{M_1-M_2} = \sqrt{\frac{\left(\Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1}\right) + \left(\Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n_2}\right)}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{9.5}$$

The *t* ratio

$$t_{obt} = \frac{M_1 - M_2}{s_{M_1 - M_2}} \quad df = n_1 + n_2 - 2 \tag{9.7}$$

Cohen's *d* for independent-samples *t* test

$$d = \frac{M_1 - M_2}{\sqrt{s_p^2}} \tag{9.8}$$

The estimate of the standard error of the difference, $s_{\bar{D}}$

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n_p}} \tag{10.1}$$

Computational formula for s_D

$$s_D = \sqrt{\frac{\Sigma D^2 - \left[(\Sigma D)^2 / n_p\right]}{n_p - 1}} \tag{10.2}$$

Dependent-samples *t* test

$$t_{obt} = \frac{(M_X - M_Y)}{s_{\bar{D}}} \quad df = n_p - 1 \tag{10.4}$$

Cohen's *d* for dependent-samples *t* test

$$d = \frac{M_X - M_Y}{s_D} \tag{10.5}$$

Delta

$$\delta = \gamma\sqrt{n} \tag{11.1}$$

Gamma

$$\gamma = \frac{\mu_{alt} - \mu_0}{\sigma} \tag{11.2}$$

Determining sample size for a single-sample *t* test

$$n = \left(\frac{\delta}{\gamma}\right)^2 \tag{11.3}$$

Computational formula for SS_{BG}

$$SS_{BG} = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right] \tag{12.4}$$

Computational formula for SS_W

$$SS_W = \Sigma X^2 - \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k}\right] \tag{12.6}$$

Computational formula for SS_T

$$SS_T = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = SS_{BG} + SS_W \tag{12.8}$$

$$F = \frac{MS_{BG}}{MS_W} \text{ where each } MS = \frac{SS}{df} \quad df_T = N - 1;$$

$$df_{BG} = k - 1; \quad df_W = N - k$$

Omega-squared, ω^2

$$\varpi^2 = \frac{SS_{BG} - df_{BG}(MS_W)}{SS_T + MS_W} \tag{12.9}$$

Eta-squared, η^2

$$\eta^2 = \frac{SS_{BG}}{SS_T} \tag{12.10}$$

Tukey's *HSD*

$$HSD = q\sqrt{\frac{MS_W}{n}} \tag{12.11}$$

Fisher's *LSD* test

$$t = \frac{M_i - M_j}{\sqrt{MS_W\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{12.12}$$

Computational formula for SS_T

$$SS_T = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \tag{13.2}$$

Computational formula for SS_W

$$SS_W = \Sigma X^2 - \left[\frac{(\Sigma X_{A_1B_1})^2}{n_{A_1B_1}} + \frac{(\Sigma X_{A_1B_2})^2}{n_{A_1B_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k}\right] \tag{13.6}$$

Computational formula for SS_A

$$SS_A = \frac{(\Sigma X_{A_1})^2}{n_{A_1}} + \frac{(\Sigma X_{A_2})^2}{n_{A_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right] \tag{13.8}$$

Computational formula for SS_B

$$SS_B = \frac{(\Sigma X_{B_1})^2}{n_{B_1}} + \frac{(\Sigma X_{B_2})^2}{n_{B_2}} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N}\right] \tag{13.10}$$

Computational formula for $SS_{A \times B}$

$$\begin{aligned} SS_{A\times B} = n_k[&(M_{A_1B_1} - M_{A_1} - M_{B_1} + M_G)^2 \\ &+ (M_{A_2B_1} - M_{A_2} - M_{B_1} + M_G)^2 \\ &+ (M_{A_1B_2} - M_{A_1} - M_{B_2} + M_G)^2 \\ &+ (M_{A_2B_2} - M_{A_2} - M_{B_2} + M_G)^2] \end{aligned} \tag{13.11}$$

$$F_A = \frac{MS_A}{MW_W};\ F_B = \frac{MS_B}{MW_W};\ F_{A\times B} = \frac{MS_{A\times B}}{MW_W}$$

where each $MS = \dfrac{SS}{df}$

$$df_T = N - 1;\ df_W = N - k;\ df_A = levels - 1;$$
$$df_B = levels - 1;\ df_{A\times B} = df_A \times df_B$$

Omega-squared, ϖ_A^2, ϖ_B^2, $\varpi_{A\times B}^2$

$$\varpi_A^2 = \frac{SS_A - (df_A)MS_W}{SS_T + MS_W} \quad \varpi_B^2 = \frac{SS_B - (df_B)MS_W}{SS_T + MS_W} \tag{12.12–13.14}$$

$$\varpi_{A\times B}^2 = \frac{SS_{A\times B} - (df_{A\times B})MS_W}{SS_T + MS_W}$$

Eta-squared, η_A^2, η_B^2, $\eta_{A \times B}^2$

$$\eta_A^2 = \frac{SS_A}{SS_A + SS_W} \quad \eta_B^2 = \frac{SS_B}{SS_B + SS_W} \quad \eta_{A \times B}^2 = \frac{SS_{A \times B}}{SS_{A \times B} + SS_W} \tag{13.15–13.17}$$

Tukey's *HSD* values, HSD_A, HSD_B, $HSD_{A \times B}$

$$HSD_A = q_A \sqrt{\frac{MS_W}{n_A}} \quad HSD_B = q_B \sqrt{\frac{MS_W}{n_B}} \quad HSD_{A \times B} = q_{A \times B} \sqrt{\frac{MS_W}{n_{A \times B}}} \tag{13.18–13.20}$$

Fisher's *LSD* test, two-way ANOVA

$$t = \frac{M_i - M_j}{\sqrt{MS_W \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \tag{13.21}$$

Computational formula for SS_T

$$SS_T = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \tag{14.1}$$

Computational formula for SS_{BG}

$$SS_{BG} = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k} - \left[\frac{(\Sigma X)^2}{N} \right] \tag{14.2}$$

Computational formula for SS_W

$$SS_W = \Sigma X^2 - \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \cdots + \frac{(\Sigma X_k)^2}{n_k} \right] \tag{14.3}$$

Computational formula for SS_{BP}

$$SS_{BP} = \frac{(P_1)^2}{k} + \frac{(P_2)^2}{k} + \cdots + \frac{(P_n)^2}{k} - \left[\frac{(\Sigma X)^2}{N} \right] \tag{14.4}$$

Formula for SS_{error}

$$SS_{error} = SS_W - SS_{BP}$$

$$F = \frac{MS_{BG}}{MS_{error}} \text{ where each } MS = \frac{SS}{df}$$

$$df_{BG} = k - 1; \quad df_{error} = (N - k) - (n - 1)$$

Repeated-measures omega-squared, ω^2

$$\varpi^2 = \frac{SS_{BG} - df_{BG}(MS_{error})}{SS_T + MS_{error}} \tag{14.5}$$

Repeated-measures eta-squared, η^2

$$\eta^2 = \frac{SS_{BG}}{SS_T} \tag{14.6}$$

Formula for Tukey's *HSD*, repeated-measures

$$HSD = q\sqrt{\frac{MS_{error}}{n}} \tag{14.7}$$

Repeated-measures Fisher's *LSD*

$$t = \frac{M_i - M_j}{\sqrt{MS_{error}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{14.8}$$

***z* score formula for the Pearson population correlation**

$$\rho = \frac{\Sigma(z_X z_Y)}{N_p} \tag{15.1}$$

The computational formula for Pearson's *r*

$$r = \frac{n_p(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{\left[n_p(\Sigma X^2) - (\Sigma X)^2\right]\left[n_p(\Sigma Y^2) - (\Sigma Y)^2\right]}} \tag{15.2}$$

Linear regression equation

$$Y_p = M_Y + b(X - M_X) \tag{16.5}$$

Computational formula for the slope

$$b = \frac{n_p(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\left[n_p(\Sigma X)^2\right] - (\Sigma X)^2} \tag{16.6}$$

Computational formula for s_e

$$s_e = \sqrt{\left[\frac{1}{n_p(n_p-2)}\right]\left[\left(n_p\Sigma Y^2 - (\Sigma Y)^2\right) - \left(\frac{\left[n_p\Sigma XY - (\Sigma X)(\Sigma Y)\right]^2}{n_p\Sigma X^2 - (\Sigma X)^2}\right)\right]} \tag{16.9}$$

Correlation formula for s_e

$$s_e = s_Y\sqrt{(1-r^2)\left[\frac{n_p}{(n_p-2)}\right]} \tag{16.10}$$

χ^2

$$\chi^2 = \sum\frac{(f_o - f_e)^2}{f_e} \tag{17.1}$$

f_e

$$f_e = \frac{f_c f_r}{N} \tag{17.2}$$

Cramér's *V*

$$\phi = \sqrt{\frac{\chi^2}{(N)\left(df_{row/column}\right)}} \tag{17.4}$$

Standardized residual

$$R = \frac{f_o - f_e}{\sqrt{f_e}} \tag{17.5}$$

Spearman rank correlation, r_s

$$r_s = 1 - \frac{6\Sigma D^2}{n_p\left(n_p^2 - 1\right)} \tag{18.1}$$

Point-biserial correlation, r_{pb}

$$r_{pb} = \frac{M_{Y_1} - M_{Y_0}}{s_y}\sqrt{\frac{n_1 n_0}{n(n-1)}} \tag{18.2}$$

U_A

$$U_A = n_A n_B + \frac{n_A(n_A+1)}{2} - \Sigma R_A \tag{18.3}$$

U_B

$$U_B = n_A n_B + \frac{n_B(n_B+1)}{2} - \Sigma R_B \tag{18.4}$$

The U to z_U transformation formula

$$z_U = \frac{U - (n_A n_B/2)}{\sqrt{\frac{n_A n_B(n_A + n_B + 1)}{12}}} \tag{18.5}$$

When using Formula 18.5, if either >20, critical values are found in the z table (0.10 = 1.65; 0.05 = 1.96; 0.01 = 2.58)

Wilcoxon signed-ranks test for large sample sizes

$$z_{obt} = \frac{T - [n(n+1)/4]}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \tag{18.6}$$

When using Formula 18.6, critical values are found in the z table (0.10 = 1.65; 0.05 = 1.96; 0.01 = 2.58). When $n < 50$, $df = n$ in Table A.12

List of Symbols

Symbols	
α	Level of significance; probability of Type I error
a	Y intercept
b	Slope
β	Probability of Type II error
B	Number of scores below an interval
$1 - \beta$	Power
C	Columns
χ^2	Chi-square
$cum\ f$	Cumulative frequency
d	Cohen's d effect size measure
D	Difference score
$\bar{D}$	Mean of difference scores
df	Degrees of freedom
df_A	Degrees of freedom for Factor A
$df_{A \times B}$	Degrees of freedom for the interaction
df_B	Degrees of freedom for Factor B
df_{BG}	Degrees of freedom between group
df_{error}	Degrees of freedom error
df_W	Degrees of freedom within group
δ	Delta
E	Exact number of scores in an interval

(*Continued*)

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

Symbols	
f	Frequency of occurrence of a score
f_c	Frequency of scores in a column
f_e	Expected frequency
f_o	Observed frequency
f_m	Number of scores in the critical interval
f_r	Frequency of scores in a row
F	Sum of all frequencies below the lower limit of a critical interval; F statistic in ANOVA
γ	gamma
η^2	Eta-squared; effect size measure
h	Interval width
H_0	Null hypothesis
H_1	Alternative hypothesis
HSD	Tukey's honestly significant difference test
k	Number of groups; number of the last group
L	Exact lower limit of a critical interval
LSD	Fisher's least significant difference test; protected t test
LL	Lower limit of an interval
μ	Population mean
μ_M	Mean of a sampling distribution of means
μ_{alt}	Mean of the alternative ("treated") population
μ_0	Population mean when null hypothesis is true
M	Sample mean
M_G	Grand mean; mean of several groups
MS	Mean square
MS_A	Mean square for Factor A
$MS_{A \times B}$	Mean square for $A \times B$ interaction
MS_B	Mean square for Factor B
MS_{BG}	Mean square between group
MS_W	Mean square within group
MS_{error}	Mean square error
N	Number of scores in a population
n	Number of scores in a sample
n_k	Number of scores in the last group
n_p	Number of pairs of scores

Symbols	
P	A person in a repeated-measures ANOVA
P	Probability
$P(A)$	Probability of event A
$P(B)$	Probability of event B
$P(A \text{ and } B)$	Probability of event A and event B co-occurring
$P(\text{not } B)$	Probability of event B not occurring
$P(A \text{ or } B)$	Probability of event A or event B
$P(A \mid B)$	Probability of event A given event B
$P(A \mid \text{not } B)$	Probability of event A occurring given event B did not occur
$P(B \mid A)$	Probability of event B given event A
PR	percentile rank
q	The studentized range statistic
Q_1	First quartile
Q_3	Third quartile
r	Pearson product moment correlation coefficient for a sample
ρ	Rho
r_{pb}	Point-biserial correlation coefficient
R	Standardized residual
r_s	Spearman rank correlation coefficient
Σ	Sigma (summation)
s	Sample standard deviation
s^2	Sample variance
s_D	Standard deviation of difference scores
$s_{\bar{D}}$	Standard error of the difference for dependent samples
s_e	Standard error of the estimate
s_p^2	Pooled variance
s_Y	Standard deviation of Y scores
s_M	Estimated standard error of the mean
$s_{M_1 - M_2}$	Estimated standard error of the difference between sample means
SS	Sum of squares
SS_{BG}	Between-group sum of squares
SS_{BS}	Between-subjects sum of squares
SS_{error}	Sum of squares error

(*Continued*)

Symbols	
SS_W	Within-group sum of squares
σ	Population standard deviation
σ^2	Population variance
σ_M	Standard error of the mean (population)
$\sigma_{M_1 - M_2}$	Standard error of the difference between means (population)
T	Wilcoxon T statistic
t	t statistic
t_{obt}	t test
UL	Upper limit of an interval
U	Mann–Whitney U statistic
ϕ	Cramér's V; effect size measure for chi-square
ω^2	omega-squared
X	A raw score from an X distribution
$\bar{X}$	Sample mean
X_H	Highest score in a distribution
X_L	Lowest score in a distribution
X_p	X given a percentile rank
x	Deviation score
Y_p	Y predicted
z	z score
z_{obt}	z test
z_U	U to z transformation for Mann–Whitney U test

Index

Statistical Applications for the Behavioral and Social Sciences, Second Edition.
K. Paul Nesselroade, Jr. and Laurence G. Grimm.

Companion website: http://www.wiley.com/go/Nesselroade/Statis_Apps_behavioral_sciences

d