The Essentials of Statistics

A Tool for Social Research 4e Joseph F. Healey

THE ESSENTIALS OF STATISTICS

A Tool for Social Research

Fourth Edition

Joseph F. Healey

Christopher Newport University

CENGAGE Learning®

Australia • Brazil • Japan • Korea • Mexico • Singapore • Spain • United Kingdom • United States

CENGAGE Learning

***The Essentials of Statistics: A Tool for Social Research*, Fourth Edition**
Joseph F. Healey

Product Director: Marta Lee-Perriard

Product Manager: Jennifer Harrison

Content Developer: Lori Bradshaw

Product Assistant: Chelsea Meredith

Media Developer: John Chell

Marketing Manager: Kara Kindstrom

Content Project Manager: Cheri Palmer

Art Director: Michelle Kunkler

Manufacturing Planner: Judy Inouye

Production, Composition, and Illustration: MPS Limited

Text Researcher: Lumina Datamatics

Copy Editor: MPS Limited

Text Designer: Diane Beasley

Cover Designer: Ramsdell Design, LLC

Cover Image: © matthi/Shutterstock.com

Library of Congress Control Number: 2014959328

ISBN: 978-1-305-09383-6

Cengage Learning
20 Channel Center Street
Boston, MA 02210
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at **www.cengage.com/global**.

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage Learning Solutions, visit **www.cengage.com**.

Purchase any of our products at your local college store or at our preferred online store **www.cengagebrain.com**.

Printed in the United States of America
Print Number: 01 Print Year: 2015

Brief Contents

Detailed Contents

Preface

Statistics are part of the everyday language of sociology and other social sciences (including political science, social work, public administration, criminal justice, urban studies, and gerontology). These disciplines are research-based and routinely use statistics to express knowledge and to discuss theory and research. To join the conversations being conducted in these disciplines, you must be literate in the vocabulary of research, data analysis, and scientific thinking. Knowledge of statistics will enable you to understand the professional research literature, conduct quantitative research yourself, contribute to the growing body of social science knowledge, and reach your full potential as a social scientist.

Although essential, learning (and teaching) statistics can be a challenge. Students in social science statistics courses typically have a wide range of mathematical backgrounds and an equally diverse set of career goals. They are often puzzled about the relevance of statistics for them, and, not infrequently, there is some math anxiety to deal with. This text introduces statistical analysis for the social sciences while addressing these realities.

The text makes minimal assumptions about mathematical background (the ability to read a simple formula is sufficient preparation for virtually all of the material in the text), and a variety of special features help students analyze data successfully. The text has been written especially for sociology and social work programs but is sufficiently flexible to be used in any program with a social science base.

The text is written at an intermediate level and its intent is to show the relevance and value of statistics for the social sciences. I emphasize interpretation and understanding statistics in the context of social science research, but I have not sacrificed comprehensive coverage or statistical correctness. Mathematical explanations are kept at an elementary level, as is appropriate in a first exposure to social statistics. For example, I do not treat formal probability theory per se in the text.[1] Rather, the background necessary for an understanding of inferential statistics is introduced, informally and intuitively, in Chapters 5 and 6 while considering the concepts of the normal curve and the sampling distribution.

The text does not claim that statistics are "fun" or that the material can be mastered without considerable effort. At the same time, students are not overwhelmed with abstract proofs, formula derivations, or mathematical theory, which can needlessly frustrate the learning experience at this level.

[1]A presentation of probability is available at the website for this text for those who are interested.

Goals of the Text

The primary goal of this text is to develop basic statistical literacy. The statistically literate person understands and appreciates the role of statistics in the research process, is competent to perform basic calculations, and can read and appreciate the professional research literature in his or her field as well as any research reports he or she may encounter outside academia.

The goal of achieving basic statistical literacy has not changed since the first edition of this text. However, in recognition of the fact that "mere computation" has become less of a challenge in this high-tech age, this edition continues to increase the stress on interpretation and computer applications while deemphasizing computation. This will be apparent in several ways:

- In recognition of the fact that modern technology has rendered hand calculation increasingly obsolete, the end-of-chapter problems feature smaller, easier-to-handle datasets, although some more challenging problems are also included.
- A section called "Using SPSS" has been added to most chapters to demonstrate how to use a computerized statistical package to produce the statistics covered in that chapter.
- The end-of-chapter problems now include SPSS-based exercises, and research projects using SPSS are included at the end of almost all chapters.
- To accommodate the increased use of SPSS, several new datasets have been added to the text and the General Social Survey dataset has been updated to 2012.

The three aspects of basic statistical literacy provide a framework for discussing the additional features of this text.

1. **An Appreciation of Statistics.** A statistically literate person understands the relevance of statistics for social research, can analyze and interpret the meaning of a statistical test, and can select an appropriate statistic for a given purpose and a given set of data. This textbook develops these skills, within the constraints imposed by the introductory nature of the course, in several ways.
   - *The relevance of statistics.* Chapter 1 includes a discussion of the role of statistics in social research and stresses their usefulness as ways of analyzing and manipulating data and answering research questions. Throughout the text, each example problem is framed in the context of a research situation. A question is posed and then, with the aid of a statistic, answered. This central theme of usefulness is further reinforced by a series of "Applying Statistics" boxes, each of which illustrates some specific way statistics can be used to answer questions, and by the "Using Statistics" feature that opens every chapter.

     End-of-chapter problems are labeled by the social science discipline from which they are drawn: ***SOC*** for sociology, ***SW*** for social work, ***PS*** for political science, ***CJ*** for criminal justice, ***PA*** for public administration, and ***GER*** for gerontology. Identifying problems with specific disciplines allows students to more easily see the relevance of statistics to their own academic interests. (Not incidentally, they will also see that the disciplines have a large subject matter in common.)

Also, a series of boxed features entitled "Statistics in Everyday Life" appear in each chapter and highlight the relevance of statistics in the real world and provide examples of everyday applications.

- *Interpreting statistics.* For most students, interpretation—saying what statistics mean—is a big challenge. The ability to interpret statistics can be developed only by exposure and experience. To provide exposure, I have been careful, in the example problems, to express the meaning of the statistic in terms of the original research question. To provide experience, the end-of-chapter problems call for an interpretation of the statistic calculated. To provide examples, many of the answers to odd-numbered computational problems in the back of the text are expressed in words as well as numbers.
- *Using statistics: Ideas for research projects*. Appendix E offers ideas for independent data-analysis projects for students. The projects require students to use SPSS to analyze a dataset. They can be assigned at intervals throughout the semester or at the end of the course. Each project provides an opportunity for students to practice and apply their statistical skills and, above all, to exercise their ability to understand and interpret the statistics they produce.

2. **Computational Competence.** Students should emerge from their first course in statistics with the ability to perform elementary forms of data analysis. While computers have made computation less of an issue today, computation is inseparable from statistics, so I have included a number of features to help students cope with these mathematical challenges.
   - *"One Step at a Time"* boxes for each statistic break down computation into individual steps for maximum clarity and ease.
   - *Extensive problem sets* are provided at the end of each chapter. For the most part, these problems use fictitious data and are designed for ease of computation.
   - *Solutions* to odd-numbered computational problems are provided so that students may check their answers.
   - *SPSS* gives students access to the computational power of the computer. This is explained in more detail later.

3. **The Ability to Read the Professional Social Science Literature.** The statistically literate person can comprehend and critically appreciate research reports written by others. The development of this skill is a particular problem at the introductory level because (1) the vocabulary of professional researchers is so much more concise than the language of the textbook, and (2) the statistics featured in the literature are more advanced than those covered at the introductory level. This text helps to bridge this gap by
   - Always expressing the meaning of each statistic in terms of answering a social science research question.

- Providing a series of boxed inserts called "Becoming a Critical Consumer" that help students decipher the statistics they are likely to encounter in everyday life as well as in the professional literature.

Additional Features

A number of other features make the text more meaningful for students and more useful for instructors.

- *Readability and clarity.* The writing style is informal and accessible to students without ignoring the traditional vocabulary of statistics. Problems and examples have been written to maximize student interest and to focus on issues of concern and significance. For the more difficult material (such as hypothesis testing), students are first walked through an example problem before being confronted by formal terminology and concepts. Each chapter ends with a summary of major points and formulas and a glossary of important concepts. A list of frequently used formulas inside the covers and a glossary can be used for quick reference.
- *Organization and coverage.* The text is divided into four parts, with most of the coverage devoted to univariate descriptive statistics, inferential statistics, and bivariate measures of association. The distinction between description and inference is introduced in the first chapter and maintained throughout the text. In selecting statistics for inclusion, I have tried to strike a balance between the essential concepts with which students must be familiar and the amount of material students can reasonably be expected to learn in their first (and perhaps only) statistics course, all the while bearing in mind that different instructors will naturally wish to stress different aspects of the subject. Thus, the text covers a full gamut of the usual statistics, with each chapter broken into subsections so that instructors may choose the particular statistics they wish to include.
- *Learning objectives.* Learning objectives are stated at the beginning of each chapter. These are intended to serve as "study guides" and to help students identify and focus on the most important material.
- *Using Statistics.* At the beginning of each chapter, some applications of the statistics to be introduced are presented to give students a context for appreciating the material and some further examples of the usefulness of statistics.
- *Review of mathematical skills.* A comprehensive review of all of the mathematical skills that will be used in this text is included as a Prologue. Students who are inexperienced or out of practice with mathematics can study this review early in the course and/or refer to it as needed. A self-test is included so that students can check their level of preparation for the course.
- *Statistical techniques and end-of-chapter problems are explicitly linked.* After a technique is introduced, students are directed to specific problems for practice and review. The "how-to-do-it" aspects of calculation are immediately and clearly reinforced.
- *End-of-chapter problems are organized progressively.* Simpler problems with small datasets are presented first. Often, explicit instructions or hints accompany

the first several problems in a set. The problems gradually become more challenging and require more decision making by the student (e.g., choosing the most appropriate statistic for a certain situation). Thus, each problem set develops problem-solving abilities gradually and progressively.

- *Computer applications.* This text integrates SPSS, the leading social science statistics package, to help students take advantage of the power of the computer. Appendix F provides an introduction to SPSS, and demonstrations are integrated into the chapters. SPSS-based problems are included at the end of chapters, and research projects using SPSS are presented in the "You Are the Researcher" feature.
- *Realistic, up-to-date data.* The databases for computer applications in the text include a shortened version of the 2012 General Social Survey, a dataset that includes census and crime data for the 50 states, and a dataset that includes demographic data for 99 nations. These databases will give students the opportunity to practice their statistical skills on "real-life" data. All databases are described in Appendix G.

Additional Course Design Resources

- *Online PowerPoint® Slides.* A revised series of PowerPoint slides allows instructors to deliver class lectures and presentations discussing chapter-by-chapter content.
- *Online Instructor's Manual/Testbank.* The Instructor's Manual includes chapter summaries, a test item file of multiple-choice questions, answers to even-numbered computational problems, and step-by-step solutions to selected problems. In addition, the Instructor's Manual includes cumulative exercises (with answers) that can be used for testing purposes. To access these instructor resources, please log in to your account at **https://login.cengage.com**.
- *Aplia*™ is an online interactive learning solution that can be assigned as part of the course. Aplia integrates a variety of media and tools such as video, tutorials, practice tests, and an interactive e-book, and provides students with detailed, immediate feedback on every question. For more information about how to use Aplia in your course, please work with your local Cengage Learning Consultant.

Changes to the Fourth Edition

The following are the most important changes to this edition of *Essentials*:

- SPSS has been moved to a more central place in the text:
  - Almost all chapters have new sections ("Using SPSS") that illustrate how to produce the statistics covered in the chapter.
  - SPSS problems have been added to the end-of-chapter problems throughout the text. In some chapters (e.g., Chapters 12 and 13), the SPSS problems replace problems using hand calculators.

- For statistics that require complex computation—such as Pearson's *r* (Chapter 12) and partial correlation, multiple correlation, and regression (Chapter 13)—explanations and examples are now SPSS-based.

- The datasets used in the text have been expanded and updated. The datasets are used throughout the text in the new "Using SPSS" sections, in the end-of-chapter problems, and in the "You Are the Researcher" projects at the end of most chapters. The datasets are available for downloading at the website for this text: **www.cengagebrain.com**; they include
  - A General Social Survey (GSS) dataset (*GSS2012.sav*), which has been updated to 2012.
  - A dataset that includes census and crime data on the 50 states (*States.sav*).
  - A dataset that includes mostly demographic data for 99 nations (*Intl-POP.sav*).
  - A fourth dataset (*CrimeTrends84-10.sav*) is used only for the graphing exercises in Chapter 2.
- Former Chapters 11 and 12 have been combined into a single chapter (Chapter 11, entitled "Bivariate Association for Nominal- and Ordinal-Level Variables"). This new chapter de-emphasizes phi and the mechanics of computation for gamma but still fully treats the analysis of association for variables organized in bivariate tables.
- Chapter 2 has been reorganized and now begins with frequency distributions.
- Boxplots have been added to Chapter 4.
- All chapters now begin with a "Using Statistics" box that cites examples of how the statistics presented in the chapter can be applied to social research and to everyday life.
- Most of the "Statistics in Everyday Life" boxes have been updated or changed.
- The "Becoming a Critical Consumer" inserts have been updated and many have been shortened or broken into separate boxes
- The datasets used for examples, in the boxed features, and in the end-of-chapter problems have been updated.
- Titles have been added to all boxed features to clarify content and purpose.

Numerous other changes have been made throughout the text, most of them minor. All are intended to clarify explanations and make the material more accessible to students. As with previous editions, my goal is to offer a comprehensive, flexible, and student-oriented book that will provide a challenging first exposure to social statistics.

Acknowledgments

This text, in one form or another, has been in development for 30 years. An enormous number of people have made contributions, both great and small, to this project and, at the risk of inadvertently omitting someone, I am bound to at least attempt to acknowledge my many debts.

Much of whatever integrity and quality this book has is a direct result of the very thorough (and often highly critical) reviews that have been conducted over the years. I am consistently impressed by the sensitivity of my colleagues to the needs of the students. Whatever failings are contained in the text are, of course, my responsibility and are probably the result of my occasional decisions not to follow the advice of my colleagues.

I would like to thank the instructors who made statistics understandable to me (Professors Satoshi Ito, Noelie Herzog, and Ed Erikson) and all of my colleagues at Christopher Newport University for their support and encouragement (especially Professors F. Samuel Bauer, Stephanie Byrd, Cheryl Chambers, Robert Durel, Marcus Griffin, Mai Lan Gustafsson, Kai Heiddemann, Ruth Kernodle, Michael Lewis, Marion Manton, Eileen O'Brien, Lea Pellet, Eduardo Perez, Virginia Purtle, Andrea Timmer, and Linda Waldron). Also, I thank all of my students for their patience and thoughtful feedback, and I am 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 Appendices B, C, and D from their book *Statistical Tables for Biological, Agricultural and Medical Research* (6th edition, 1974).

Finally, I want to acknowledge the support of my family and rededicate this work to them. I have the extreme good fortune to be a member of an extended family that is remarkable in many ways and that continues to increase in size. Although I cannot list everyone, I would especially like to thank the older generation (my mother, Alice T. Healey, may she rest in peace), my wife Patricia A. Healey, the next generation (my sons, Kevin and Christopher, my daughters-in-law, Jennifer and Jessica, my step-son Christopher Schroen, and step-daughters Kate Cowell and her husband Matt and Jennifer Schroen), and the youngest generation (Benjamin, Caroline, and Isabelle Healey and Abigail Cowell).

Prologue: Basic Mathematics Review

You will probably be relieved to hear that this text, your first exposure to statistics for social science research, does not stress computation per se. While you will encounter many numbers to work with and numerous formulas to solve, the major emphasis will be on understanding the role of statistics in research and the logic we use to answer research questions empirically. You will also find that, in this text, the example problems and many of the homework problems have been intentionally simplified so that the computations will not unduly impede the task of understanding the statistics themselves.

On the other hand, you may regret to learn that there is, inevitably, some arithmetic that you simply cannot avoid if you want to master this material. It is likely that some of you haven't had any math in a long time, others have convinced themselves that they just cannot do math under any circumstances, and still others are just rusty and out of practice. All of you will find that mathematical operations that might seem complex and intimidating at first glance can be broken down into simple steps. If you have forgotten how to cope with some of these steps or are unfamiliar with these operations, this prologue is designed to ease you into the skills you will need to do all of the computations in this textbook. Also, you can use this section for review whenever you feel uncomfortable with the mathematics in the chapters to come.

Calculators and Computers

A calculator is a virtual necessity for this text. Even the simplest, least expensive model will save you time and effort and is definitely worth the investment. However, I recommend that you consider investing in a more sophisticated calculator with memories and preprogrammed functions, especially the statistical models that can compute means and standard deviations automatically. Calculators with these capabilities are available for around $20.00 to $30.00 and will almost certainly be worth the small effort it will take to learn to use them.

In the same vein, there are several computerized statistical packages (or **statpaks**) commonly available on college campuses that can further enhance your statistical and research capabilities. The most widely used of these is the Statistical Package for the Social Sciences (**SPSS**). Statistical packages like SPSS are many times more powerful than even the most sophisticated handheld calculators, and it will be well worth your time to learn how to use them because they will eventually save you time and effort.

SPSS is introduced in Appendix F of this text and is integrated into almost all the chapters. There are demonstrations that show you, step by step, how to use the program to generate the statistics covered in the chapter and end-of-chapter problems that require you to apply the program. Furthermore, the "You Are the Researcher" feature at the end of most chapters gives you the opportunity to use SPSS in some simplified social research projects.

There are many other programs that can help you calculate statistics with a minimum of effort and time. Even spreadsheet programs such as Microsoft® Excel, which is included in many versions of Microsoft Office, have some statistical capabilities. You should be aware that all of these programs (other than the simplest calculators) will require some effort to learn, but the rewards will be worth the effort.

In summary, you should find a way at the beginning of this course—with a calculator, a statpak, or both—to minimize the tedium of mere computing. This will permit you to devote maximum effort to the truly important goal of increasing your understanding of the meaning of statistics in particular and social research in general.

Variables and Symbols

Statistics are a set of techniques by which we can describe, analyze, and manipulate variables. A **variable** is a trait that can change value from case to case or from time to time. Examples of variables include height, weight, level of prejudice, and political party preference. The possible values or scores associated with a given variable might be numerous (for example, income) or relatively few (for example, gender). I will often use symbols, usually the letter X, to refer to variables in general or to a specific variable.

Sometimes we will need to refer to a specific value or set of values of a variable. This is usually done with the aid of subscripts. So the symbol X_1 (read "X-sub-one") would refer to the first score in a set of scores, X_2 ("X-sub-two") to the second score, and so forth. Also, we will use the subscript i to refer to all the scores in a set. Thus, the symbol X_i ("X-sub-i") refers to all of the scores associated with a given variable (for example, the test grades of a particular class).

Operations

You are all familiar with the four basic mathematical operations of addition, subtraction, multiplication, and division and the standard symbols ($+$, $-$, $\times$, $\div$) used to denote them. I should remind you that multiplication and division can be symbolized in a variety of ways. For example, the operation of multiplying some number a by some number b may be symbolized in (at least) six different ways:

$$a \times b$$

$$a \cdot b$$

$$a * b$$

$$ab$$
$$a(b)$$
$$(a)(b)$$

In this text, we will commonly use the "adjacent symbols" format (that is, *ab*), the conventional times sign (×), or adjacent parentheses to indicate multiplication. On most calculators and computers, the asterisk (*) is the symbol for multiplication.

The operation of division can also be expressed in several different ways. In this text, in addition to the standard symbol for division, we will use either of these two methods:

$$a/b \text{ or } \frac{a}{b}$$

Several formulas require us to find the square of a number. To do this, multiply the number by itself. This operation is symbolized as X^2 (read "*X* squared"), which is the same thing as $(X)(X)$. If X has a value of 4, then

$$X^2 = (X)(X) = (4)(4) = 16$$

or we could say that "4 squared is 16."

The square root of a number is the value that, when multiplied by itself, results in the original number. So the square root of 16 is 4 because (4)(4) is 16. The operation of finding the square root of a number is symbolized as

$$\sqrt{X}$$

A final operation with which you should be familiar is summation, or the addition of the scores associated with a particular variable. When a formula requires the addition of a series of scores, this operation is usually symbolized as ΣX_i. "Σ" is uppercase Greek letter sigma and stands for "the summation of." So the combination of symbols ΣX_i means "the summation of all the scores" and directs us to add all the scores for that variable. If four people had family sizes of 2, 4, 5, and 7, then the summation of these four scores for this variable could be symbolized as

$$\Sigma X_i = 2 + 4 + 5 + 7 = 18$$

The symbol Σ is an operator, just like the + and × signs. It directs us to add all of the scores on the variable indicated by the X symbol.

There are two other common uses of the summation sign, and, unfortunately, the symbols denoting these uses are not, at first glance, sharply different from each other or from the symbol used earlier. Some careful attention to these various meanings should minimize the confusion.

The first set of symbols is ΣX^2, which means "the sum of the squared scores." This quantity is found by *first* squaring each of the scores and *then* adding the squared scores together. A second common set of symbols will be $(\Sigma X_i)^2$, which means "the sum of the scores, squared." This quantity is found by *first* summing the scores and *then* squaring the total.

These distinctions might be confusing at first, so let's use an example to help clarify the situation. Suppose we had a set of three scores: 10, 12, and 13. So

$$X_i = 10, 12, 13$$

The sum of these scores is

$$\Sigma X_i = 10 + 12 + 13 = 35$$

The sum of the squared scores would be

$$\Sigma X^2 = (10)^2 + (12)^2 + (13)^2 = 100 + 144 + 169 = 413$$

Take careful note of the order of operations here. First the scores are squared one at a time, and then the squared scores are added. This is a completely different operation from squaring the sum of the scores:

$$(\Sigma X_i)^2 = (10 + 12 + 13)^2 = (35)^2 = 1225$$

To find this quantity, first the scores are summed and then the total of all the scores is squared. The squared sum of the scores (1225) is *not* the same as the sum of the squared scores (413).

In summary, the operations associated with each set of symbols can be summarized as follows:

Symbols	Operations
ΣX_i	Add the scores.
ΣX_i^2	First square the scores and then add the squared scores.
$(\Sigma X_i)^2$	First add the scores and then square the total.

Operations with Negative Numbers

A number can be either positive (if it is preceded by a + sign or by no sign at all) or negative (if it is preceded by a − sign). Positive numbers are greater than 0, and negative numbers are less than 0. It is very important to keep track of signs because they will affect the outcome of virtually every mathematical operation. This section briefly summarizes the relevant rules for dealing with negative numbers.

First, adding a negative number is the same as subtraction. For example,

$$3 + (-1) = 3 - 1 = 2$$

Second, subtraction changes the sign of a negative number:

$$3 - (-1) = 3 + 1 = 4$$

Note the importance of keeping track of signs here. If you neglected to change the sign of the negative number in the second expression, you would get the wrong answer.

For multiplication and division, you need to be aware of various combinations of negative and positive numbers. Ignoring the case of all positive numbers, this leaves several possible combinations. A negative number times a positive number results in a negative value:

$$(-3)(4) = -12$$
$$(3)(-4) = -12$$

A negative number multiplied by a negative number is always positive:

$$(-3)(-4) = 12$$

Division follows the same patterns. If there is a negative number in the calculations, the answer will be negative. If both numbers are negative, the answer will be positive. So

$$\frac{-4}{2} = -2$$

and

$$\frac{4}{-2} = -2$$

but

$$\frac{-4}{-2} = 2$$

Negative numbers do not have square roots, because multiplying a number by itself cannot result in a negative value. Squaring a negative number always results in a positive value (see the multiplication rules earlier).

Accuracy and Rounding Off

A possible source of confusion in computation involves accuracy and rounding off. People work at different levels of precision and, for this reason alone, may arrive at different answers to problems. This is important because our answers can be at least slightly different if you work at one level of precision and I (or your instructor or your study partner) work at another. You may sometimes think you've gotten the wrong answer when all you've really done is round off at a different place in the calculations or in a different way.

There are two issues here: when to round off and how to round off. My practice is to work in as much accuracy as my calculator or statistics package will allow and then round off to two places of accuracy (two places beyond, or to the right of, the decimal point) only at the very end. If a set of calculations is lengthy and requires the reporting of intermediate sums or subtotals, I will round the subtotals off to two places as I go.

In terms of how to round off, begin by looking at the digit immediately to the right of the last digit you want to retain. If you want to round off to 100ths (two places beyond the decimal point), look at the digit in the 1000ths place (three places beyond the decimal point). If that digit is 5 or more, round up. For example, 23.346 would round off to 23.35. If the digit to the right is less than 5, round down. So, 23.343 would become 23.34.

Let's look at an additional example of how to follow these rules of rounding. If you are calculating the mean value of a set of test scores and your calculator shows a final value of 83.459067 and you want to round off to two places, look at the digit three places beyond the decimal point. In this case the value is 9 (greater than 5), so we would round the second digit beyond the decimal point up and report the mean as 83.46. If the value had been 83.453067, we would have reported our final answer as 83.45.

Formulas, Complex Operations, and the Order of Operations

A mathematical formula is a set of directions, stated in general symbols, for calculating a particular statistic. To "solve a formula," you replace the symbols with the proper values and then perform a series of calculations. Even the most complex formula can be simplified by breaking the operations down into smaller steps.

Working through the steps requires some knowledge of general procedure and the rules of precedence of mathematical operations. This is because the order in which you perform calculations may affect your final answer. Consider the following expression:

$$2 + 3(4)$$

If you add first, you will evaluate the expression as

$$5(4) = 20$$

but if you multiply first, the expression becomes

$$2 + 12 = 14$$

Obviously, it is crucial to complete the steps of a calculation in the correct order.

The basic rules of precedence are to find all squares and square roots first, then do all multiplication and division, and finally complete all addition and subtraction. So the expression

$$8 + 2 \times 2^2/2$$

would be evaluated as

$$8 + 2 \times \frac{4}{2} = 8 + \frac{8}{2} = 8 + 4 = 12$$

The rules of precedence may be overridden by parentheses. Solve all expressions within parentheses before applying the rules stated earlier. For most formulas in this text, the order of calculations will be controlled by the parentheses.

Consider the following expression:

$$(8 + 2) - \frac{4}{(5 - 1)}$$

Resolving the parenthetical expressions first, we would have

$$(8 + 2) - \frac{4}{(5 - 1)} = (10) - \frac{4}{4} = 10 - 1 = 9$$

A final operation you will encounter in some formulas in this text involves denominators of fractions that themselves contain fractions. In this situation, solve the fraction in the denominator first and then complete the division. For example,

$$\frac{15 - 9}{6/2}$$

would become

$$\frac{15 - 9}{6/2} = \frac{6}{3} = 2$$

When you are confronted with complex expressions such as these, don't be intimidated. If you're patient with yourself and work through them step by step, beginning with the parenthetical expressions, even the most imposing formulas can be managed.

Exercises

You can use the following problems as a self-test on the material presented in this review. If you can handle these problems, you're ready to do all of the arithmetic in this text. If you have difficulty with any of these problems, please review the appropriate section of this prologue. You might also want to use this section as an opportunity to become more familiar with your calculator. The answers are given immediately following these exercises, along with commentary and some reminders.

1. Complete each of the following:

- **a.** $17 \times 3 =$
- **b.** $17(3) =$
- **c.** $(17)(3) =$
- **d.** $17/3 =$
- **e.** $(42)^2 =$
- **f.** $\sqrt{113} =$

2. For the set of scores (X_i) of 50, 55, 60, 65, and 70, evaluate each of the following expressions:

$\Sigma X_i =$

$\Sigma X_i^2 =$

$(\Sigma X_i)^2 =$

3. Complete each of the following:

- **a.** $17 + (-3) + (4) + (-2) =$
- **b.** $(-27)(54) =$
- **c.** $(-14)(-100) =$
- **d.** $-34/(-2) =$
- **e.** $322/(-11) =$
- **f.** $\sqrt{-2} =$
- **g.** $(-17)^2 =$

4. Round off each of the following to two places beyond the decimal point:

- **a.** 17.17532
- **b.** 43.119
- **c.** 1076.77337
- **d.** 32.4641152301
- **e.** 32.4751152301

5. Evaluate each of the following:

- **a.** $(3 + 7)/10 =$
- **b.** $3 + 7/10 =$
- **c.** $\dfrac{(4 - 3) + (7 + 2)(3)}{(4 + 5)(10)} =$
- **d.** $\dfrac{22 + 44}{15/3} =$

Answers to Exercises

1. **a.** 51 **b.** 51 **c.** 51
(The obvious purpose of these first three problems is to remind you that there are several different ways of expressing multiplication.)
d. 5.67 (Note the rounding off.) **e.** 1764
f. 10.63

2. The first expression is "the sum of the scores," so this operation would be

$$\Sigma X_i = 50 + 55 + 60 + 65 + 70 = 300$$

The second expression is the "sum of the squared scores." So

$$\Sigma X_i^2 = (50)^2 + (55)^2 + (60)^2 + (65)^2 + (70)^2$$
$$\Sigma X_i^2 = 2500 + 3025 + 3600 + 4225 + 4900$$
$$\Sigma X_i^2 = 18{,}250$$

The third expression is "the sum of the scores, squared":

$$(\Sigma X_i)^2 = (50 + 55 + 60 + 65 + 70)^2$$
$$(\Sigma X_i)^2 = (300)^2$$
$$(\Sigma X_i)^2 = 90{,}000$$

Remember that ΣX_i^2 and $(\Sigma X_i)^2$ are two completely different expressions with very different values.

3. **a.** 16
b. −1458
c. 1400
d. 17
e. −29.27
f. Your calculator probably gave you some sort of error message for this problem, because negative numbers do not have square roots.
g. 289

4. **a.** 17.18
b. 43.12
c. 1076.77
d. 32.46
e. 32.48

5. **a.** 1
b. 3.7 (Note the importance of parentheses.)
c. 0.31
d. 13.2

1 Introduction

LEARNING OBJECTIVES

By the end of this chapter, you will be able to:

1. Describe the limited but crucial role of statistics in social research.
2. Distinguish among three applications of statistics and identify situations in which each is appropriate.
3. Identify and describe three levels of measurement and cite examples of variables from each.

USING STATISTICS

One of the most important themes of this text is that statistics are useful tools for analyzing and understanding information and for communicating our conclusions to others. To stress this theme, each chapter will begin with a list of situations in which statistics can be (and should be) usefully applied. We will focus on general examples in this introductory chapter, but in the rest of the text, this section will highlight the usefulness of the specific statistics presented in that chapter.

Statistics can be used to:

- Demonstrate the connection between smoking and cancer.
- Measure political preferences, including the popularity of specific candidates for office.
- Track attitudes about gay marriage, abortion, and other controversial issues over time.
- Compare the cost of living (housing prices and rents, the cost of groceries and gas, health care, and so forth) between different localities (cities, states, and even nations).
- Measure the efficiency and productivity of workers in a wide variety of occupations, including factory workers, computer programmers, and professional athletes.

Why Study Statistics?

Students sometimes wonder about the value of studying statistics. What, after all, do numbers and formulas have to do with understanding people and society? The value of statistics will become clear as we move from chapter to chapter but, for now, we can demonstrate the importance of statistics by considering **research**: a disciplined and careful attempt to answer questions, examine ideas, and test theories. Research can take numerous forms, and statistics are relevant for **quantitative research** projects, or projects that collect information in the form of numbers or **data**. **Statistics** are mathematical techniques used by social scientists to analyze data in order to answer questions and test theories.

What is so important about learning how to manipulate data? On one hand, some of the most important and enlightening works in the social sciences do not utilize statistics at all. There is nothing magical about data and statistics. The mere presence of numbers guarantees nothing about the quality of a scientific inquiry.

On the other hand, data can be the most trustworthy information available to the researcher and, consequently, deserve special attention. Data that have been carefully collected and thoughtfully analyzed can be the strongest, most objective foundations for building theory and enhancing our understanding of the social world.

Let us be very clear about one point: It is never enough merely to gather data (or any kind of information, for that matter). Even the most carefully collected numerical information does not and cannot speak for itself. The researcher must be able to use statistics effectively to organize, evaluate, and analyze the data. Without a good understanding of statistical analysis, the researcher will be unable to make sense of the data. Without an appropriate application of statistics, data are useless.

Statistics are an indispensable tool for the social sciences because they provide researchers with an array of useful techniques for evaluating ideas and testing theory. The next section describes the relationships among theory, research, and statistics in more detail.

The Role of Statistics in Scientific Inquiry

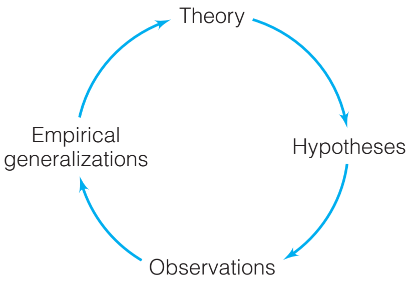
Figure 1.1 graphically represents the role of statistics in the research process. The diagram is based on the thinking of Walter Wallace and is a useful conception of how any scientific enterprise grows and develops.

One point the diagram makes is that scientific theory and research continually shape each other. Statistics are one of the most important means by which research and theory interact. Let's take a closer look at the process.

A Journey Through the Scientific Process

Because Figure 1.1 is circular, with no beginning or end, we could start our journey at any point. For the sake of convenience, let's begin at the top, with theory, and follow the arrows around the circle.

FIGURE 1.1 The Wheel of Science



Source: Adapted from Walter Wallace, *The Logic of Science in Sociology* (Chicago: Aldine-Atherton, 1971)

Theory. A **theory** is an explanation of the relationships among phenomena. People naturally (and endlessly) wonder about problems in society (such as prejudice, poverty, child abuse, and serial murders), and they develop explanations ("low levels of education cause prejudice") in their attempts to understand. Unlike our everyday, informal explanations, scientific theory is subject to a rigorous testing process. Let's take prejudice as an example to illustrate how the research process works.

Why are some people prejudiced against other groups? One possible answer to this question is provided by a theory that was published over 50 years ago by social psychologist Gordon Allport and that has been tested on a number of occasions.[1]

The theory states that prejudice will decrease in situations in which members of different groups have equal status and cooperate with each other to work toward mutual goals. The more equal and cooperative the contact, the more likely people will see each other as individuals and not as representatives of a particular group. For example, we might predict that members of a racially mixed athletic team who work together to win games will experience a decline in prejudice. On the other hand, when different groups compete for jobs, housing, or other resources, prejudice will tend to increase. Allport's theory is not a complete explanation of prejudice, of course, but it will serve to illustrate a sociological theory.

Variables and Causation. Note that Allport's theory is stated in terms of causal relationships between two variables. A **variable** is any trait that can change values from case to case; examples include gender, age, income, and political party affiliation as well as prejudice.

[1]Allport, Gordon, 1954. *The Nature of Prejudice*. Reading, MA: Addison-Wesley. This theory is often called "the contact hypothesis." For recent attempts to test this theory, see: McLaren, Lauren. 2003. "Anti-Immigrant Prejudice in Europe: Contact, Threat Perception, and Preferences for the Exclusion of Migrants." *Social Forces*, 81:909–937; Pettigrew, Thomas. 1997. "Generalized Intergroup Contact Effects on Prejudice." *Personality and Social Psychology Bulletin*, 23:173–185; and Pettigrew, T. F., Tropp, L. R., Wagner, U., & Christ, O. 2011. "Recent advances in intergroup contact theory." *International Journal of Intercultural Relations*, 35: 271–280.

A theory may identify some variables as causes and others as effects or results. In the language of science, the causes are **independent variables** and the effects or results are **dependent variables**. In our theory, equal-status contact would be the independent variable (or the cause) and prejudice would be the dependent variable (the result or effect). In other words, the theory argues that prejudice *depends on* (or is caused by) the extent to which a person participates in equal-status, cooperative contacts with members of other groups.

Diagrams can be a useful way of representing the relationships between variables:

$$\text{Equal-status contact} \rightarrow \text{Prejudice}$$
$$\text{Independent variable} \rightarrow \text{Dependent variable}$$
$$X \rightarrow Y$$

The arrow represents the direction of the causal relationship, and X and Y are general symbols for the independent and dependent variables, respectively.

Hypotheses. So far, we have a theory of prejudice and independent and dependent variables. Is the theory true or false? To find out, we need to compare our theory with the facts; we need to do some research.

Our next steps would be to define our terms and ideas. One problem we often face in research is that theories are complex and abstract, and we need to be very specific to conduct a valid test. Often, we do this by deriving a hypothesis from the theory: A **hypothesis** is a specific and exact statement about the relationship between variables.

For example, if we wished to test Allport's theory, we would have to say exactly what we mean by prejudice and we would need to describe "equal-status, cooperative contact" in detail. We would also review the research literature to help develop and clarify our definitions and our understanding of these concepts.

As our hypothesis takes shape, we begin the next step of the research process, during which we decide exactly how we will gather data. We must decide how cases will be selected and tested, how variables will be measured, and a host of related matters. Ultimately, these plans will lead to the observation phase (the bottom of the wheel of science), where we actually measure social reality. Before we can do this, we must have a very clear idea of what we are looking for and a well-defined strategy for conducting the search.

Making Observations and Using Statistics. To test Allport's theory of prejudice, we might begin with people from different racial or ethnic groups. We might place some subjects in cooperative situations and others in competitive situations. We would need to measure levels of prejudice before and after each type of contact. We might do this by administering a survey that asks subjects to agree or disagree with statements such as "Greater efforts must be made to racially integrate the public school system." Our goal would be to see whether the people exposed to the cooperative-contact situations actually become less prejudiced.

Now, finally, we come to the use of statistics in our research. During the observation phase, we will collect a great deal of numerical information or data.

If we have a sample of 100 people, we will have 200 completed surveys measuring prejudice: 100 completed before the contact situation and 100 filled out afterward. Try to imagine dealing with 200 completed surveys. If we ask each respondent just five questions to measure their prejudice, we will have a total of 1000 separate pieces of information to deal with. What will we do? We'll have to organize and analyze this information, and statistics will become very helpful at this point. They will supply us with many ideas about "what to do" with the data. We will begin to look at some of the options in the next chapter; for now, let me stress two points about statistics.

First, statistics are crucial. Simply put, without statistics, quantitative research is impossible. We need statistics to analyze data and to shape and refine our theories of the social world.

Second, and somewhat paradoxically, the role of statistics is limited. As Figure 1.1 makes clear, scientific research proceeds through several mutually interdependent stages. Statistics become directly relevant only at the end of the observation stage. Before any statistical analysis can be legitimately applied, however, earlier phases must have been successfully completed. If the researcher has asked poorly conceived questions or has made serious errors of design or method, then even the most sophisticated statistical analysis is valueless. As useful as they can be, statistics cannot substitute for rigorous conceptualization, detailed and careful planning, or creative use of theory. Statistics cannot salvage a poorly conceived or designed research project. They cannot make sense out of garbage.

On the other hand, inappropriate statistical applications can limit the usefulness of an otherwise carefully done project. Only by successfully completing *all* phases of the process can a quantitative research project hope to contribute to understanding. A reasonable knowledge of the uses and limitations of statistics is as essential to the education of the social scientist as is training in theory and methodology.

Empirical Generalizations. Our statistical analysis would focus on assessing our theory, but we would also look for other trends in the data. For example, if we found that equal-status, cooperative contact reduces prejudice in general, we might go on to ask whether the pattern applies to males as well as to females and to the well educated as well as to the poorly educated. As we probed the data, we might begin to develop some generalizations based on the empirical patterns we observe. For example, what if we found that contact reduced prejudice for younger respondents but not for older respondents? Could it be that younger people are less rigid in their thinking and have attitudes and feelings that are more open to change? As we developed tentative explanations, we would begin to revise or elaborate the theory that guides the research project.

A New Theory? If we changed our theory because of our empirical generalizations, a new research project would probably be needed to test the revised theory, and the wheel of science would begin to turn again. We (or perhaps other researchers) would go through the entire process once again with this new—and, hopefully, improved—theory. This second project might result in further revisions that would require still more research, and the wheel of science would

STATISTICS
IN EVERYDAY LIFE

Push Polls
Political campaigns sometimes use "push polls" to sway public opinion. These polls are designed to influence opinions, sometimes by starting or circulating rumors and innuendo. For example, they may try to brand an opponent as untrustworthy by asking questions like "Would you still support the candidate if you found out that he is an alcoholic?" Even when completely fabricated, such a question can create negative associations in the minds of voters. Statistics may be used to "analyze" data gathered by push polls (or by marketing campaigns that use similar techniques), but the results have little or no scientific validity.

continue to turn as long as scientists were able to suggest revisions or develop new insights. Every time the wheel turned, our understanding of the phenomenon under consideration would (hopefully) improve.

Summary: The Dialog Between Theory and Research. This description of the research process does not include white-coated, clipboard-carrying scientists who, in a blinding flash of inspiration, discover some fundamental truth about reality and shout "Eureka!" The truth is that, in the normal course of science, we rarely are in a position to declare a theory true or false. Rather, evidence will gradually accumulate over time, and ultimate judgments of truth will be the result of many years of hard work, research, and debate.

Let's briefly review our imaginary research project. We began with a theory and examined the various stages of the research project that would be needed to test that theory. We wound up back at the top of the wheel, ready to begin a new project guided by a revised theory. We saw how theory can motivate research and how our observations might cause us to revise theory and, thus, motivate a new research project: Theory stimulates research and research shapes theory. This constant interaction is the lifeblood of science and the key to enhancing our understanding of the social world.

The dialog between theory and research occurs at many levels and in multiple forms. Statistics are one of the most important links between these two realms. Statistics permit us to analyze data, to identify and probe trends and relationships, to develop generalizations, and to revise and improve our theories. As you will see throughout this text, statistics are limited in many ways. They are also an indispensable part of the research enterprise, an essential tool for conducting research and shaping theory. *(For practice in describing the relationship between theory and research and the role of statistics in research, see problems 1.1 and 1.2.)*

The Goals of This Text

Clearly, statistics are a crucial part of social science research and every social scientist needs some training in statistical analysis. In this section, we address how much training is necessary and the purpose of that training.

First, this textbook takes the view that statistics are tools: useful, but not ends in themselves. Thus, we will not take a "mathematical" approach to the subject, although we will cover enough material so that you can develop a basic understanding of why statistics do what they do. Instead, I will present statistics as tools that can be used to answer important questions, and our focus will be on how these techniques are applied in the social sciences.

Second, you will soon become involved in advanced coursework in your major fields of study, and you will find that much of the professional literature assumes at least basic statistical literacy. Furthermore, after graduation, many of you will find yourselves in positions—either in a career or in graduate school—where some understanding of statistics will be very helpful or perhaps even required. Very few of you will become statisticians per se, but you must have a grasp of statistics in order to read and critically appreciate the research literature of your discipline. As a student in the social sciences and in many careers related to the social sciences, you simply cannot realize your full potential without a background in statistics.

Within these constraints, this textbook is an introduction to statistics as they are used in the social sciences. The general goal of the text is to develop an appreciation—a healthy respect—for statistics and their place in the research process. You should emerge from this experience with the ability to use statistics intelligently and to know when other people have done so—and when they have not! You should be familiar with the advantages and limitations of the more commonly used statistical techniques, and you should know which techniques are appropriate for a given purpose. Finally, you should develop sufficient statistical and computational skills and enough experience in the interpretation of statistics to be able to carry out some elementary forms of data analysis by yourself.

Descriptive and Inferential Statistics

As noted earlier, statistics are tools used to analyze data and answer research questions. Two general classes of statistical techniques, introduced in this section, are available to accomplish this task.

Descriptive Statistics

The first class of techniques, known as **descriptive statistics**, is relevant in several different situations:

1. When a researcher needs to summarize or describe the distribution of a single variable. These statistics are called *univariate* ("one variable") descriptive statistics.
2. When the researcher wishes to describe the relationship between two or more variables. These statistics are called *bivariate* ("two variable") or *multivariate* (more than two variable) descriptive statistics.

Univariate Descriptive Statistics. Many of these statistics are probably familiar to you. For example, percentages, averages, and graphs can all be used to describe single variables.

STATISTICS
IN EVERYDAY LIFE

Using Descriptive Statistics
American society is increasingly connected to the Internet. In 2014, about 87% of all American adults used the Internet at least occasionally—a dramatic increase from 53% in 2000. However, connectedness is dependent on social class: 99% of the most affluent Americans used the Internet vs. only 77% of the least affluent. Are the less affluent being left behind? Will their lower use of this increasingly essential resource affect the education and job prospects of their children and thus perpetuate their lower income over the generations? What additional information would you need to answer these questions?

Sources: U.S. Bureau of the Census. 2012. *Statistical Abstract of the United States: 2012.* Available at http://www.census.gov/prod/2011pubs/12statab/infocomm.pdf

Pew Research Center. 2014. http://www.pewinternet.org/files/2014/02/12-internet-users-in-2014.jpg

To illustrate the usefulness of univariate descriptive statistics, consider the following problem: Suppose you wanted to summarize the income distribution of a community of 10,000 families. How would you do it? Obviously, you couldn't simply list all the incomes and let it go at that. You would want to use some summary measures of the overall distribution—perhaps a graph, an average, or the percentage of incomes that are low, moderate, or high. Whatever method you choose, its function is to reduce these thousands of items of information into a few clear, concise, and easily understood numbers. The process of using a few numbers to summarize many, called **data reduction**, is the basic goal of univariate descriptive statistics. Part I of this text is devoted to these statistics.

Bivariate and Multivariate Descriptive Statistics. The second type of descriptive statistics is designed to help us understand the relationship between two or more variables. These statistics, called **measures of association**, allow us to quantify the strength and direction of a relationship. We can use these statistics to investigate two matters of central theoretical and practical importance to any science: causation and prediction. Measures of association help us disentangle the connections between variables and trace the ways in which some variables might affect others. We can also use them to predict scores on one variable from scores on another.

For example, suppose you were interested in the relationship between two variables—amount of study time and grades—and had gathered data from a group of college students. By calculating the appropriate measure of association, you could determine the strength of the relationship and its direction. Suppose you found a strong, positive relationship. This would indicate that "study time" and "grade" were closely related (strength of the relationship) and that as one increased in value, the other also increased (direction of the relationship). You could make predictions from one variable to the other ("the longer the study time, the higher the grade").

Measures of association can give us valuable information about relationships between variables and help us understand how one variable causes another. One important point to keep in mind about these statistics, however, is that they cannot, by themselves, *prove* that two variables are causally related. Even if a measure of association shows a very strong relationship between study time and grades, we cannot conclude that one variable causes the other. Correlation is not the same thing as causation, and the mere existence of a correlation cannot prove that a causal relationship exists between variables. We will consider bivariate associations or correlations in Part III of this text, and we will cover multivariate analysis in Part IV.

Inferential Statistics

This second class of statistical techniques becomes relevant when we wish to generalize to a **population**, the total collection of all cases in which the researcher is interested and wishes to understand better. Examples of possible populations would be voters in the United States, all parliamentary democracies, or all unemployed people in Atlanta.

Populations can theoretically range from enormous ("all humanity") to quite small (all sophomores on your campus) but are usually fairly large. Social scientists almost never have the resources or time to test every case in a population, hence the need for **inferential statistics**. This class of techniques involves using information from **samples** (carefully chosen subsets of the population, often called random samples), to make inferences about populations. Because they have fewer cases, samples are cheaper to assemble, and—if the proper techniques are followed—generalizations based on samples can be very accurate representations of populations.

Many of the concepts and procedures involved in inferential statistics may be unfamiliar, but most of us are experienced consumers of inferential statistics—most familiarly, perhaps, in the form of public-opinion polls and election projections. When a public-opinion poll reports that 42% of the American electorate plans to vote for a certain presidential candidate, it is essentially reporting a generalization to a population ("the American electorate"—which numbers over 130 million people) from a carefully drawn sample (usually between 1000 and 3000 respondents). Inferential statistics will occupy our attention in Part II of this book. *(For practice in describing different statistical applications, see problems 1.3 and 1.7.)*

STATISTICS
IN EVERYDAY LIFE

Using Inferential Statistics

In 2014, a sample of 1028 adult Americans was asked about gay marriage. A majority of 55% (up from 27% in 1996) said that they believed that marriages between same-sex couples should be recognized as valid by the law while 42% (down from 68% in 1996) disagreed. Some say that American society is polarized on the issue of gay marriage. Do these statistics support that view? How?

Source: Gallup Polls. Available at http://www.gallup.com/poll/169640/sex-marriage-support-reaches-new-high.aspx

Level of Measurement

We have many statistical techniques from which to choose, as you will begin to see in the next chapter. How do we select the best statistic for a given situation?

The most basic and important guideline for choosing statistics is the **level of measurement**, or the mathematical nature of the variables under consideration. Some variables, such as age and income, have numerical scores (years, dollars) and can be analyzed in a variety of ways, using many different statistics. For example, we could summarize these variables with a mean or average and make statements like "The average income of residents of this city is $43,000" or "The average age of students on this campus is 19.7."

Other variables, such as gender and zip code, have "scores" that are really just labels, not numbers at all, and we have fewer options for analyzing them. The average would be meaningless as a way of describing these variables. Your personal zip code might *look* like a number, but it is merely an arbitrary label that happens to be expressed in digits. These "numbers" cannot be added or divided, and statistics like the average zip code of a group of people are pointless.

Level of measurement is crucial because our statistical analysis must match the mathematical characteristics of our variables. One of the first steps in any quantitative research project is to determine the level of measurement of our variables, and we will consider the level of measurement of variables throughout this text and each time a statistical technique is introduced.

There are three levels of measurement. In order of increasing sophistication, they are nominal, ordinal, and interval-ratio. Each will be discussed separately.

The Nominal Level of Measurement

Variables measured at the nominal level have non-numerical "scores" or categories. Examples of variables at this level include gender, zip code, political party affiliation, and religious preference. Statistical analysis with nominal-level variables is limited to comparing the relative sizes of the categories and making statements such as "There are more females than males in this dorm" or "The most common zip code on this campus is 20158."

Let's look at some examples of nominal-level variables and take a moment to consider some terminology. Earlier in this chapter, we defined a variable as any trait that can vary from case to case. A variable consists of categories or scores—the traits that can vary. We "measure" a variable (e.g., gender) by classifying a case into one of the categories (e.g., male or female) or assigning a score to it. Several examples of nominal-level variables and their scores or categories are presented in Table 1.1.

Note that there is a number assigned to each of the categories in Table 1.1 (e.g., 1 for Protestant, 2 for Catholic), as is commonly done in quantitative research, especially when the data are being prepared for computer analysis. Recall that, for nominal-level variables, these "numbers" are just labels and

TABLE 1.1 Some Examples of Nominal-Level Variables and Their Categories

Variables →	Gender	Political Party Preference	Religious Preference
Categories →	1. Male 2. Female	1. Democrat 2. Republican 3. Other 4. Independent	1. Protestant 2. Catholic 3. Jew 4. None 5. Other

cannot be treated mathematically: They cannot be added, divided, or otherwise manipulated. The categories or scores do not make up a mathematical scale; they are different from each other but not "more or less" or "higher or lower" than each other. Males and females differ in terms of gender, but neither category has more or less gender than the other. In the same way, a zip code of 54398 is different from but not "higher" than a zip code of 13427.

Nominal variables are rudimentary, but there are criteria that we need to observe in order to measure them adequately. These criteria, in fact, apply to variables measured at all levels. They are stated in Table 1.2 and illustrated in Table 1.3. We consider and illustrate the criteria one at a time.

1. The first criterion ("mutually exclusive") means that there can be no overlap between categories and no confusion or ambiguity about where a case belongs. To illustrate, consider Scale A in Table 1.3, which violates this principle because of the overlap between the categories "Protestant" and "Episcopalian."

2. The second criterion ("exhaustive") means that there must be a place for every case or score. Scale B in Table 1.3 violates this criterion because there is no category for people who have no religious affiliation and for those who belong to religions other than the three stated. Often, we include an "Other" category, as in Scale D, to make a list exhaustive.

TABLE 1.2 Criteria for Stating the Categories of Variables

The categories of variables must	
1.	Be mutually exclusive (Each case must fit into one and only one category.)
2.	Be exhaustive (There must be a category for every case.)
3.	Include elements that are similar (The cases in each category must be similar to each other.)

TABLE 1.3 Four Scales for Measuring Religious Affiliation

Scale A	Scale B	Scale C	Scale D
Protestant	Protestant	Protestant	Protestant
Episcopalian	Catholic	Non-Protestant	Catholic
Catholic	Jew		Jew
Jew			None
None			Other
Other			

3. The third criterion ("similar") means that the categories should include cases that are truly comparable. To put it another way, we should avoid lumping apples with oranges. There are no hard-and-fast guidelines for judging whether a category includes similar items. The researcher must make that decision in terms of the specific purpose of the research, and categories that are too broad for some purposes may be perfectly adequate for others. Scale C in Table 1.3 violates this criterion because it uses a category (Non-Protestant, which would include Catholics, Jews, Buddhists, Atheists, and so forth) that seems too broad for meaningful research.

   Scale D in Table 1.3 is probably the most common way religious preference has been measured in North America, and this scale would be adequate in many situations. However, it may be too general for some research purposes. For example, an investigation of moral questions such as support for legal abortion would probably have to distinguish among the various Protestant denominations, and an effort to document religious diversity certainly would need to add categories for Buddhists, Muslims, and numerous other religious faiths.

The Ordinal Level of Measurement

Variables measured at the ordinal level are more sophisticated than nominal-level variables. They have scores or categories that can be ranked from high to low, so, in addition to classifying cases into categories, we can describe the categories in terms of "more or less" with respect to each other. Thus, with ordinal-level variables, not only can we say that one case is different from another; we can also say that one case is higher or lower, more or less than another.

For example, the variable socioeconomic status is usually measured at the ordinal level, often using categories such as those in Table 1.4. Individual cases can be compared in terms of the categories into which they are classified. Thus, an individual classified as a 4 (upper class) would be ranked higher than an individual classified as a 2 (working class), and a lower-class person (1) would rank lower than a middle-class person (3). Other variables that are usually measured at the ordinal level include attitude and opinion scales, such as those that measure prejudice, alienation, or political conservatism.

TABLE 1.4 Measuring Socioeconomic Status

If you were asked to describe yourself, into which of the following classes would you say you belong?

Score	Class
1	Lower class
2	Working class
3	Middle class
4	Upper class

The major limitation of the ordinal level of measurement is that the scores have no absolute or objective meaning: They only represent position with respect to other scores. We can distinguish between high and low scores, but the distance between the scores cannot be described in precise numerical terms. Thus, in terms of the scale shown in Table 1.4, we know that a social class score of 4 is more than a score of 2 but we do not know whether it is twice as much as 2.

Our options for statistical analysis with ordinal-level variables are limited by the fact that we don't know the exact numerical distances from score to score. For example, addition (and other mathematical operations) assumes that the intervals between scores are exactly equal. If the distances from score to score are not equal, $2 + 2$ might equal 3 or 5 or even 15.

Strictly speaking, statistics such as the average, or mean (which requires that the scores be added together and then divided by the number of scores), are not permitted with ordinal-level variables. The most sophisticated mathematical operation fully justified with an ordinal variable is ranking categories and cases (although, as we will see, it is not unusual for social scientists to take liberties with this criterion).

The Interval-Ratio Level of Measurement[2]

The scores of variables measured at the interval-ratio level are actual numbers that can be analyzed with all possible statistical techniques. This means that we can add or multiply the scores, compute averages or square roots, or perform any other mathematical operation.

There are two crucial differences between ordinal-level and interval-ratio-level variables. First, interval-ratio-level variables have equal intervals from score to score. For example, age is an interval-ratio variable because the unit of measurement (years) has equal intervals (the distance from year to year is 365 days). Similarly, if we ask people how many siblings they have, we would produce a variable with equal intervals: Two siblings are 1 more than 1, and 13 is 1 more than 12.

Second, interval-ratio variables have true zero points. That is, the score of 0 for these variables is not arbitrary; it indicates the absence or complete lack of whatever is being measured. For example, the variable "number of siblings" has a true zero point because it is possible to have no siblings at all. Similarly, it is possible to have 0 years of education, no income at all, a score of 0 on a multiple-choice test, and to be 0 years old (although not for very long). Other examples of interval-ratio variables include number of children, life expectancy, and years married. All mathematical operations are permitted for variables measured at the interval-ratio level.

[2]Many statisticians distinguish between the interval level (equal intervals) and the ratio level (true zero point). I find the distinction unnecessarily cumbersome in an introductory text and will treat these two levels as one.

BECOMING A CRITICAL CONSUMER: Introduction

The goal of this text is to develop your ability to understand and analyze statistical information, and a special feature called "Becoming a Critical Consumer" will help you reach this goal. In this feature, we examine the "everyday" statistics you might encounter in the media, in casual conversation, and in the professional social science research literature. In this first installment, I briefly outline the activities that will be included in this feature. We'll start with social science research and then examine other applications of statistics.

As you probably already know, articles published in social science journals often use statistics that are, at this point in your education, completely indecipherable. Compared to this text, the language of the professional researcher is compact and dense. Space in research journals is expensive, and the typical research project analyzes many variables. Thus, a large volume of information must be summarized in very few words. Researchers may express in just a word or two a point that will take us a paragraph or more to state. Also, professional researchers assume a certain level of statistical knowledge in their audience: they write for colleagues, not for undergraduate students.

How can you bridge the gap that separates you from this literature? How can you understand these articles, which seem so challenging? The (unfortunate but unavoidable) truth is that a single course in statistics will not close the gap entirely. However, this text will help you read much of the research literature and give you the ability to critically analyze statistical information. It will help you decode research articles by explaining their typical reporting style and considering examples.

Simultaneously, you will develop your ability to critically analyze the statistics you encounter in everyday life. In this age of information, statistical literacy is not just for academics or researchers. A critical perspective on statistics can help you think more critically, assess the torrent of information, opinion, and facts that washes over us every day, and make better decisions on a broad range of issues.

For example, what do statements like the following really mean?

- The Democratic candidate for governor is projected to win 55% of the vote.
- The average life expectancy is 77 years.
- The number of cohabiting couples in this city has increased by 300% since 1980.
- There is a strong correlation between church attendance and divorce: the more frequent the church attendance, the lower the divorce rate.

How would you evaluate these claims? The truth is elusive—how can we know it when we see it? The same skills that help you read the professional research literature will also help you sort out the statistics you encounter in the media and in casual conversation. These boxed inserts will help you develop a more critical and informed approach to all statistics.

Statistical literacy will not always lead you to the truth, of course, but it will enhance your ability to analyze and evaluate information, to sort through claims and counterclaims, and to appraise them sensibly.

ONE STEP AT A TIME Determining the Level of Measurement of a Variable

Step	***Operation***
1.	Inspect the scores of the variable *as they are actually stated*, keeping in mind the definition of the three levels of measurement (see Table 1.5).
2.	Change the order of the scores. Do they still make sense? If the answer is *yes*, the variable is **nominal**. If the answer is *no,* proceed to step 3.

(*continued*)

Illustration: Gender is a nominal variable, and its scores can be stated in any order:

1. Male
2. Female

or

1. Female
2. Male

The two statements of the scores are equally sensible. On a nominal-level variable, no score is higher or lower than any other score and the order in which they are stated is arbitrary.

3. Is the distance between the scores unequal or undefined? If the answer is *yes*, the variable is **ordinal**. If the answer is *no*, proceed to step 4.

Illustration: Consider the following scale, which measures support for capital punishment.

1. Strongly support
2. Somewhat support
3. Neither support or oppose
4. Somewhat oppose
5. Strongly oppose

People who "strongly" support the death penalty are more in favor than people who "somewhat" support it, but the distance from one level of support to the next (from a score of 1 to a score of 2) is undefined. We do not have enough information to ascertain how much more or less one score is than another.

4. If you answered *no* in steps 2 and 3, the variable is **interval-ratio**. Variables at this level have scores that are actual numbers: They have an order with respect to each other and are a defined, equal distance apart. They also have a true zero point. Examples of interval-ratio variables include age, income, and number of siblings.

Source: This system for determining level of measurement was suggested by Michael G. Bisciglia, Louisiana State University.

Some Final Points

Table 1.5 summarizes the discussion of the three levels of measurement. Note that the number of permitted mathematical operations increases as we move from nominal to ordinal to interval-ratio levels of measurement. Ordinal-level variables are more sophisticated and flexible than nominal-level variables, and interval-ratio-level variables permit the broadest range of mathematical operations.

Let us end this section by making three points. The first stresses the importance of level of measurement, and the other two discuss some common points of confusion in applying this concept.

1. Knowing the level of measurement of a variable is crucial because it tells us which statistics are appropriate and useful. Not all statistics can be used with all variables. As shown in Table 1.5, each level of measurement allows different mathematical operations and, thus, different statistics. For example, computing an average requires addition and division, and finding a median

TABLE 1.5 Basic Characteristics of the Three Levels of Measurement

Levels	Examples	Measurement Procedures	Mathematical Operations Permitted
Nominal	Gender, race, religion, marital status	Classification into categories	Counting number in each category, comparing sizes of categories
Ordinal	Social class, attitude and opinion scales	Classification into categories plus ranking of categories with respect to each other	All of the above plus judgments of "greater than" and "less than"
Interval-Ratio	Age, number of children, income	All of the above plus description of scores in terms of equal units	All of the above plus all other mathematical operations (addition, subtraction, multiplication, division, square roots, etc.)

(or middle score) requires that the scores be ranked from high to low. Addition and division are appropriate only for interval-ratio-level variables, and ranking is possible only for variables that are at least ordinal in level of measurement. Your first step in dealing with a variable and selecting appropriate statistics is *always* to determine its level of measurement.

2. In determining level of measurement, always examine the way in which the scores of the variable are *actually stated*. This is particularly a problem with variables that are interval-ratio but that have been measured at the ordinal level.

To illustrate, consider income as a variable. If we asked respondents to list their exact income in dollars, we would generate scores that are interval-ratio in level of measurement. Measured in this way, the variable would have a true zero point (no income at all) and equal intervals from score to score (1 dollar). It is more convenient for respondents, however, simply to check the appropriate category from a list of income ranges, as in Table 1.6. The four scores or categories in Table 1.6 are ordinal in level of measurement because they are unequal in size and have no true zero point. It is common for researchers to sacrifice precision (income in actual dollars) for the convenience of the respondents. You should be careful to look at the way in which the variable is measured before making a decision about its level of measurement.

TABLE 1.6 Measuring Income at the Ordinal Level

Score	Income Range
1	Less than $24,999
2	$25,000 to $49,999
3	$50,000 to $99,999
4	$100,000 or more

3. There is a mismatch between the variables that are usually of most interest to social scientists (race, gender, marital status, attitudes, and opinions) and the most powerful and interesting statistics (such as the mean). The former are typically nominal or ordinal in level of measurement, but the more sophisticated statistics require measurement at the interval-ratio level.

   This mismatch creates some very real difficulties for social science researchers. On one hand, researchers will want to measure variables at the highest, most precise level of measurement possible. If income is measured in exact dollars, for example, researchers can make very precise descriptive statements about the differences between people: "Ms. Smith earns $12,547 more than Mr. Jones." If the same variable is measured in broad, unequal categories, such as those in Table 1.6, comparisons between individuals would be less precise and provide less information: "Ms. Smith earns more than Mr. Jones."

   On the other hand, given the nature of the disparity, researchers are more likely to treat variables as if they were higher in level of measurement than they actually are. In particular, variables measured at the ordinal level, especially when they have many possible categories or scores, are often treated as if they were interval-ratio and are analyzed with the more powerful, flexible, and interesting statistics available at the higher level. This practice is common, but researchers should be cautious in assessing statistical results and developing interpretations when the level of measurement criterion has been violated.

In conclusion, level of measurement is a very basic characteristic of a variable, and we will always consider it when presenting statistical procedures. Level of measurement is also a major organizing principle for the material that follows, and you should make sure that you are familiar with these guidelines. *(For practice in determining the level of measurement of a variable, see problems 1.4 through 1.9.)*

SUMMARY

1. The purpose of statistics is to organize, manipulate, and analyze data so that researchers can test theories and answer questions. Along with theory and methodology, statistics are a basic tool used by social scientists to enhance their understanding of the social world.
2. There are two general classes of statistics. Descriptive statistics are used to summarize the distribution of a single variable and the relationships between two or more variables. We use inferential statistics to generalize to populations from random samples.
3. Variables may be measured at any of three different levels. At the nominal level, we can compare category sizes. At the ordinal level, scores can be ranked from high to low. At the interval-ratio level, all mathematical operations are permitted.

GLOSSARY

Data. Information expressed as numbers.

Data reduction. Summarizing many scores with a few statistics.

Dependent variable. A variable that is identified as an effect or outcome. The dependent variable is thought to be caused by the independent variable.

Descriptive statistics. The branch of statistics concerned with (1) summarizing the distribution of a single variable or (2) measuring the relationship between two or more variables.

Hypothesis. A specific statement, derived from a theory, about the relationship between variables.

Independent variable. A variable that is identified as a cause. The independent variable is thought to cause the dependent variable.

Inferential statistics. The branch of statistics concerned with making generalizations from samples to populations.

Level of measurement. The mathematical characteristic of a variable and the major criterion for selecting statistical techniques. Variables can be measured at any of three levels, each permitting certain mathematical operations and statistical techniques. The characteristics of the three levels are summarized in Table 1.5.

Measures of association. Statistics that summarize the strength and direction of the relationship between variables.

Population. The total collection of all cases in which the researcher is interested.

Quantitative research. Research projects that collect data or information in the form of numbers.

Research. Any process of gathering information systematically and carefully to answer questions or test theories. Statistics are useful for research projects that collect numerical information or data.

Sample. A carefully chosen subset of a population. In inferential statistics, information is gathered from a sample and then generalized to a population.

Statistics. A set of mathematical techniques for organizing and analyzing data.

Theory. A generalized explanation of the relationship between two or more variables.

Variable. Any trait that can change values from case to case.

PROBLEMS

1.1 In your own words, describe the role of statistics in the research process. Using the "wheel of science" as a framework, explain how statistics link theory with research.

1.2 Find a research article in any social science journal. Choose an article on a subject of interest to you, and don't worry about being able to understand all of the statistics that are reported.

- **a.** How much of the article is devoted to statistics?
- **b.** Is the research based on a sample from some population? How large is the sample? How were subjects or cases selected? Can the findings be generalized to some population?
- **c.** What variables are used? Which are independent and which are dependent? Determine the level of measurement of each variable.
- **d.** What statistical techniques are used? Try to follow the statistical analysis and see how much you can understand. Save the article and read it again after you finish this course and see if you do any better.

1.3 Distinguish between descriptive and inferential statistics. Describe a research situation that would use each type.

1.4 For each of the following items from a public-opinion survey, indicate the level of measurement.

- **a.** In what country were you born? ____________
- **b.** What is your age? ____________
- **c.** How many years of school have you completed? ____________
- **d.** What is your occupation? ____________
- **e.** If you were asked to use one of these four names for your social class, in which would you say you belonged?
  ____________ Upper ____________ Middle
  ____________ Working ____________ Lower
- **f.** What is your grade-point average? ____________
- **g.** What is your major? ____________
- **h.** The only way to deal with the drug problem is to legalize all drugs.
  ____________ Strongly agree
  ____________ Agree
  ____________ Undecided
  ____________ Disagree
  ____________ Strongly disagree
- **i.** What is your astrological sign? ____________
- **j.** How many brothers and sisters do you have? ____________

1.5 Following are brief descriptions of how researchers measured a variable. For each situation, determine the level of measurement of the variable.

a. Race or ethnicity. Check all that apply:
__________ Black
__________ White
__________ Hispanic
__________ Asian or Pacific Islander
__________ Native American
__________ Other (Please specify: ________)

b. Honesty. Subjects were observed as they passed by a spot where an apparently lost wallet was lying. The wallet contained money and complete identification. Subjects were classified into one of the following categories:
__________ Returned the wallet with money
__________ Returned the wallet but kept the money
__________ Did not return the wallet

c. Social class. What was your family situation when you were 16 years old?
__________ Very well off compared to other families
__________ About average
__________ Not so well off

d. Education. How many years of schooling have you completed? __________

e. Racial integration on campus. Students were observed during lunchtime at the cafeteria for a month. The number of students sitting with students of other races was counted for each meal period.

f. Number of children. How many children have you ever had? __________

g. Student seating patterns in classrooms. On the first day of class, instructors noted where each student sat. Seating patterns were remeasured every two weeks until the end of the semester. Each student was classified as:
__________ Same seat as at last measurement
__________ Adjacent seat
__________ Different seat, not adjacent
__________ Absent

h. Physicians per capita. The number of physicians was counted in each of 50 cities. The researchers used population data to compute the number of physicians per capita.

i. Physical attractiveness. A panel of ten judges rated each of 50 photos of a mixed-race sample of males and females for physical attractiveness on a scale from 0 to 20, with 20 being the highest score.

j. Number of accidents. The number of traffic accidents per year for each of 20 intersections was recorded. Also, each accident was rated as:
__________ Minor damage, no injuries
__________ Moderate damage, personal injury requiring hospitalization
__________ Severe damage and injury

1.6 For each research project listed here, identify the variables and classify them in terms of level of measurement and whether they are independent or dependent.

a. For a research project in a political science course, a student collected information for 50 nations. She used infant mortality rates (the number of infant deaths per 100,000 population) as a measure of quality of life and the percentage of all adults who are permitted to vote as a measure of democratization. Her hypothesis was that quality of life is higher in more democratic nations.

b. A highway engineer wonders whether increasing the speed limit on a heavily traveled highway will result in more accidents. He plans to collect information on traffic volume, number of accidents, and number of fatalities for the six-month periods before and after the speed limit is changed.

c. Students are planning a program to promote "safe sex" and awareness of other health concerns on campus. To measure the effectiveness of the program, they plan to survey students about their knowledge of safe sex practices before and after the program.

d. A graduate student asks 500 female students whether they have experienced any sexual harassment on campus. Each student is asked to estimate the frequency of these incidents as either "often, sometimes, rarely, or never." The researcher also gathers data on age and major to see whether there is any connection between these variables and frequency of sexual harassment.

e. A supervisor in the solid waste management division of city government is assessing two different methods of trash collection. One area of the city is served by trucks with two-man crews who do "backyard" pickups, and the rest of the city is served by "high-tech" single-person trucks with curbside pickup. The assessment measures include the number of complaints received from the two different areas over a six-month period, the amount of time per day required to service each area, and the cost per ton of trash collected.

f. Does tolerance for diversity vary by race or ethnicity? Samples of white, black, Asian, Hispanic, and Native Americans have been given a survey that measures their interest in and appreciation of cultures other than their own. Degree of tolerance is scored as "high, moderate, or low."

g. States have drastically cut their budgets for mental health care. Will this increase the number of homeless people? A researcher contacts a number of agencies serving the homeless in each state and develops an estimate of the size of the homeless population before and after the cuts.

h. The adult bookstore near campus has been raided and closed by the police. Your social research class has decided to poll a sample of students to find out whether he or she supports the store's closing, how many times each has visited the store, and whether he or she agrees that "pornography causes sexual assaults on women." The class also collects information on the gender, political philosophy, and major of the students to see if these characteristics affect opinions.

1.7 For each of the following research situations, identify the level of measurement of all variables. Also, decide which statistical applications are used: descriptive statistics (single variable), descriptive statistics (two or more variables), or inferential statistics. Remember that it is quite common for a given situation to require more than one type of application.

a. The administration of your university is proposing a change in parking policy. You select a random sample of students and ask each one how strongly they favor or oppose the change. You then use the results to estimate the support for the change in the entire student body.

b. You ask everyone in your social research class for their highest letter grade in any math course and their grade (percentage of items correct) on a recent statistics test. You compare the two sets of scores to see whether there is a relationship.

c. Your aunt is running for mayor and hires you to question a random sample of voters about their concerns. Specifically, she wants to use this information to characterize the entire city in terms of political party affiliation, gender, and what percentage favor widening of the main street in town.

d. Several years ago, a state reinstituted the death penalty for first-degree homicide. Did this reduce the homicide rate? A researcher has gathered information on the number of homicides in the state for the two-year periods before and after the change.

e. A local automobile dealer is concerned about customer satisfaction. He mails a survey form to all customers for the past year and asks whether they are satisfied, very satisfied, or not satisfied with their purchases.

1.8 For each of the first 20 items in the General Social Survey (see Appendix G), indicate the level of measurement.

1.9 Identify all variables in the research situations listed below and classify them by their level of measurement. Which variables are independent and which are dependent?

a. A researcher is wondering about racial preferences in dating among college students and asks a large sample of undergraduates about their own racial self-identification. She then asks the respondents to rank some racial-ethnic categories (stated as white, black, Latino, Asian) in terms of desirability as potential dates.

b. For high school students, does GPA affect sexual activity? A sample has been interviewed about the number of different romantic relationships they have had, the number of times they have had sexual intercourse, and their high school grade-point average.

c. Several hundred voting precincts across the nation have been classified in terms of percentage of minority voters, voting turnout, and percentage of local elected officials who are members of minority groups. Do precincts with higher percentages of minority voters have lower turnout? Do precincts with higher percentages of minority elected officials have higher turnout?

d. As nations become more affluent (as measured by per capita income), does the percentage of children enrolled in school increase? Is this relationship different for boys and girls?

e. Does the level of support for gun control vary by number of years of schooling? Does this relationship vary by gender, region of the country, or political party preference? Support for gun control was measured by a five-point scale that ranged from "strongly in favor" to "strongly opposed."

YOU ARE THE RESEARCHER

Introduction

The best way—maybe the only way—to learn statistics and to appreciate their importance is to apply and use them. This includes selecting the correct statistic for a given purpose, doing the calculations, and interpreting the result. I have included extensive end-of-chapter problems to give you multiple opportunities to select and calculate statistics and say what they mean. Most of these problems have been written so that they can be solved with just a simple hand calculator. I've purposely kept the number of cases unrealistically low so that the tedium of lengthy calculations would not interfere with the learning process. These problems present an important and useful opportunity to develop your statistical skills.

As important as they are, these end-of-chapter problems are simplified and several steps removed from the complex realities of social science research. To provide a more realistic statistical experience, I have included a feature called "You Are the Researcher," in which you will walk through many of the steps of a research project, making decisions about how to apply your growing knowledge of research and statistics and interpreting the statistical output you generate.

To conduct these research projects, you will analyze a shortened version of the 2012 General Social Survey (GSS). Please visit **www.cengagebrain.com** to download a copy of this data base. The GSS is a public-opinion poll that has been conducted on nationally representative samples of citizens of the United States since 1972. The full survey includes hundreds of questions covering a broad range of social and political issues. The version supplied with this text has a limited number of variables and cases but is still actual, "real-life" data, so you have the opportunity to practice your statistical skills in a more realistic context.

Even though the version of the GSS that we use for this text is shortened, it is still a large data set, with almost 1500 respondents and 49 variables, too large for even the most advanced hand calculator. To analyze the GSS, you will learn how to use a computerized statistical package called the Statistical Package for the Social Sciences (SPSS). A statistical package is a set of computer programs designed to analyze data. The advantage of these packages is that, because the programs are already written, you can capitalize on the power of the computer with minimal computer literacy and virtually no programming experience. Be sure to read Appendix F before attempting any data analysis.

In most of the research exercises, which begin in Chapter 2, you will make the same kinds of decisions as do professional researchers and move through some of the steps of a research project. You will select variables and statistics, generate and analyze output, and express your conclusions. When you finish these exercises, you will be well prepared to conduct your own research project (within limits, of course) and perhaps make a contribution to the ever-growing social science research literature.

Part I Descriptive Statistics

Part I consists of four chapters, each devoted to a different application of univariate descriptive statistics. Chapter 2 covers "basic" descriptive statistics, including frequency distributions, percentages, ratios, rates, and graphs. This material is relatively elementary and at least vaguely familiar to most people. Although these statistics are "basic," they are not necessarily simple or obvious, and the explanations and examples should be considered carefully before attempting the end-of-chapter problems or using them in actual research.

Chapters 3 and 4 cover measures of central tendency and dispersion, respectively. Measures of central tendency describe the typical case or average score (e.g., the mean), while measures of dispersion describe the amount of variety or diversity among the scores (e.g., the range, or the distance from the high score to the low score). These two types of statistics are presented in separate chapters to stress the point that centrality and dispersion are independent, separate characteristics of a variable. You should realize, however, that both measures are necessary and commonly reported together, along with some of the statistics presented in Chapter 2. To reinforce this idea, many of the problems at the end of Chapter 4 require the computation of a measure of central tendency from Chapter 3.

Chapter 5 is a pivotal chapter in the flow of the text. It takes some of the statistics from Chapters 2 through 4 and applies them to the normal curve, a concept of great importance in statistics. The normal curve is a line chart (see Chapter 2) that can be used to describe the position of scores using means (Chapter 3) and standard deviations (Chapter 4). Chapter 5 also uses proportions and percentages (Chapter 2).

In addition to its role in descriptive statistics, the normal curve is a central concept in inferential statistics, the topic of Part II of this text. Thus, Chapter 5 ends the presentation of univariate descriptive statistics and lays essential groundwork for the material to come.

2 Basic Descriptive Statistics: Tables, Percentages, Ratios and Rates, and Graphs

LEARNING OBJECTIVES

By the end of this chapter, you will be able to:

1. Explain how descriptive statistics can make data understandable.
2. Construct and analyze frequency distributions for variables at each of the three levels of measurement.
3. Compute and interpret percentages, proportions, ratios, rates, and percentage change.
4. Analyze bar and pie charts, histograms, and line graphs.
5. Use SPSS to generate and analyze frequency distributions and graphs.

USING STATISTICS

The statistical techniques presented in this chapter are used to summarize the scores on a single variable. They can be used to:

- Organize information into easy-to-read tables, charts, and graphs.
- Express the percentage of people in a community that belong to various religions, including people with no religious preference and atheists.
- Express the structure of opinion on controversial issues (e.g., the strength of opposition to cohabitation or support for gun control) and track changes over time.
- Report changes in the crime rate from year to year.

Research results do not speak for themselves. Researchers use statistics to organize and manipulate data so that their meaning can be understood by their readers. The purpose of descriptive statistics is to express research findings clearly and effectively.

In this chapter, we consider some commonly used techniques for presenting research results, including tables, percentages, rates, and graphs. These univariate descriptive statistics are not mathematically complex (although they are not as simple as they might seem at first glance), but they can be extremely useful tools for organizing and analyzing results and communicating conclusions.

Very often, the first step in a quantitative research project is to examine the variables and see how scores are distributed. One of the most useful ways to do this is to construct tables, or **frequency distributions**, that report the number of cases in each category, for all variables. These tables can be used with variables at any level of measurement. We will begin with nominal-level variables and consider not only frequency distributions but also some statistics that can increase the clarity of results.

Frequency Distributions for Nominal-Level Variables

Constructing frequency distributions for nominal-level variables is typically very straightforward. Count the number of times each category or score of the variable occurs and display the frequencies in table format. Table 2.1, for example, displays the distribution of gender for 113 respondents.

Note that the table has a title and clearly labeled categories, and it reports the total number of cases (*N*) at the bottom of the frequency column. These items must be included in *all* frequency distributions.

For some tables, the researcher might have to make some choices about the number of categories to be included. For example, recall Table 1.3, which displayed several different ways of measuring religious affiliation in North America. As we noted in Chapter 1, it is probably most common to use five categories (see Scale D of Table 1.3) to display this variable, but what if a researcher wanted to emphasize the diversity of religions? We could do this by expanding the "Other" category and including more religions. But where do we stop? There are hundreds—maybe thousands—of different religions in the United States. Increasing the number of categories will present a more accurate picture of religious diversity, but clarity and ease of communication may suffer.

Tables 2.2 and 2.3 illustrate the choices that sometimes have to be made. Table 2.2 uses the "standard" five categories to display religious membership in the United States today. Table 2.3 presents more detail by adding the three largest "Other" religions. The number of categories would be further increased if we included more "Other" religions or if we subdivided the Protestant category into the various denominations (Methodist, Lutheran, Episcopalian, and so forth).

At what point do we have "enough" detail? When does the picture presented by the table become too cluttered, too complex, and unclear? These are questions that must be answered in the context of the purpose of the research project. If you wanted to stress the numerical predominance of Protestants in the United States, Table 2.2 might be preferred. On the other hand, if you wanted to stress the diversity of religious affiliations, Table 2.3 would be preferable. There are no hard-and-fast rules, and the choice between greater detail (more categories) and more clarity (fewer categories) may be confronted with variables at all three levels of measurement.

TABLE 2.1 Gender (fictitious data)

Gender	Frequency
Males	53
Females	60
	$N = 113$

TABLE 2.2 Self-Described Religious Identifications of Adult Americans, 2008

Religious Group	Frequency
Protestant	116,203,000
Catholic	57,199,000
Jewish	2,680,000
Other	6,116,000
None	34,169,000
	N = 216,367,000

Source: U.S. Bureau of the Census. 2012. *Statistical Abstract of the United States: 2012.* p. 61. Retrieved from http://www.census.gov/prod/2011pubs/12statab/pop.pdf

TABLE 2.3 Self-Described Religious Identifications of Adult Americans, 2008

Religious Group	Frequency
Protestant	116,203,000
Catholic	57,199,000
Jewish	2,680,000
Muslim	1,349,000
Buddhist	1,189,000
Unitarian	586,000
Other	2,992,000
None	34,169,000
	N = 216,367,000

Source: U.S. Bureau of the Census. 2012. *Statistical Abstract of the United States: 2012.* p. 61. Retrieved from http://www.census.gov /prod/2011pubs/12statab/pop.pdf

Enhancing Clarity: Percentages and Proportions

It is likely that you have noted a difficulty with Tables 2.2 and 2.3: The frequencies involved are large and, therefore, may be difficult to comprehend. For example, imagine yourself trying to explain these tables in terms of the frequencies: "116,203,000 of 216,367,000 adult Americans described themselves as Protestant." There is nothing wrong with this statement—it is literally true—but the same point could have been more clearly conveyed using a **percentage**: "About 54% of adult Americans describe themselves as Protestant."

Percentages are extremely useful statistics because they supply a frame of reference by standardizing the raw frequencies to the base 100. The mathematical definition of a percentage is

FORMULA 2.1

$$\text{Percentage: } \% = \left(\frac{f}{N}\right) \times 100$$

Where f = frequency, or the number of cases in a specific category
N = the number of cases in all categories

To illustrate the computation of percentages, consider the data presented in Table 2.1. How can we find the percentage of males in the sample? Note that there are 53 males ($f = 53$) and a total of 113 cases ($N = 113$). So

$$\% = \left(\frac{f}{N}\right) \times 100 = \left(\frac{53}{113}\right) \times 100 = (0.4690) \times 100 = 46.90\%$$

Using the same procedure, we can find the percentage of females:

$$\% = \left(\frac{f}{N}\right) \times 100 = \left(\frac{60}{113}\right) \times 100 = (0.5310) \times 100 = 53.10\%$$

Percentages are easier to read and comprehend than raw frequencies, and a column for percentages is commonly added to frequency distributions for variables at all levels of measurement. For example, Table 2.4 reports the same information as Table 2.3, but the percentages make it much easier to read.

The advantage of using percentages is particularly obvious when we want to compare groups of different sizes. For example, consider the information presented in Table 2.5. Which college has the higher *relative* number of social science majors? Because the total enrollments are so different, comparisons are difficult to make from the raw frequencies. Percentages eliminate the difference in size of the two campuses by standardizing both distributions to the base of 100. The same data are presented in Table 2.6 with percentages.

TABLE 2.4 Self-Described Religious Identifications of Adult Americans, 2008

Religious Group	Frequency	Percentages
Protestant	116,203,000	53.71%
Catholic	57,199,000	26.44%
Jewish	2,680,000	1.24%
Muslim	1,349,000	0.62%
Buddhist	1,189,000	0.55%
Unitarian	586,000	0.27%
Other	2,992,000	1.38%
None	34,169,000	15.79%
	N = 216,367,000	100.00%

Source: U.S. Bureau of the Census. 2012. *Statistical Abstract of the United States: 2012*. p. 61. Retrieved from http://www.census.gov/prod/2011pubs/12statab/pop.pdf

TABLE 2.5 Declared Major Fields on Two College Campuses (fictitious data)

Major	College A	College B
Business	103	3120
Natural sciences	82	2799
Social sciences	137	1884
Humanities	93	2176
	N = 415	N = 9979

TABLE 2.6 Declared Major Fields on Two College Campuses (fictitious data)

Major	College A	College B
Business	24.82%	31.27%
Natural sciences	19.76%	28.05%
Social sciences	33.01%	18.88%
Humanities	22.41%	21.81%
	100.00%	100.01%
	(415)	(9979)

Table 2.6 makes it easier to identify differences as well as similarities. College A has a higher percentage of social science majors than College B (even though the absolute number of social science majors is less than at College B) and about the same percentage of humanities majors. How would you describe the differences in the remaining two major fields?

Social scientists use **proportions** as well as percentages. Proportions vary from 0.00 to 1.00: They standardize results to a base of 1.00 instead of to the base of 100 used for percentages. A proportion is the same as a percentage except that we do *not* multiply by 100.

FORMULA 2.2

$$\text{Proportion} = \left(\frac{f}{N}\right)$$

Percentages can be converted to proportions by dividing by 100 and, conversely, proportions can be converted to percentages by multiplying by 100. The two statistics are equivalent expressions of the same message and are interchangeable. For example, we could state the relative number of males in Table 2.1 as a proportion:

$$\text{Proportion} = \left(\frac{f}{N}\right) = \left(\frac{53}{113}\right) = 0.47$$

How can we choose between these statistics? Percentages are easier for most people (including statisticians) to comprehend and generally would be preferred. Proportions are used less frequently, generally when we are working with probabilities (see Chapter 5). The preference for percentages is based solely on ease of communication: Proportions and percentages are equally valid ways of expressing results. Be sure to familiarize yourself with the guidelines stated in Applying Statistics 2.1. *(For practice in computing and interpreting percentages and proportions, see problems 2.1 and 2.2.)*

ONE STEP AT A TIME **Finding Percentages and Proportions**

Step	***Operation***
1.	Find the values for f (number of cases in a category) and N (number of cases in all categories). Remember that f will be the number of cases in a *specific category* (e.g., males on your campus) and N will be the number of cases in *all* categories (e.g., all students, males and females, on your campus) and that f will be smaller than N, except when the category and the entire group are the same (e.g., when all students are male). Proportions cannot exceed 1.00, and percentages cannot exceed 100%.
2.	For a proportion, divide f by N.
3.	For a percentage, multiply the value you calculated in step 2 by 100.

Applying Statistics 2.1 Using Percentages and Proportions

These statistics are relatively simple, but there are several guidelines that should be observed when using them.

1. When the number of cases (*N*) is small (say, fewer than 20), it is usually preferable to report the actual frequencies rather than percentages or proportions. With a small number of cases, the percentages are unstable and can change drastically with relatively minor changes. For example, if you begin with 10 males and 10 females (that is, 50% of each gender) and then add another female, the percentage distributions will change to 52.4% female and 47.6% male. Of course, as the number of observations increases, each additional case will have a smaller impact. If we started with 500 males and 500 females and then added one more female, the percentage of females would change by less than a tenth of a percent (from 50% to 50.05%).
2. *Always* report the number of observations along with proportions and percentages. This permits the reader to judge the adequacy of the sample size and, conversely, helps to prevent the researcher from lying with statistics. Statements like "Two out of three people prefer courses in statistics to any other course" might sound impressive, but the claim would lose its gloss if you learned that only three people were tested. *You should be extremely suspicious of reports that fail to mention the number of cases.*
3. Percentages and proportions can be calculated for variables at the ordinal and nominal levels of measurement, in spite of the fact that they require division. This is not a violation of the level-of-measurement guideline (see Table 1.5). Percentages and proportions do not require the division of the *scores* of the variable (as would be the case in computing the average score on a test, for example). Rather, the *number of cases* in a particular category (f) is divided by the total number of cases in the sample (*N*). Saying that "53.10% of the sample is female" is equivalent to saying that "60 of the 113 respondents are female," but in a more convenient and understandable way.

Applying Statistics 2.2 Communicating with Statistics

Not long ago, in a large social service agency, the following conversation took place between the executive director of the agency and a supervisor of one of the divisions.

> *Executive director:* Well, I don't want to seem abrupt, but I've only got a few minutes. Tell me, as briefly as you can, about this staffing problem you claim to be having.
> *Supervisor:* Ma'am, we just don't have enough people to handle our workload. Of the 177 full-time employees of the agency, only 50 are in my division. Yet 6231 of the 16,722 cases handled by the agency last year were handled by my division.
> *Executive director (smothering a yawn)*: Very interesting. I'll certainly get back to you on this matter.

How could the supervisor have presented his case more effectively? Because he wants to compare two sets of numbers (his staff versus the total staff and the workload of his division versus the total workload of the agency), proportions or percentages would be a more forceful way of presenting results. What if the supervisor had said, "Only 28.25% of the staff is assigned to my division, but we handle 37.26% of the total workload of the agency"? Is this a clearer message?

The first percentage is found by

$$\% = \left(\frac{f}{N}\right) \times 100 = \left(\frac{50}{177}\right) \times 100$$
$$= (0.2825) \times 100 = 28.25\%$$

and the second percentage is found by

$$\% = \left(\frac{f}{N}\right) \times 100 = \left(\frac{6231}{16{,}722}\right) \times 100$$
$$= (0.3726) \times 100 = 37.26\%$$

STATISTICS
IN EVERYDAY LIFE

College Majors

Tables 2.5 and 2.6 present fictional data on major fields of study of college students. What does this variable look like nationally? How has it changed over the years? The accompanying table presents the six most common majors, along with an "other" category.

Bachelor's Degrees Awarded in the United States by Major, 2011–2012

Major	Percentage of All Degrees
Business	20%
Humanities	17%
Social Science and Behavioral Sciences	16%
Natural Sciences and Mathematics	8%
Computer Science and Engineering	8%
Education	6%
Other	25%
Total	100%

The number of B.A. degrees has more than doubled since 1970. Among many other changes, the percentage of business majors has increased from about 14% to 20%, the percentage of education majors has declined precipitously, and the percentage of "other" majors has ballooned from less than 10% to 25%.

Source: National Center for Education Statistics, retrieved from https://nces.ed.gov/programs/digest/d13/tables/dt13_318.20.asp

Frequency Distributions for Ordinal-Level Variables

Frequency distributions for ordinal-level variables are constructed in the same way as for nominal-level variables. Table 2.7 reports the frequency distribution of responses of students to a survey item measuring support for birth control on the campus. Note that a percentage column has been added to the table to increase clarity.

TABLE 2.7 Support for Birth Control on a University Campus (fictitious data)

Do you strongly agree, agree, disagree, or strongly disagree that the University Health Center should provide condoms and other "safe sex" items on demand and at no additional cost to students?

Response	Frequency	Percentage
Strongly agree	350	25.55%
Agree	462	33.72%
Disagree	348	25.40%
Strongly disagree	210	15.33%
	1370	100.00%

TABLE 2.8 Support for Birth Control on a University Campus with Categories Collapsed (fictitious data)

Do you strongly agree, agree, disagree, or strongly disagree that the University Health Center should provide condoms and other "safe sex" items on demand and at no additional cost to students?

Response	Frequency	Percentage
Strongly agree or Agree	812	59.27%
Disagree or Strongly disagree	558	40.73%
	1370	100.00%

The table reports that opinions were fairly evenly distributed. The single most popular response was "Agree" (33.72%), and a majority of students (59.27%) agreed or strongly agreed that condoms and other "safe sex" devices should be available. If the researcher wanted to emphasize this pattern or make the table more compact, categories could be collapsed as in Table 2.8. However, there is a price to pay for the simpler, more compact expression of results: The exact breakdown of degrees of agreement and disagreement is lost. As we saw when discussing the display of religious affiliations in Tables 2.2 and 2.3, the researcher must strike a balance between more detail (more categories) and greater clarity (fewer categories). *(For practice in constructing and interpreting frequency distributions for nominal- and ordinal-level variables, see problems 2.5 and 2.12.)*

Frequency Distributions for Interval-Ratio-Level Variables

In general, the construction of frequency distributions for variables measured at the interval-ratio level is more complex than for nominal and ordinal variables. Interval-ratio variables usually have a wide range of scores, and this means that the researcher must collapse or group categories to produce reasonably compact tables. Once again, we see that the researcher needs to decide between more detail and more clarity.

For example, suppose you wished to report the distribution of the variable "age" for a community. In most communities, there would be a very broad range of ages, from newborns to people in their 90s or older. If you simply reported the number of times that each year of age (or score) occurred, you could wind up with a frequency distribution that had 80, 90, or even more categories; such a table would be very difficult to read. The scores (years) must be grouped into larger categories to increase ease of comprehension. How large should these categories be? How many categories should be included in the table? Should we provide more information (a greater number of narrow categories) or more clarity (a smaller number of wide categories)?

Constructing Frequency Distributions for Interval-Ratio-Level Variables

For ease of illustration, let's consider a frequency distribution for a small college class of 20 students. Because of the narrower age range of college students, we can

TABLE 2.9 Age of Students in a College Class (fictitious data)

Interval width set at 1 year of age	
Ages	Frequency
18	5
19	6
20	3
21	2
22	1
23	1
24	1
25	0
26	1
	$N = 20$

use categories of only one year (these categories are often called **class intervals** when working with interval-ratio data). The frequency distribution is constructed by listing the ages in order, counting the number of times each score (year of age) occurs, and then totaling the number of cases in each category. Table 2.9 presents the information and reveals a concentration of cases in the 18 and 19 class intervals.

Even though this table is fairly clear, assume that you desire a more compact (less detailed) summary. To obtain this, you will have to group scores into wider class intervals. Increasing the interval width (say, to two years) will reduce the number of intervals and produce a more compact expression. The grouping of scores in Table 2.10 clearly emphasizes the predominance of younger respondents. This trend can be stressed even more by adding a column for percentages.

Stated Limits

Note that the class intervals in Table 2.10 have been stated with an apparent gap between them. That is, the **stated class limits** are separated by a distance of one unit. At first glance, these gaps may appear to violate the principle of exhaustiveness introduced in Chapter 1 (see Table 1.2) but, because age has been measured in whole numbers, the "gaps" are actually not a problem. Given the level

TABLE 2.10 Age of Students in a College Class (fictitious data)

Interval width set at 2 years of age		
Age	Frequency	Percentage
18–19	11	55.0%
20–21	5	25.0%
22–23	2	10.0%
24–25	1	5.0%
26–27	1	5.0%
	$N = 20$	100.0%

of precision of the measurement (in whole years, as opposed to, say, tenths of a year), no case could have a score that falls between the intervals. The class intervals in Table 2.10 are exhaustive and mutually exclusive, and each of the 20 respondents can be sorted into one and only one age category.

However, consider the difficulties if age had been measured with greater precision. If age had been measured in tenths of a year, into which class interval in Table 2.10 would a 19.5-year-old subject be placed? We avoid this ambiguity by always stating the limits of the class intervals at the same level of precision as the data. Thus, if age were being measured in tenths of a year, the limits of the class intervals in Table 2.10 would be stated in tenths of a year. For example:

17.0–18.9
19.0–20.9
21.0–22.9
23.0–24.9
25.0–26.9

To maintain mutual exclusivity between categories, do not overlap the class intervals. If you state the limits of the class intervals at the same level of precision as the data (which might be in whole numbers, tenths, hundredths, etc.) and maintain a "gap" between intervals, you will always produce a frequency distribution where each case can be assigned to one and only one category.

Midpoints

On occasion—when constructing certain graphs, for example—you will need to work with the **midpoints** of the class intervals. Midpoints are exactly halfway between the upper and lower limits of a class interval and can be found by dividing the sum of the upper and lower limits by 2. To illustrate, Table 2.11 displays midpoints for two different sets of class intervals. See the "One Step at a Time" box for detailed instructions. *(For practice in finding midpoints, see problems 2.8b and 2.9b.)*

TABLE 2.11 Finding Midpoints

Class interval width set at 3	
Class Interval	Midpoint
0–2	1.0
3–5	4.0
6–8	7.0
9–11	10.0

Class interval width set at 6	
Class Interval	Midpoint
100–105	102.5
106–111	108.5
112–117	114.5
118–123	120.5

ONE STEP AT A TIME **Finding Midpoints**

Step	*Operation*
1.	Find the upper and lower limits of the lowest interval in the frequency distribution. For the top set of intervals in Table 2.11, the lowest interval (0–2) includes scores of 0, 1, and 2. The upper limit of this interval is 2 and the lower limit is 0.
2.	Add the upper and lower limits and divide by 2. For the interval 0−2: (0 + 2)/2 = 1. The midpoint for this interval is 1.
3.	Midpoints for other intervals can be found by repeating steps 1 and 2 for each interval. As an alternative, you can find the midpoint for any interval by adding the value of the interval width to the midpoint of the next lower interval. For example, the lowest interval in the top panel of Table 2.11 is 0–2 and the midpoint is 1. Intervals are 3 units wide (that is, they each include three scores), so the midpoint for the next higher interval (3−5) is 1 + 3, or 4. The midpoint for the interval 6–8 is 4 + 3, or 7, and so forth.

Cumulative Frequency and Cumulative Percentage

Two commonly used adjuncts to the basic frequency distribution for interval-ratio data are the **cumulative frequency** and **cumulative percentage** columns. These columns allow us to tell at a glance how many cases fall at or below a given score or class interval in the distribution. To add these columns, follow the instructions in the "One Step at a Time" box. Table 2.12 shows a cumulative frequency column and a cumulative percentage column added to Table 2.10.

These cumulative columns are quite useful in situations where the researcher wants to make a point about how cases are spread across the range of scores. For example, Table 2.12 shows clearly that the great majority of students in the class are younger than 21. If the researcher wishes to stress this fact, then these cumulative columns are quite handy.

Most realistic research situations will be concerned with many more than 20 cases and/or many more categories than our tables have, and the cumulative percentage column would usually be preferred to the cumulative frequencies column.

TABLE 2.12 Age of Students in a College Class

Age	Frequency	Cumulative Frequency	Percentage	Cumulative Percentage
18–19	11	11	55.0%	55.0%
20–21	5	16	25.0%	80.0%
22–23	2	18	10.0%	90.0%
24–25	1	19	5.0%	95.0%
26–27	1	20	5.0%	100.0%
	$N = 20$		100.0%	

ONE STEP AT A TIME **Adding Cumulative Frequency and Percentage Columns to Frequency Distributions**

Step ***Operation***

Cumulative Frequency Column

1. Begin with the lowest class interval (the interval with the lowest scores). The entry in the cumulative frequency column will be the same as the number of cases in this interval.
2. Go to the next class interval. The cumulative frequency for this interval is the number of cases in the interval plus the number of cases in the lower interval.
3. Continue adding (or accumulating) cases from interval to interval until you reach the interval with the highest scores, which will have a cumulative frequency equal to *N*.

Step ***Operation***

Cumulative Percentage Column

1. Compute the percentage of cases in each category and then follow the pattern for the cumulative frequencies. The entry for the lowest class interval will be the same as the percentage of cases in the interval.
2. For the next higher interval, the cumulative percentage is the percentage of cases in the interval plus the percentage of cases in the lower interval.
3. Continue adding (or accumulating) percentages from interval to interval until you reach the interval with the highest scores, which will have a cumulative percentage of 100%.

Unequal Class Intervals

As a general rule, the class intervals of frequency distributions should be equal in size in order to maximize clarity and ease of comprehension. For example, all of the class intervals in Table 2.12 are 2 years wide. There are several other possibilities for stating class intervals, and we will examine each situation separately.

Open-Ended Intervals. What would happen to the frequency distribution in Table 2.12 if we added one more student who was 47 years of age? We would now have 21 cases and there would be a large gap between the oldest respondent (now 47) and the second oldest (age 26). If we simply added the older student to Table 2.12, we would have to include nine new class intervals (28–29, 30–31, 32–33, and so forth) with zero cases in them before we got to the 46–47 interval. This would waste space and probably be unclear and confusing. An alternative way to handle a few very high (or low) scores would be to add an "open-ended" interval to the frequency distribution, as in Table 2.13.

The open-ended interval in Table 2.13 allows us to present the information compactly. We could handle an extremely low score by adding an open-ended interval as the lowest class interval (e.g., "17 and younger"). Of course, there is a price to pay for this efficiency—the table omits information about the exact scores included in the open-ended interval—so this technique should not be used indiscriminately.

Intervals of Unequal Size. On some variables, most scores are tightly clustered together but others are strewn across a broad range of scores. Consider the

TABLE 2.13 Age of Students in a College Class (*N* = 21)

Age	Frequency	Cumulative Frequency
18–19	11	11
20–21	5	16
22–23	2	18
24–25	1	19
26–27	1	20
28 and older	1	21
	N = 21	

distribution of income in the United States. In 2012, most households (about 54%) reported annual incomes between $25,000 and $100,000 and a sizeable grouping (24.4%) earned less than that. The problem (from a statistical point of view) comes with more affluent households, those with incomes above $100,000. There are fewer households in these categories (21.6%) but we must still account for them.

If we used equal intervals of, say, $10,000, in Table 2.14, we would need 30 or 40 or more intervals to include all of the more affluent households, and many of our intervals in the higher income ranges—especially those over $150,000—would have few or zero cases. The resulting table would be very large and hard to comprehend.

In situations such as this, we can use intervals of unequal size to summarize the variable more efficiently, as in Table 2.14. Some of the intervals in Table 2.14 are $10,000 wide, others are $25,000 or $50,000 wide, and two (the lowest and highest intervals) are open-ended. Tables that use intervals of mixed widths might be a little confusing for the reader, but the trade-off in compactness and efficiency can be considerable. *(For practice in constructing and interpreting frequency distributions for interval-ratio-level variables, see problems 2.5 to 2.9 and 2.12.)*

TABLE 2.14 Distribution of Income by Household, United States, 2012

Income	Percentage of Households	Cumulative Percentage
Less than $10,000	7.7%	7.7%
$10,000 to $14,999	5.6%	13.3%
$15,000 to $24,999	11.1%	24.4%
$25,000 to $34,999	10.4%	34.8%
$35,000 to $49,999	13.8%	48.6%
$50,000 to $74,999	18.0%	66.6%
$75,000 to $99,999	11.9%	78.5%
$100,000 to $149,999	12.4%	90.9%
$150,000 to $199,000	4.6%	95.5%
$200,000 and above	4.6%	100.1%
	100.1%	
	(*N* = 115,969,540)	

Source: U.S. Bureau of the Census. 2014. *American Community Survey, 2012*. Retrieved from http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_1YR_DP03&prodType=table

Summary: Frequency Distributions for Interval-Ratio-Level Variables

Remember that there are no absolute rules for constructing frequency distributions for interval-ratio variables. As long as you state the class intervals with a "gap" and have an interval for every score, you are free to make your own decisions about the best balance between detail (more categories) and clarity (fewer categories). The "One Step at a Time" box provides some general guidelines that may be helpful in constructing a frequency distribution by hand.

Also, it is likely that you have access to computers and software such as SPSS to help you construct these tables. The "Using SPSS" section presents instructions for using the software to generate a frequency distribution.

ONE STEP AT A TIME Constructing Frequency Distributions for Interval-Ratio Variables

Step	*Operation*
1.	Decide how many class intervals (k) you wish to use. One reasonable convention suggests that the number of intervals should be about 10 ($k = 10$). Many research situations may require fewer than 10 intervals, and it is common to find frequency distributions with as many as 15 or 20 intervals. Only rarely will more than 20 intervals be used, since the resultant frequency distribution would be too large for easy comprehension.
2.	Find the range (R) of the scores by subtracting the low score from the high score.
3.	Find the size of the class intervals (i) by dividing R (from step 2) by k (from step 1): $$i = R/k$$ Round the value of i to a convenient whole number. This will be the interval size or width.
4.	State the lowest interval so that its lower limit is equal to or below the lowest score. Your highest interval will be equal to or greater than the highest score. Generally, intervals should be equal in size, but unequal and open-ended intervals may be used when convenient.
5.	State the limits of the class intervals at the same level of precision as you have used to measure the data. Do not overlap intervals. You will thereby define the class intervals so that each case can be sorted into one and only one category.
6.	Count the number of cases in each class interval, and report these subtotals in a column labeled "Frequency." Report the total number of cases (N) at the bottom of this column. The table may also include a column for percentages, cumulative frequencies, and cumulative percentages.
7.	Inspect the frequency distribution carefully. Has too much detail been obscured? If so, reconstruct the table with a greater number of class intervals (or, use a smaller interval size). Is the table too detailed? If so, reconstruct the table with fewer class intervals (or, use wider intervals). Are there too many intervals with no cases in them? If so, consider using open-ended intervals or intervals of unequal size. Remember that the frequency distribution results from a number of decisions you make in a rather arbitrary manner. If the appearance of the table seems less than optimal given the purpose of the research, redo the table until you are satisfied that you have struck the best balance between detail and conciseness.
8.	Give your table a clear, concise title, and number the table if your report contains more than one. All categories and columns must also be clearly labeled.

Applying Statistics 2.3 Frequency Distributions

The following list shows the ages of 50 prisoners enrolled in a work-release program. Is this group young or old? A frequency distribution will provide an accurate picture of the overall age structure.

18	60	57	27	19
20	32	62	26	20
25	35	75	25	21
30	45	67	41	30
37	47	65	42	25
18	51	22	52	30
22	18	27	53	38
27	23	32	35	42
32	37	32	40	45
55	42	45	50	47

We will use about 10 intervals to display these data. By inspection we see that the youngest prisoner is 18 and the oldest is 75. The range (R) is thus 57. Interval size will be 57/10, or 5.7, which we can round off to either 5 or 6. Let's use a six-year interval beginning at 18. The limits of the lowest interval will be 18–23. Now we must state the limits of all the other intervals, count the number of cases in each interval, and display these counts in a frequency distribution.

Columns may be added for percentages, cumulative percentages, and/or cumulative frequency. The complete distribution, with a column added for percentages, is

Ages	Frequency	Percentages
18–23	10	20%
24–29	7	14%
30–35	9	18%
36–41	5	10%
42–47	8	16%
48–53	4	8%
54–59	2	4%
60–65	3	6%
66–71	1	2%
72–77	1	2%
	$N = 50$	100%

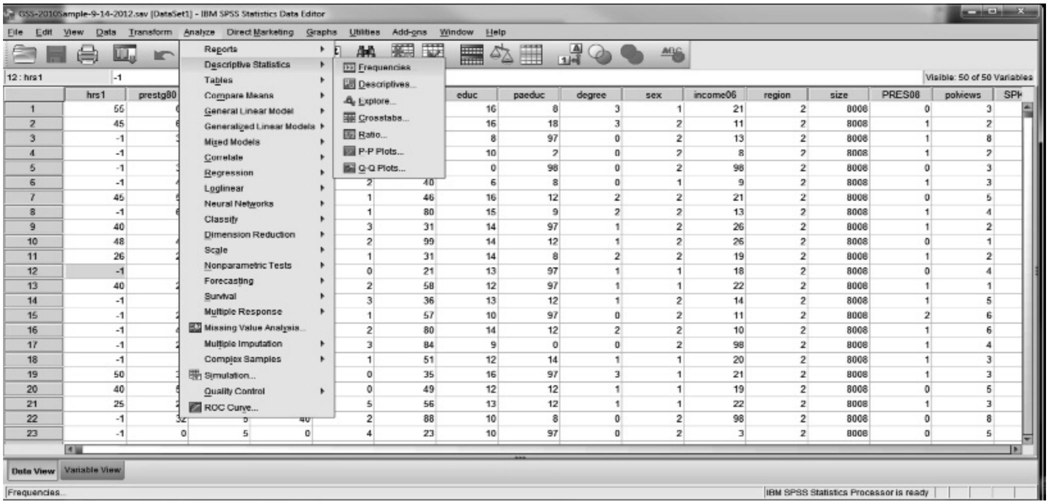
The prisoners are fairly evenly spread across the age groups 18 to 35. The frequencies then begin to decline and there is a noticeable lack of prisoners in the oldest age groups.

Using SPSS to Produce Frequency Distributions

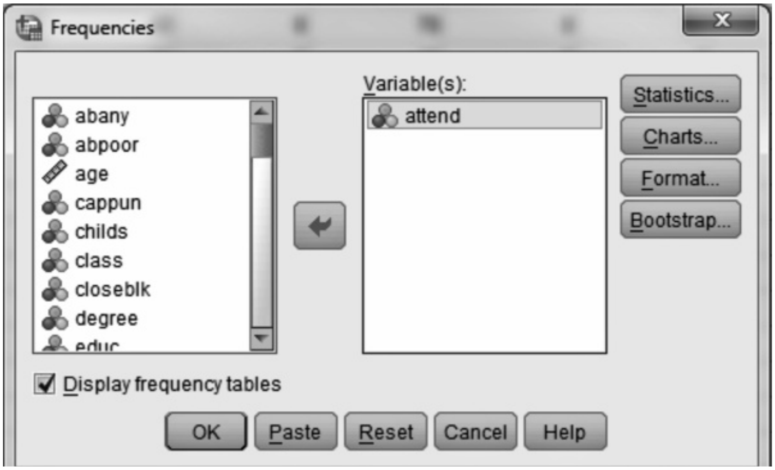
In this first installment of "Using SPSS," we will use the program to create a frequency distribution for a variable called *attend*, a measure of how often Americans attend religious services.

Follow these steps to get a frequency distribution:

1. Find and click the SPSS icon on your desktop.
2. Load the General Social Survey 2012 database.
   a. Find the **File** command on the far left of the menu bar and then click **File → Open → Data.**
   b. Find the 2012 General Social Survey (*GSS2012*) database file supplied with this text. You can download this file from the website for this text if you haven't already.
3. From the menu bar across the top of the SPSS window, click **Analyze**, **Descriptive Statistics**, and **Frequencies**.



4. Find the variable (*attend*) in the box on the left of the "Frequencies" window and click the arrow to move the variable name to the box on the right. What you see on the screen will be similar to this:



The window on the left may display variables by name (e.g., *abany*, *abpoor*) or by label (e.g., ABORTION IF WOMAN WANTS FOR ANY REASON). If labels are displayed, you may switch to variable names by clicking **Edit**, **Options**, and then making the appropriate selections on the "General" tab. See Appendix F and Table F.2 for further information.

5. Click **OK** and the frequency distribution will be sent to the "SPSS Output" window.

The output is as follows:

How Often R Attends Religious Services

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NEVER	364	25.0	25.1	25.1
	LT ONCE A YEAR	83	5.7	5.7	30.8
	ONCE A YEAR	194	13.3	13.4	44.2
	SEVRL TIMES A YR	135	9.3	9.3	53.5
	ONCE A MONTH	97	6.7	6.7	60.2
	2-3X A MONTH	126	8.6	8.7	68.8
	NRLY EVERY WEEK	55	3.8	3.8	72.6
	EVERY WEEK	289	19.8	19.9	92.6
	MORE THN ONCE WK	108	7.4	7.4	100.0
	Total	1451	99.6	100.0	
Missing	DK,NA	6	.4		
Total		1457	100.0		

Note, first of all, that this variable is ordinal in level of measurement. There is a true zero point ("Never") but attendance rates are grouped into broad, unequal categories.

Secondly, note that this table looks very much like the frequency distributions presented in this chapter. The categories (or scores or class intervals) are listed on the left and frequencies appear in the next column. We see immediately that the most common score for church attendance is "never" (364 respondents) followed by "every week" (289 respondents).

The percentages in the next column (Percent) are calculated from the entire sample, including the six respondents whose response was DK (Don't Know) or NA (No Answer) and who should be regarded as "missing" and excluded from all calculations. We will ignore this column and move to the Valid Percent column, the percentages in which all missing cases are eliminated. A glance down this column reinforces the conclusion that the two most common scores are "never" (25.1%) and "every week" (19.9%). Finally, the Cumulative Percent column shows that most respondents (53.5%) attended church "Several Times a Year" or less.

Overall, how would you describe the pattern in this table? Is attendance at religious services an important value in the United States? Are Americans divided or polarized in their religious observances? How do you think that the pattern of attendance would vary by denomination or by age or gender?

Ratios, Rates, and Percentage Change

Ratios, rates, and percentage change are statistics that are used to summarize results simply and clearly. They may be used independently or with frequency distributions, and they may be computed for variables at any level of measurement. Although they are similar to each other, each statistic has a specific application and purpose, and we will consider them one at a time.

Ratios

Ratios are computed by dividing the frequency in one category by the frequency in another. The formula for a ratio is

FORMULA 2.3

$$\text{Ratio} = \frac{f_1}{f_2}$$

where f_1 = the number of cases in the first category
f_2 = the number of cases in the second category

Ratios are especially useful for comparing the relative sizes of different categories of a variable. To illustrate, suppose that you were interested in the relative sizes of the various religious denominations and found that a particular community included 1370 Protestant families and 930 Catholic families. To find the ratio of Protestants (f_1) to Catholics (f_2), divide 1370 by 930:

$$\text{Ratio} = \frac{f_1}{f_2} = \frac{1370}{930} = 1.47$$

The ratio of 1.47 means that there are 1.47 Protestant families for every Catholic family.

Ratios can be very economical ways of expressing relative size. In our example, it is obvious from the raw data that Protestants outnumber Catholics, but ratios tell us exactly how much one category outnumbers the other.

Ratios are often multiplied by some power of 10 to eliminate decimal points. For example, the ratio of Protestants to Catholics might be multiplied by 100 and reported as 147 instead of 1.47. This would mean that there are 147 Protestant families for every 100 Catholic families in the community. To ensure clarity, the comparison units for the ratio are often expressed as well. Based on a unit of ones, the ratio of Protestants to Catholics would be expressed as 1.47:1. Based on hundreds, the same statistic might be expressed as 147:100. *(For practice in computing and interpreting ratios, see problems 2.1 and 2.2.)*

Applying Statistics 2.4 Ratio

In Table 2.5, how many natural science majors are there compared to social science majors at College B? This question could be answered with frequencies, but a more easily understood way of expressing the answer would be with a ratio. The ratio of natural science to social science majors would be

$$\text{Ratio} = \frac{f_1}{f_2} = \frac{2799}{1884} = 1.49$$

For every social science major, there are 1.49 natural science majors at College B.

Rates

Rates provide still another way of summarizing the distribution of a single variable. A rate is defined as the number of actual occurrences of a phenomenon divided by the number of possible occurrences per some unit of time. Rates are usually multiplied by some power of 10 to eliminate decimal points.

For example, the crude death rate for a population is defined as the total number of deaths in that population (actual occurrences) divided by the number of people in the population (possible occurrences) per year. This quantity is then multiplied by 1000. This formula can be expressed as

$$\text{Crude death rate} = \frac{\text{Number of deaths}}{\text{Total population}} \times 1000$$

If there were 100 deaths during a given year in a town of 7000, the crude death rate would be

$$\text{Crude death rate} = \frac{100}{7000} \times 1000 = (0.01429) \times 1000 = 14.29$$

Or, for every 1000 people, there were 14.29 deaths during this particular year.

Rates are commonly used to measure crime. For a particular crime, the rate is the number of occurrences divided by the size of the population, multiplied by 100,000. For example, if a city of 237,000 people experienced 120 auto thefts during a particular year, the auto theft rate would be

$$\text{Auto theft rate} = \frac{120}{237{,}000} \times 100{,}000 = (0.0005063) \times 100{,}000 = 50.63$$

Or, for every 100,000 people, there were 50.63 auto thefts during the year in question. *(For practice in computing and interpreting rates, see problems 2.3 and 2.4a.)*

Applying Statistics 2.5 Rates

In 2010, there were 2500 births in a city of 167,000. In 1970, when the population of the city was only 133,000, there were 2700 births. Is the birth rate rising or falling?

Although this question can be answered from the preceding information, the trend in birth rates will be much more obvious if we compute birth rates for both years. Like crude death rates, crude birth rates are usually multiplied by 1000 to eliminate decimal points. For 1970:

$$\text{Crude Birth Rate} = \frac{2700}{133{,}000} \times 1000 = 20.30$$

In 1970, there were 20.30 births for every 1000 people in the city. For 2010:

$$\text{Crude Birth Rate} = \frac{2500}{167{,}000} \times 1{,}000 = 14.97$$

In 2010, there were 14.97 births for every 1000 people in the city. With the help of these statistics, the decline in the birth rate is clearly expressed.

Percentage Change

Measuring social change, in all its variety, is an important task for all social sciences. One very useful statistic for this purpose is the **percentage change**, which tells us how much a variable has increased or decreased over a certain span of time.

To compute this statistic, we need the scores of a variable at two different points in time. The scores could be in the form of frequencies, rates, or percentages. The percentage change will tell us how much the score has changed at the later time relative to the earlier time. Using death rates as an example once again, imagine a society suffering from a devastating outbreak of disease in which the death rate rose from 16 deaths per 1000 in the year 2000 to 24 deaths per 1000 in 2010. Clearly, the death rate is higher in 2010, but by how much relative to 2000?

The formula for the percent change is

FORMULA 2.4

$$\text{Percent change} = \left(\frac{f_2 - f_1}{f_1}\right) \times 100$$

where f_1 = first score, frequency, or value
f_2 = second score, frequency, or value

In our example, f_1 is the death rate in 2000 ($f_1 = 16$) and f_2 is the death rate in 2010 ($f_2 = 24$). The formula tells us to subtract the earlier score from the later score and then divide by the earlier score. The result expresses the size of the change in scores ($f_2 - f_1$) relative to the score at the earlier time (f_1). The value is then multiplied by 100 to express the change in the form of a percentage:

$$\text{Percent change} = \left(\frac{24 - 16}{16}\right) \times 100 = \left(\frac{8}{16}\right) \times 100 = (0.50) \times 100 = 50\%$$

The death rate in 2010 is 50% higher than in 2000. This means that the 2010 rate was equal to the 2000 rate *plus* half of the earlier score. If the rate had risen to 32 deaths per 1000, the percent change would have been 100% (the rate would have doubled), and if the death rate had fallen to 8 per 1000, the percentage change would have been −50%. Note the negative sign: It means that the death rate has decreased by 50% and the 2010 rate is half the size of the 2000 rate.

An additional example should make the computation and interpretation of the percentage change clearer. Suppose we wanted to compare the projected population growth rates for various nations through 2050. The necessary information is presented in Table 2.15, which shows the actual population for each nation in 2013 and the projected population for 2050. The "Increase/Decrease" column shows how many people will be added or lost over the time span.

Which nation will grow the fastest? Casual inspection will give us some information, but we can make precise comparisons using percentage change. The values in the right-hand column of Table 2.15 were computed by subtracting the

TABLE 2.15 Projected Population Growth for Six Nations, 2012–2050

Nation	Population, 2013 (f_1)	Population, 2050 (f_2)	Increase/Decrease ($f_2 - f_1$)	Percentage Change
China	1,357,400,000	1,314,400,000	−43,000,000	−3.17
U.S.	313,200,000	399,800,000	86,600,000	27.65
Nigeria	173,600,000	440,400,000	266,800,000	153.69
Mexico	117,600,000	150,000,000	32,400,000	27.55
U.K.	64,100,000	78,800,000	14,700,000	22.93
Canada	35,300,000	48,400,000	13,100,000	37.11

Source: Population Reference Bureau, 2014. *2013 World Population Data Sheet.* Accessed from http://www.prb.org/Publications/Datasheets/2013/2013-world-population-data-sheet/data-sheet.aspx

2013 population (f_1) from the 2050 population (f_2), dividing by the 2013 population, and multiplying by 100.

China has the largest population of these six nations, but it will actually lose about 3% of its population by 2050. The United States population will increase by almost 28%, and the populations of Mexico and the United Kingdom will increase by about one-quarter of their 2013 populations. Canada will grow more rapidly—by about 37%—but Nigeria has by far the highest growth rate: It will add the most people and its population will increase by over 150%. This means that the 2050 population of Nigeria will be 2.5 times larger than in 2013. *(For practice in computing and interpreting percent change, see problem 2.4b.)*

Applying Statistics 2.6 Percentage Change

The American family has been changing rapidly over the past several decades. One major change has been an increase in the number of married women and mothers with young children who hold jobs outside the home. For example, in 1970, 30.3% of married women with children under the age of 6 worked outside the home. In 2012, this percentage had risen to 64.8%. How large is this change?

Obviously, the 2012 percentage is much higher. Calculating the percentage change will tell us the exact magnitude of the change. The 1970 percentage is f_1 and the 2012 figure is f_2, so

$$\text{Percent Change} = \left(\frac{64.8 - 30.3}{30.3}\right) \times 100$$

$$= \left(\frac{34.5}{30.3}\right) \times 100 = 113.86$$

In the years between 1970 and 2012, the percentage of married women with children younger than 6 who worked outside the home increased by 113.86%. This is an extremely large change—more than doubling the earlier percentage—and signals major changes in this social institution.

Sources: U.S Department of Labor, 2014. "Employment Characteristics of Families, 2012. Retrieved from http://www.bls.gov/news.release/famee.nr0.htm U.S. Bureau of the Census. 2012. *Statistical Abstract of the United States, 2012*. Washington, DC: Government Printing Office. p. 385. Retrieved from http://www.census.gov/prod/2011pubs/12statab/labor.pdf

ONE STEP AT A TIME Finding Ratios, Rates, and Percent Change

Step ***Operation***

Ratios

1. Determine the values for f_1 and f_2. The value for f_1 will be the number of cases in the first category (e.g., the number of males on your campus), and the value for f_2 will be the number of cases in the second category (e.g., the number of females on your campus).
2. Divide the value of f_1 by the value of f_2.
3. You may multiply the value you calculated in step 2 by some power of 10 when reporting results.

Rates

1. Determine the number of actual occurrences (e.g., births, deaths, homicides, assaults). This value will be the numerator of the formula.
2. Determine the number of possible occurrences. This value will usually be the total population for the area in question and will be the denominator of the formula.
3. Divide the number of actual occurrences (step 1) by the number of possible occurrences (step 2).
4. Multiply the value you calculated in step 3 by some power of 10. Conventionally, birth rates and death rates are multiplied by 1000 and crime rates are multiplied by 100,000.

Percentage Change

1. Determine the values for f_1 and f_2. The former will be the score at time 1 (the earlier time) and the latter will be the score at time 2 (the later time).
2. Subtract f_1 from f_2.
3. Divide the quantity you found in step 2 by f_1.
4. Multiply the quantity you found in step 3 by 100.

STATISTICS IN EVERYDAY LIFE

Traffic Fatalities

Are American highways becoming more dangerous? Here are some statistical facts: In 1994 there were 40,716 motor vehicle deaths and that number fell to 32,367 in 2011. This is a decrease of 20.5%, and that is certainly good news, but do the raw numbers tell the whole story? Wouldn't this comparison be more meaningful and informative if it took account of changes in population, number of cars and drivers on the road, or changing driving habits?

One way to make the statistic more useful would be to compute a rate of traffic deaths based on the number of miles Americans drive each year: For every 100 million miles driven, the rate of traffic deaths fell from 1.7 in 1990 to 1.1 in 2011, an even more dramatic decrease of about 35.3%. What factors might account for this increase in highway safety?

Source: National Highway Traffic Safety Administration. 2014. Retrieved from http://www-fars.nhtsa.dot.gov/Main/index.aspx

BECOMING A CRITICAL CONSUMER: Lying with Statistics

The numbers may not lie, but people can shade the truth, confuse the issue, mislead, misunderstand, and deceive. Part of becoming statistically literate is learning how to analyze numbers critically and defend yourself against their misuse. Entire books have been written about lying with statistics: the best we can do in this limited space is to alert you to some common misuses and remind you to be prepared to use your own critical abilities.

Perhaps the most challenging examples of "lying" with statistics are situations in which statistics are accurate but present an incomplete or biased picture. Economist Charles Wheelan* cites an excellent example of this. The personal income tax rate in his state was raised from 3% to 5%. The political party that favored the increase presented it as a change of a mere 2 percentage points, a characterization that is technically true. Opponents, on the other hand, calculated percentage change, which showed a huge 67% tax increase:

$$\text{Percent Change} = \left(\frac{5-3}{3}\right) \times 100$$

$$= \left(\frac{2}{3}\right) \times 100 = 66.67$$

Both statistics are true, but which is more relevant and appropriate?

*Wheelen, Charles. 2013. *Naked Statistics: Stripping the Dread from the Data*. New York: Norton, p. 29.

Using Graphs to Present Data

Researchers frequently use graphs to present their data in a visually dramatic way. These devices are particularly useful for communicating the overall shape of a distribution and for highlighting any clustering of cases in a particular range of scores.

Many types of graphs are available, but we will examine just four. The first two, pie charts and bar charts, are appropriate for variables at any level of measurement. The last two, histograms and line charts (or frequency polygons), are used with interval-ratio variables.

All of the graphs described here are easily produced with SPSS, Microsoft Excel, and other software programs commonly available on college campuses. These programs are relatively flexible and easy to use. I will demonstrate how to use SPSS to produce graphs in the next section and in the "You Are the Researcher" section at the end of this chapter.

Pie Charts

A **pie chart** is an excellent way of displaying the relative sizes of the categories of a variable. Figure 2.1 shows the distribution of religious identification in the United States first shown in Table 2.2. The table is reproduced here as Table 2.16 with percentages added.

The pie chart divides a circle into "slices" proportional to the relative frequencies of the categories. The largest slice represents the Protestant group, with almost 54% of the cases, and the smallest slice represents the Jewish group, the smallest category. You might think of a pie chart as a visual frequency distribution, and Figure 2.1 dramatically and clearly displays the relative sizes of the various religions.

FIGURE 2.1 Self-Described Religious Identifications of Adult Americans, 2008.

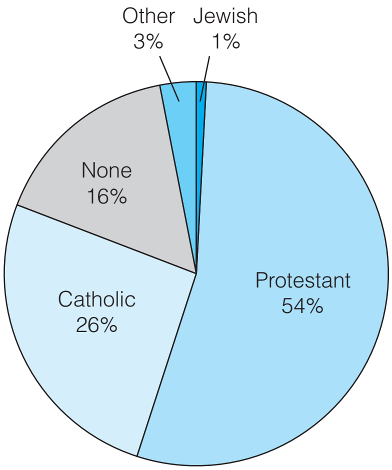


TABLE 2.16 Self-Described Religious Identifications of Adult Americans, 2008

Religious Group	Frequency	Percentage
Protestant	116,203,000	53.71
Catholic	57,199,000	26.44
Jewish	2,680,000	1.24
Other	6,116,000	2.83
None	34,169,000	15.79
	N = 216,367,000	100.01

Source: See Table 2.2.

Bar Charts

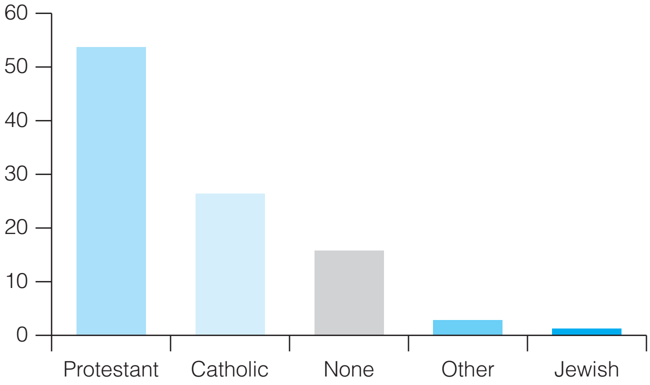
Like pie charts, **bar charts** are straightforward. The categories of the variable are arrayed along the horizontal axis (or *abscissa*) and frequencies, or percentages if you prefer, along the vertical axis (or *ordinate*). The heights of the bars are proportional to the relative frequencies of the categories and, like pie charts, bar charts are visual equivalents of frequency distributions. Figure 2.2 reproduces the data on religious identification displayed in Figure 2.1 and Table 2.16.

Figure 2.2 would be interpreted in exactly the same way as the pie chart in Figure 2.1, and researchers are free to choose between these two methods of displaying data. However, if a variable has more than four or five categories, the bar chart would be preferred, since the pie chart gets very crowded with many categories. *(For practice in constructing and interpreting pie charts and bar charts, see problem 2.13.)*

Histograms

Histograms look a lot like bar charts and, in fact, are constructed in much the same way. However, the bars in a histogram touch each other in a continuous

FIGURE 2.2 Self-Described Religious Identifications of Adult Americans, 2008 (percentages)

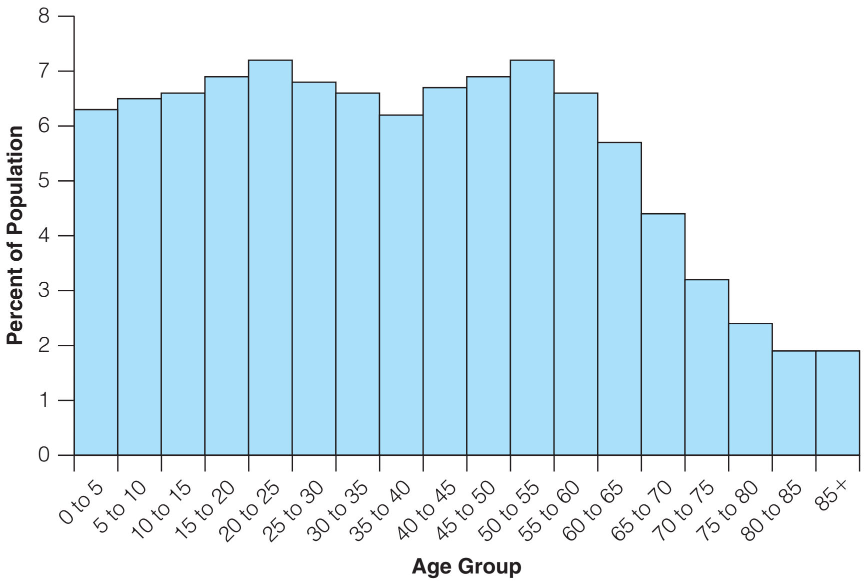


series from the lowest to the highest scores. These graphs are most appropriate for interval-ratio-level variables that have many scores covering a wide range.

As an example, Figure 2.3 uses a histogram to display the distribution of age for the population of the United States in 2012. Note that age is an interval-ratio variable that has been collapsed into categories of 5 years, except for the oldest category (85 and older), which is an open-ended interval.

The bars in the graph are 5 years wide, and their uneven heights reflect the varying number of respondents for each 5-year group. As you would expect, there is a sharp drop in the graph starting at about age 60 and continuing to the oldest age group, a reflection of the higher death rate in older populations.

FIGURE 2.3 Age Distribution, United States 2012



Note that the graph has peaks in the 50 to 55 age groups (people born between 1962 and 1967) and the 20 to 25 age groups (people born between 1992 and 1997). The older age group is part of the "baby boom," produced by the unusually high birth rate between the end of World War II and the mid-1960s. The second, or younger, peak is partly an echo of the first. The two peaks are 30 years, or about a generation, apart and many members of the younger group are the children of the older group.

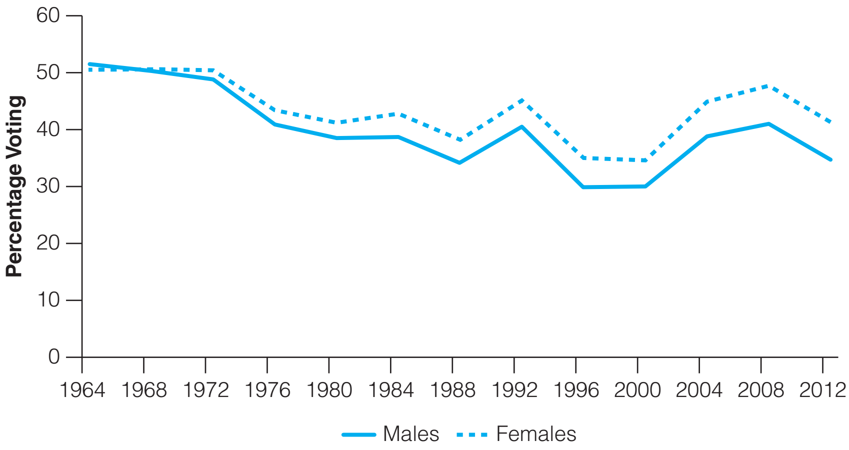
Line Charts

Line charts (or **frequency polygons**) are similar to histograms but use dots connected with lines and positioned at the midpoint of the intervals rather than bars to represent the frequencies. The height of the dot reflects the number of cases in the interval. These graphs are especially appropriate for interval-ratio-level variables with many scores. Figure 2.4 displays a line chart showing voter turnout in presidential elections for male and female young adult voters between 1964 and 2012.

Note that the two lines move in parallel over the time period and that females consistently have a higher turnout than males. Note also that turnout for both groups was at its highest in the 1960s and for the 1972 election and generally has drifted down since that time, although there were spikes in turnout in 1992 (when President Clinton was first elected) and, especially, in 2008 (President Obama's first election).

Histograms and line charts are alternative ways of displaying essentially the same message. Thus, the choice between the two techniques is left to the aesthetic pleasures of the researcher. *(For practice in constructing and interpreting histograms and line charts, see problems 2.10 and 2.13.)*

FIGURE 2.4 Turnout in Presidential Election Years for Young (age 18 to 24) Male and Female Voters, 1964–2012

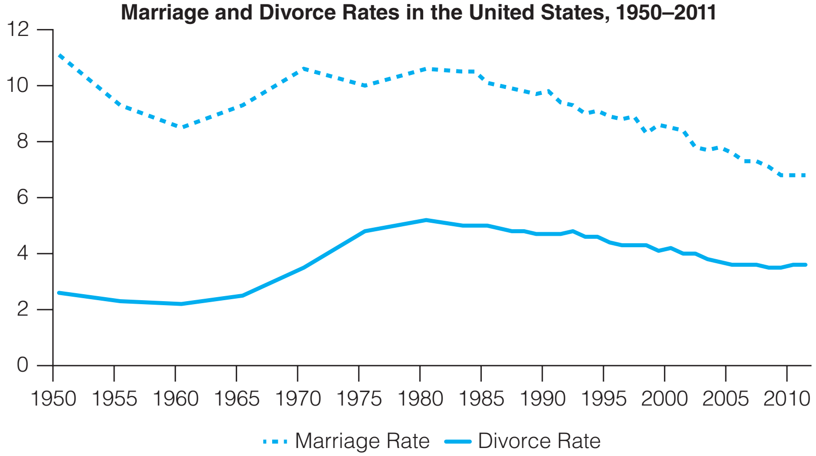


STATISTICS IN EVERYDAY LIFE

Marriage and Divorce

There is a great deal of concern in U.S. society about the future of the American family and, more specifically, our high divorce rate. The U.S. divorce rate is higher than that of other Western industrialized nations but it may surprise you to learn that the rate has been falling since the early 1980s. Why? One reason is that the marriage rate has also been falling—you can't get divorced unless you get married! The line chart displays both trends.

Marriage and Divorce Rates in the United States, 1950–2011



Rates are per 1000 population

Sources: 1950–2008: U.S. Bureau of the Census, 2012. Retrieved from http://www.census.gov/prod/2011pubs/12statab/vitstat.pdf 2009–2011: Centers for Disease Control, 2014. Retrieved from http://www.cdc.gov/nchs/nvss/marriage_divorce_tables.htm

This information could have been presented in a table. What are the advantages of the "picture" presented in a graph? For example, can you see a difference in the rate of decline of the two lines? Is one rate falling faster than the other? What are the implications of this difference? What else does the graph imply about family life in the United States? What sociological variables might help to account for these changes?

Are there any disadvantages of using graphs rather than tables? What kinds of variables are most appropriately displayed in line charts?

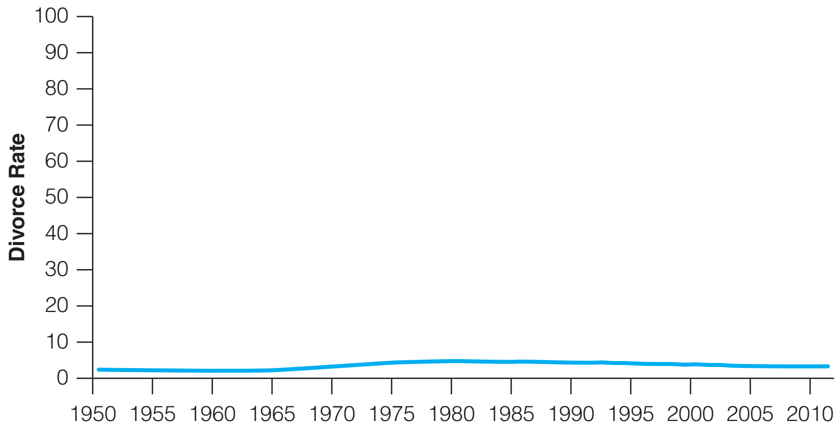
BECOMING A CRITICAL CONSUMER: Tricks with Graphs

Charts and graphs can be powerful ways to summarize data but, like any statistical tool, they can be manipulated and distorted. One way to change the message conveyed by a graph is to alter the horizontal or vertical axes. For example, look again at the graph displaying marriage and divorce rates in "Statistics in Everyday Life" and note that the divorce rate rises and falls over the time period. This picture can be altered by using only some dates on the horizontal axis. Using only the oldest (1950) and most recent (2011) dates would give the impression that the divorce rate had *increased* from 2.6 to 3.6.

(*continued*)

The vertical axis (divorce rate) can also be manipulated. If someone wanted to create the impression that the divorce rate had essentially remained unchanged over the time period, they could increase the scale of the vertical axis. In the graph below, the scale has been increased from "0 to 12" to "0 to 100" and the line now seems essentially flat. Since divorce rates are generally less than 5.0, the larger scale is unwarranted. You should be especially wary of graphs that use extended or exaggerated scales.

Divorce Rate, 1950–2011, with an Exaggerated Vertical Axis



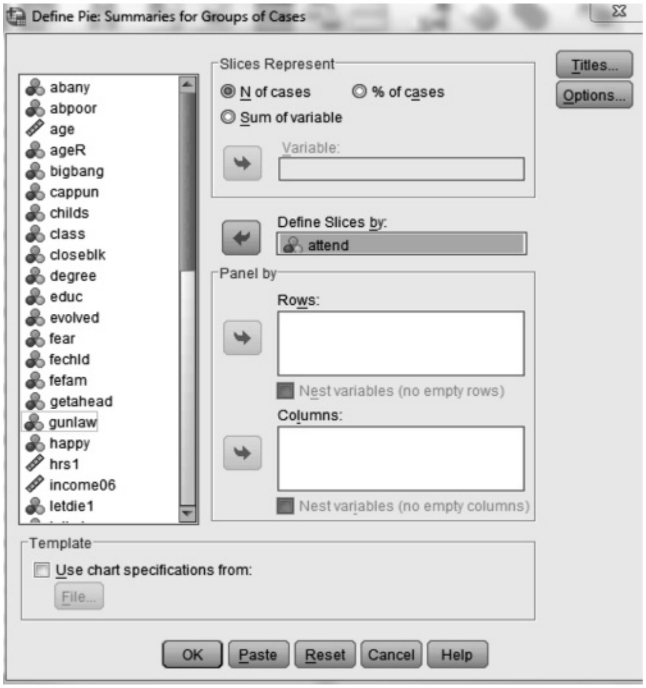
Using SPSS to Produce Graphs

SPSS can produce a variety of graphs. Here, we will illustrate the procedures by producing a pie chart and histogram using the General Social Survey (GSS). We will use *attend* again for the pie chart and *age* for the line chart. We can compare the distribution of age for the GSS sample with the population distribution presented in Figure 2.3.

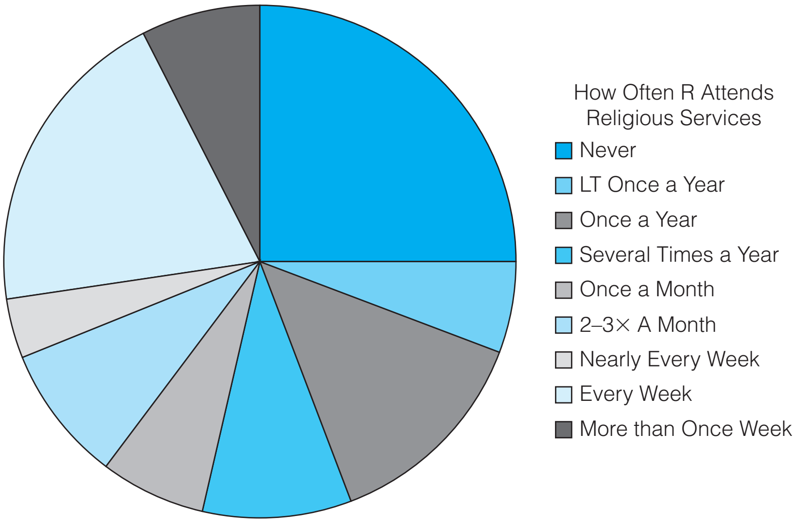
Follow these steps to get a pie chart:

1. Find and click the SPSS icon on your desktop.
2. Load the database.
   a. Find the **File** command on the far left of the menu bar and then click **File → Open Data.**
   b. Find the 2012 General Social Survey (*GSS2012.sav*) database file supplied with this text. You can download this file from the website for this text if you haven't already.
3. From the menu bar across the top of the SPSS window, click **Graphs, Legacy Dialog,** and **Pie**.
4. The "Pie Charts" window will appear with "Summaries for groups of cases" preselected. This is what we want, so click **Define**.

5. A new window will appear. Find *attend* in the list of variables on the left and click the arrow to move the variable name to the "Define Slices by:" box. When you are done, the window will look like this:



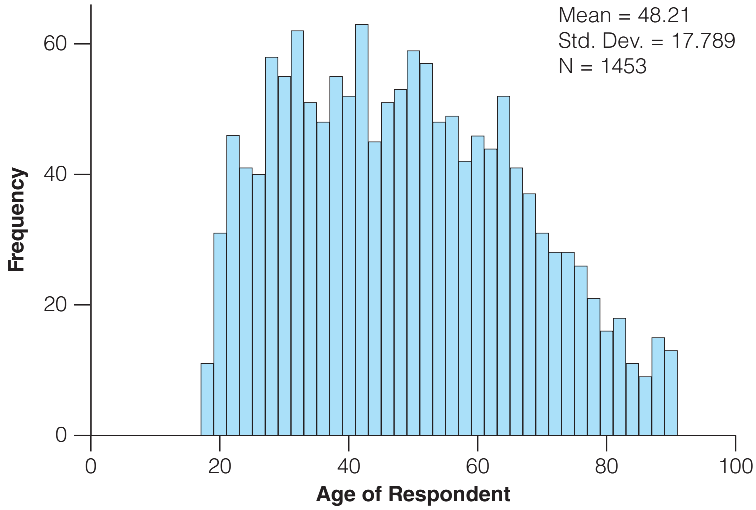
6. Click **OK** and the pie chart will be sent to the "SPSS Output" window:



Remember that a pie chart is essentially a visual frequency distribution. As we saw before, the most common score (largest "slice") is "never" and the "every week" slice is nearly as large.

To create a histogram, complete the following steps:

1. Click **Graphs, Legacy Dialog**, and **Histogram**.
2. In the "Histogram" window, find *age* and move it to the "Variable:" box.
3. Click **OK** and the graph will be sent to the "SPSS Output" window:



Note that this graph is more irregular and jagged than Figure 2.3. This is partly because of sample size. This graph is based on about 1500 respondents, and Figure 2.3 includes over 300 million. Also note that since the General Social Survey is given only to adults, this graph has no respondents younger than 18.

BECOMING A CRITICAL CONSUMER: Statistics in the Professional Literature

A recent report calls adolescent use of cigarettes, alcohol, and drugs "the #1 public health problem in the United States," especially because of the associations between substance abuse and accidental injuries and deaths, mental health problems, unintended pregnancies, and involvement in criminal activities. The authors of the report use a number of simple statistics to make their point, including percentages, tables, and graphs. Here, we will briefly illustrate their argument and examine their statistics.

How extensive is the problem? According to the report, substance abuse begins before high school and accelerates through the grades. The report makes this point in a series of tables and graphs, two of which are reproduced here.

The table on the following page shows that alcohol use is common for high school freshmen, which suggests that the pattern began earlier in their lives. Binge drinking (consumption of five or more drinks at a time) is also common and increases to one-third of students for seniors.

(*continued*)

BECOMING A CRITICAL CONSUMER (*continued*)

High School Students Who Have Ever Used Alcohol and Who Binge Drink by Grade

Grade	Ever Used Alcohol	Binge Drink
9th	63.4%	15.3%
10th	71.1%	22.3%
11th	77.8%	28.3%
12th	79.6%	33.5%

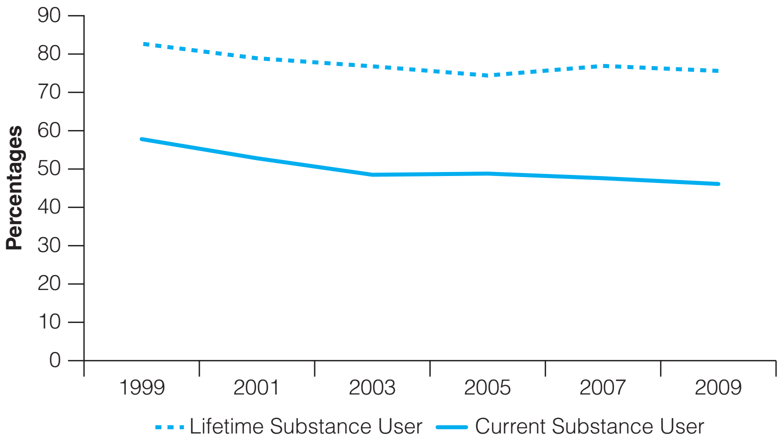
National Center on Addiction and Substance Abuse at Columbia University. 2011. p. 24

The following figure shows that rates of substance use—including cigarettes, alcohol, marijuana, and cocaine—have declined somewhat over the past decade but remain quite high. In the most recent year, three-quarters of high school students reported that they have used a substance at least once ("lifetime substance use") and a little less than half report current use.

The report is intended for the general public, parents, and policy makers and uses only the kinds of statistics and graphs covered in this chapter. However, the authors build a powerful case—thoroughly anchored in empirical evidence—for the dire importance of the problem they address.

Source: National Center on Addiction and Substance Abuse at Columbia University. 2011. *Adolescent Substance Use: America's #1 Health Problem*. Retrieved from http://www .casacolumbia.org/upload/2011/20110629adolescentsubstanceuse.pdf

Trends in Substance Use Among High School Students



SUMMARY

1. We considered several different ways of summarizing the distribution of a single variable and, more generally, reporting the results of our research. Our emphasis throughout was on the need to communicate the results clearly and concisely.
2. Frequency distributions are tables that summarize the entire distribution of some variable. Statistical analysis almost always starts with the construction and review of these tables for each variable. Columns for percentages, cumulative frequencies, and/or cumulative percentages often enhance the readability of frequency distributions.
3. Percentages and proportions, ratios, rates, and percentage changes give us different ways of expressing our results in terms of relative frequency. Percentages and proportions report the relative occurrence of some category of a variable compared with the distribution as a whole. Ratios compare two categories with each other, and rates report the actual occurrences of some

phenomenon compared with the number of possible occurrences per some unit of time. Percentage change shows the relative increase or decrease in a variable over time.

4. Pie charts and bar charts, histograms, and line charts (or frequency polygons) are graphs that express the information contained in the frequency distribution in a compact and visually dramatic way.

SUMMARY OF FORMULAS

FORMULA 2.1 Percentage: $$\% = \left(\frac{f}{N}\right) \times 100$$

FORMULA 2.2 Proportion: $$p = \left(\frac{f}{N}\right)$$

FORMULA 2.3 Ratio: $$\text{Ratio} = \frac{f_1}{f_2}$$

FORMULA 2.4 Percent Change: $$\text{Percent Change} = \left(\frac{f_2 - f_1}{f_1}\right) \times 100$$

GLOSSARY

Bar chart. A graph used for nominal and ordinal variables with only a few categories. Categories are represented by bars of equal width, the height of each corresponding to the number (or percentage) of cases in the category.

Class intervals. The categories used in the frequency distributions for interval-ratio variables.

Cumulative frequency. An optional column in a frequency distribution that displays the number of cases within an interval and all preceding intervals.

Cumulative percentage. An optional column in a frequency distribution that displays the percentage of cases within an interval and all preceding intervals.

Frequency distribution. A table that displays the number of cases in each category of a variable.

Frequency polygon. See **Line chart**.

Histogram. A graph used for interval-ratio variables. Class intervals are represented by contiguous bars of equal width (equal to the class limits), the height of each corresponding to the number (or percentage) of cases in the interval.

Line chart. A graph used for interval-ratio variables. Class intervals are represented by dots placed over the midpoints, the height of each corresponding to the number (or percentage) of cases in the interval. All dots are connected by straight lines.

Midpoint. The point exactly halfway between the upper and lower limits of a class interval.

Percentage. The number of cases in a category of a variable divided by the number of cases in all categories of the variable, the entire quantity multiplied by 100.

Percent change. A statistic that expresses the magnitude of change in a variable from time 1 to time 2.

Pie chart. A graph used for nominal and ordinal variables with only a few categories. A circle (the pie) is divided into segments proportional in size to the percentage of cases in each category of the variable.

Proportion. The number of cases in one category of a variable divided by the number of cases in all categories of the variable.

Rate. The number of actual occurrences of some phenomenon or trait divided by the number of possible occurrences per some unit of time. Rates are usually multiplied by some power of 10.

Ratio. The number of cases in one category divided by the number of cases in some other category.

Stated class limits. The class intervals of a frequency distribution.

PROBLEMS

2.1 *SOC* The table that follows reports the marital status of 20 respondents in two different apartment complexes. (*HINT: As you solve these problems, make sure that you have the correct numbers in the numerator and denominator. For example, problem 2.1a asks for "the percentage of respondents in each* *complex who are married," and the denominators will be 20 for these two fractions. Problem 2.1d, on the other hand, asks for the "percentage of the single respondents who live in Complex B," and the denominator for this fraction will be 4 + 6, or 10.*)

Status	Complex A	Complex B
Married	5	10
Unmarried ("living together")	8	2
Single	4	6
Separated	2	1
Widowed	0	1
Divorced	1	0
	20	20

a. What percentage of the respondents in each complex are married?
b. What is the ratio of single to married respondents at each complex?
c. What proportion of the residents of each complex are widowed?
d. What percentage of the single respondents live in Complex B?
e. What is the ratio of the "unmarried/living together" to the "married" at each complex?

2.2 *SOC* At St. Winnefred Hospital, the numbers of males and females in various job categories are as follows:

Job	Males	Females	Totals
Doctor	83	15	98
Nurse	62	116	178
Orderly	151	12	163
Lab technician	32	30	62
Administrator	12	1	13
Clerk	72	141	213
Totals	412	315	727

Read each question carefully before constructing the fraction and solving for the answer. (*HINT: Be sure you place the proper number in the denominator of each fraction. For example, some problems use the total number of males or females as the denominator, whereas others use the total number of subjects.*)

a. What percentage of nurses are male?
b. What proportion of orderlies are female?
c. For doctors, what is the ratio of males to females?
d. What percentage of the total staff are males?
e. What is the ratio of males to females for the entire sample?
f. What proportion of the clerks are male?
g. What percentage of the sample are administrators?
h. What is the ratio of nurses to doctors?
i. What is the ratio of female lab technicians to male lab technicians?
j. What proportion of the males are clerks?

2.3 *CJ* The town of Shinbone, Kansas, has a population of 211,732 and experienced 47 bank robberies, 13 murders, and 23 auto thefts during the past year. Compute a rate for each type of crime per 100,000 population. (*HINT: Make sure that you set up the fraction with size of population in the denominator.*)

2.4 *CJ* The numbers of homicides in five U.S. states and five Canadian provinces for the years 1997 and 2012 are presented in two accompanying tables.

a. Calculate the homicide rate per 100,000 population for each state and each province for each year. Relatively speaking, which state and which province had the highest homicide rates in each year? Which society seems to have the higher homicide rate? Write a paragraph describing these results.
b. Using the rates you calculated in part a, calculate the percent change between 1997 and 2012 for each state and each province. Which states and provinces had the largest increase and decrease? Which society seems to have the largest change in homicide rates? Summarize your results in a paragraph.

	1997		2012	
State	Homicides	Population	Homicides	Population
New Jersey	338	8,053,000	388	8,864,590
Iowa	52	2,852,000	45	3,074,186
Alabama	426	4,139,000	342	4,882,023
Texas	1327	19,439,000	1144	26,059,203
California	2579	32,268,000	1884	38,041,430

Source: Federal Bureau of Investigation. *Uniform Crime Reports*. Retrieved from http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2012/crime-in-the-u.s.-2012/tables/4tabledatadecoverviewpdf/table_4_crime_in_the_united_states_by_region_geographic_division_and_state_2011-2012.xls

	1997		2012	
Province	Homicides	Population	Homicides	Population
Nova Scotia	24	936,100	17	945,100
Quebec	132	7,323,600	108	8,043,600
Ontario	178	11,387,400	162	13,490,500
Manitoba	31	1,137,900	52	1,274,100
British Columbia	116	3,997,100	71	4,678,100

Source: Statistics Canada. Retrieved from http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/legal12a-eng.htm

2.5 *SOC* The scores of 15 respondents on four variables are as reported here. These scores were taken from a public-opinion survey called the General Social Survey (GSS), the data set used for the computer exercises in this text. For the actual questions, see Appendix G. The numerical codes for the variables are as follows:

Sex	Support for Gun Control	Level of Education	Age
1 = Male	1 = In favor	0 = Less than HS	Actual years
2 = Female	2 = Opposed	1 = HS	
		2 = Jr. college	
		3 = Bachelor's	
		4 = Graduate	

Case Number	Gender	Support for Gun Control	Level of Education	Age
1	2	1	1	45
2	1	2	1	48
3	2	1	3	55
4	1	1	2	32
5	2	1	3	33
6	1	1	1	28
7	2	2	0	77
8	1	1	1	50
9	1	2	0	43
10	2	1	1	48
11	1	1	4	33
12	1	1	4	35
13	1	1	0	39
14	2	1	1	25
15	1	1	1	23

Construct a frequency distribution for each variable. Include a column for percentages.

2.6 *SW* A local youth service agency has begun a sex education program for teenage girls who have been referred by the juvenile courts. The girls were given a 20-item test for general knowledge about sex, contraception, and anatomy and physiology upon admission to the program and again after completion of the program. The scores of the first 15 girls to complete the program are as follows.

Case	Pretest	Posttest	Case	Pretest	Posttest
A	8	12	I	5	7
B	7	13	J	15	12
C	10	12	K	13	20
D	15	19	L	4	5
E	10	8	M	10	15
F	10	17	N	8	11
G	3	12	O	12	20
H	10	11			

Construct frequency distributions for the pretest and posttest scores. Include a column for percentages. *(HINT: There were 20 items on the test, so the maximum range for these scores is 20. If you use 10 class intervals to display these scores, the interval size will be 2. Since there are no scores of 0 or 1 for either test, you may state the first interval as 2–3. To make comparisons easier, both frequency distributions should have the same intervals.)*

2.7 *SOC* Sixteen high school students completed a class to prepare them for college entrance examinations. Their scores were as follows.

420	345	560	650
459	499	500	657
467	480	505	555
480	520	530	589

These same 16 students were given a test of math and verbal ability to measure their readiness for college-level work. Scores are reported here in terms of the percentage of correct answers for each test.

Math Test			
67	45	68	70
72	85	90	99
50	73	77	78
52	66	89	75

Verbal Test			
89	90	78	77
75	70	56	60
77	78	80	92
98	72	77	82

Display each of these variables in a frequency distribution with columns for percentages and cumulative percentages.

2.8 GER Twenty-five residents of a community for senior citizens were asked to keep track of the number of hours they watched TV during the past week.

20	2	12	7	32
7	10	7	13	17
14	15	15	9	27
5	21	4	7	6
2	18	10	5	11

a. Construct a frequency distribution to display these data.
b. What are the midpoints of the class intervals?
c. Add columns to the table to display the percentage distribution, cumulative frequencies, and cumulative percentages.
d. Write a paragraph summarizing this distribution of scores.

2.9 SOC Twenty-five students completed a questionnaire that measured their attitudes toward interpersonal violence. A high score indicates that the respondent believed that physical force could legitimately be used in many situations. Low scores mean that the respondent thought that violence could be justified in few or no situations.

52	47	17	8	92
53	23	28	9	90
17	63	17	17	23
19	66	10	20	47
20	66	5	25	17

a. Construct a frequency distribution to display these data.
b. What are the midpoints of the class intervals?
c. Add columns to the table to display the percentage distribution, cumulative frequencies, and cumulative percentages.
d. Write a paragraph summarizing this distribution of scores.

2.10 PA/CJ As part of an evaluation of the efficiency of your local police force, you have gathered the following data on police response time to calls for assistance during two different years. (Response times were rounded off to whole minutes.) Convert both frequency distributions into percentages, and write a paragraph comparing the changes in response time between the two years.

Response Time, 2000	Frequency (*f*)
21 minutes or more	20
16–20 minutes	90
11–15 minutes	185
6–10 minutes	370
Less than 6 minutes	210
	875

Response Time, 2010	Frequency (*f*)
21 minutes or more	45
16–20 minutes	95
11–15 minutes	155
6–10 minutes	350
Less than 6 minutes	250
	895

Statistical Analysis Using SPSS

2.11 SOC In this exercise, you will use SPSS to graph trends in crime in the United States. You will use the *CrimeTrends84-12* data set, which is available from the website for this text.

- Find and click the SPSS icon on your desktop.
- Load the *CrimeTrends84-12* data set.
- From the menu bar across the top of the SPSS window, click **Graphs**, then **Legacy Dialogs** and **Line**.
- The "Line Chart" dialog box will open. Click **Simple** in the top part of the window and **Values of Individual Cases** from the "Data in Chart Are" panel.
- Click **Define**.

- The "Define Simple Line" dialog box will open. In the "Category Labels" panel, click **Variable**. Next click *year* in the list of variables and click the arrow pointing to the "Variable" box.
- Click one of the measures of crime in the list of variables and click the arrow pointing to the "Line Represents" box. Click **OK** and the line chart will be created in the "SPSS Output" window.
- Repeat these steps for the other crime variables.
  - Each time you return to the "Define Simple Line" dialog box, click the variable name in the "Line Represents" box and then click the left-pointing arrow to return the variable to the list of variable names.
  - Click on a new measure of crime and click the arrow pointing to the "Line Represents" box. Click **OK** and the next line chart will be created. DO NOT change the *year* variable. Repeat until you have charts for all five measures of crime.

Write a paragraph describing each of these graphs. What similarities and differences can you observe among the three graphs? (For example, do crime rates always change in the same direction?) Note the differences in the vertical axis from chart to chart—for homicide, the axis has a much smaller range than for the other crimes, which are far more common, and a scale with smaller intervals is needed to display the rates.

2.12 *SOC* In this exercise, you will use the *GSS2012.sav* database supplied with this text to produce frequency distributions and bar charts or pie charts for three variables. The steps are the same as in the "Using SPSS: Frequency Distributions" section of this chapter, except that you will choose your own variables.

- Find and click the SPSS icon on your desktop.
- Load the 2012 GSS (*GSS2012.sav*) database supplied with this text.
- Use Appendix G to select one nominal-level, one ordinal-level (other than *attend*), and one interval-ratio-level variable. To determine level of measurement, use "ONE STEP AT A TIME: Determining the Level of Measurement of a Variable" from Chapter 1 and Table 1.5. Make sure that you examine the way in which the scores of the variable are *actually stated*.
- From the menu bar across the top of the SPSS window, click **Analyze**, **Descriptive Statistics**, and **Frequencies**.
- Choose your three variables from the list and then click the arrow to move each variable name—one at a time—to the box on the right.
- Click **OK** and the frequency distributions will be sent to the "Output" window.

Study the tables carefully, including the **Valid Percent** column and, if appropriate, the **Cumulative Percent** column. Write a sentence or two of interpretation for each table.

2.13 *SOC* In this exercise, you will use the *GSS2012.sav* database to produce pie charts and bar charts for religious denomination (*relig*) and a line chart and histogram for hours worked (*hrs1*). With the *GSS2012.sav* data loaded, click **Graphs** and then **Legacy Dialogs**. The submenu displays the types of graphs available. Each time you finish a graph, return to the "Data Editor" window and click **Graphs** and **Legacy Dialogs** again to get the next graph.

- For a pie chart:
  - Click **Pie** and then **Define** on the "Pie Chart" window.
  - On the "Define Pie:" window, move *relig* into the "Define Slices by" box.
  - Click **OK.**
- For a bar chart:
  - Click **Bar** and then **Define** on the "Bar Charts" window.
  - On the "Define Simple Bar:" window, move *relig* into the "Category Axis" box.
  - Click **OK**.
- For a line chart:
  - Click **Line** and then **Define** on the "Line Charts" window.
  - On the "Define Simple Line:" window, move *hrs1* into the "Category Axis" box.
  - Click **OK**.
- For a histogram:
  - Click **Histogram**.
  - Move *hrs1* into the "Variables" box.
  - Click **OK**.

Use the graphs to write a few sentences of description for each variable.

YOU ARE THE RESEARCHER

Is There a "Culture War" in the United States?

One of the early steps in a research project is to produce frequency distributions for all variables. If nothing else, these tables provide excellent background information and, sometimes, you can use them to begin to answer research questions. In this installment of "You Are the Researcher," you will use SPSS to produce summary tables for several variables that measure attitudes about controversial issues in U.S. society and that may map the battlefronts in what many call the American culture wars.

The United States seems to be divided on a number of religious, political, and cultural issues and values. We might characterize the opposing sides using terms like "liberal vs. conservative," "modern vs. traditional," or "progressive vs. old-school"; some of the most bitter debates across these lines include abortion, gay marriage, and gun control. As you know, debates over issues like these can be intense, bitter, and even violent: Adherents of one position may view their opponents with utter contempt, demonize them, and dismiss their arguments. How deep is this fault line in U.S. society?

We can begin to investigate these questions by choosing three variables from the 2012 General Social Survey (GSS) that seem to differentiate the sides in the American culture war. Before doing this, let's take a moment to consider the process of "picking variables." Technically, selecting a variable to represent a concept such as "culture war" is called *operationalization,* and this can be one of the most difficult steps in a research project.

We discussed operationalization in Chapter 1 when we reviewed the research process (see Figure 1.1). When we move from the "Theory" stage to the "Hypothesis" stage, we identify specific variables (such as responses to survey items) that match our concepts (like prejudice). This can be challenging because our concepts are usually quite abstract and subject to a variety of perspectives. What exactly is a culture war, and why are some positions liberal or traditional, conservative or progressive? In order to do research, we must use concrete variables to represent our abstract concepts, but which variables relate to which concepts?

Any pairing we make between variables and concepts is bound to be at least a little arbitrary. In many cases, the best strategy is to use several variables to represent the concept: If our operationalizations are reasonable, our selected variables will behave similarly and each will behave as the abstract concept would if we could measure it directly. This is why I will ask you to select three different variables to represent the culture wars. Each researcher may select different variables but, if everyone makes reasonable decisions, the chosen variables should be valid representations of the concept.

Begin by starting SPSS and opening the 2012 GSS. You can download this file from the website for this text if you haven't already.

Next, select three variables that seem relevant to the culture war by browsing through the list of variables in Appendix G. You can also see a list of variables in the database by clicking **Utilities → Variables** on the menu bar of the "Data Editor" window. After you have made your selections, complete the following steps.

Step 1: Identify Your Three Variables

Write the names of your variables as they appear in the database (for example: *abany, marital*, or *sex*). To "explain what the variable measures," look at the wording of the survey items in Appendix G or use the abbreviated statements in the **Utilities → Variables** window.

SPSS Variable Name	Explain Exactly What This Variable Measures

Step 2: Operationalization

Explain how each of your variables relates to the culture wars. Which position is liberal and which is conservative? For example, you might argue that liberals would be pro-choice on abortion and conservatives would be more supportive of gun control.

SPSS Variable Name	Explain the relationship between each variable and the culture wars. Which position is liberal and which is conservative?

Step 3: Producing Frequency Distributions

Now we are ready to generate output and get some background on the nature of disagreements over values and issues among Americans.

Generating Frequency Distributions

We produced and examined a frequency distribution for the variable ***sex*** in Appendix F and for *attend* in the "Using SPSS" demonstration in this chapter. Use the same procedures to produce frequency distributions for the three variables you used to represent the American culture wars. From the menu bar, click **Analyze, Descriptive Statistics**, and then **Frequencies**.

Find your three variables in the *Frequencies* window. Click each variable name and then, one variable at a time, click the arrow button in the middle of the screen to move the variable name to the right-hand window. SPSS will process all variables listed in the right-hand box at the same time. When you have all three of your variables listed in the right-hand box, click **OK** in the *Frequencies* window, and SPSS will rush off to create the frequency distributions you requested. The tables will be in the SPSS Viewer or "Output" window that will now be "closest" to you on the screen.

Reading SPSS Frequency Distributions

We discussed how to read frequency distributions produced by SPSS in the "Using SPSS" demonstration earlier in this chapter. Remember to disregard the "Percent" column.

Step 4: Interpreting Results

Summarize your results by reporting the percentage (not the frequencies) of respondents who endorsed each response. How large are the divisions in American values? Is there consensus on the issue measured by your variable (do the great majority endorse the same response) or is there considerable disagreement? The lower the consensus, the greater the opportunity for the issue to be included in the culture war.

SPSS name of variable 1: ____________________
Summarize the frequency distribution in terms of the percentage of respondents who endorsed each position:

Are these results consistent with the idea that there is a "war" over American values? How?

SPSS name of variable 2: ____________________
Summarize the frequency distribution in terms of the percentage of respondents who endorsed each position:

Are these results consistent with the idea that there is a "war" over American values? How?

SPSS name of variable 3: ____________________
Summarize the frequency distribution in terms of the percentage of respondents who endorsed each position:

Are these results consistent with the idea that there is a "war" over American values? How?

3 Measures of Central Tendency

LEARNING OBJECTIVES By the end of this chapter, you will be able to:

1. Explain the purposes of measures of central tendency and interpret the information they convey.
2. Calculate, explain, and compare and contrast the mode, median, and mean.
3. Explain the mathematical characteristics of the mean.
4. Select an appropriate measure of central tendency according to level of measurement and skew.
5. Use SPSS to produce means, medians, and modes.

USING STATISTICS

The statistical techniques presented in this chapter are used to find the typical case or average score on a single variable. They can be used to:

- Identify the most popular political party in a community.
- Compare the average cost of living or house or gas prices in different cities or states. For example, in 2012, the sales price listed for the typical house in San Francisco was $967,280 while the comparable price for Detroit was $129,900.
- Measure the “gender gap” in income by comparing the average wages for females and males.
- Track changes in American family life by reporting the average number of children or typical age at first marriage over time.

One benefit of frequency distributions and graphs is that they summarize the overall shape of a distribution of scores in a way that can be quickly comprehended. However, you will almost always need to report more detailed information about the distribution. Specifically, two additional kinds of statistics are extremely useful: some idea of the typical or average case in the distribution (for example, “The average starting salary for social workers in this state is $43,000”) and some idea of how much variety there is in the distribution (“In this state, starting salaries for social workers range from $35,000 to $47,000 per year”). The first kind of statistic, called **measures of central tendency**, is the subject of this chapter. The second kind of statistic, measures of dispersion, is presented in Chapter 4.

The three commonly used measures of central tendency—the mode, median, and mean—are all probably familiar to you. They summarize a distribution of scores by describing the most common score (the mode), the score of the middle case (the median), or the average score (the mean) of that distribution. These statistics are powerful because they can reduce huge arrays of data to a single, easily understood number. Remember that the central purpose of descriptive statistics is to summarize or "reduce" data.

Even though they share a common purpose, the three measures of central tendency are different statistics and they will have the same value only under certain conditions. They vary in terms of level of measurement and, perhaps more importantly, in terms of how they define central tendency. They will not necessarily identify the same score or case as "typical." Thus, your choice of an appropriate measure of central tendency will depend, in part, on the sort of information you want to convey.

The Mode

The **mode** of any distribution of scores is the value that occurs most frequently. For example, in the set of scores 58, 82, 82, 90, 98, the mode is 82 because it occurs twice and the other scores occur only once.

The mode is a simple statistic, most useful when you are interested in the most common score and when you are working with nominal-level variables. In fact, the mode is the only measure of central tendency for nominal-level variables. For example, Table 3.1 reports the number of tourists who visited the ten most popular nations in 2012. The mode of this distribution, the single most common destination (by far), is France.

You should note that the mode has several limitations. First, distributions can have no mode at all (when all scores have the same frequency) or so many modes that the statistic becomes meaningless. For example, consider the unusual but not impossible distributions of test scores in Table 3.2. Test A has no "single most common" score and, therefore, no mode. Test B has four different

TABLE 3.1 Top Ten Most Visited Nations, 2012

Nation	Number of Visitors
China	57,700,000
France	83,000,000
Germany	30,400,000
Italy	46,400,000
Malaysia	25,000,000
Russia	25,700,000
Spain	57,700,000
Turkey	35,700,000
United Kingdom	29,300,000
United States	67,000,000

Source: United Nations World Tourism Organization. Retrieved from http://dtxtq4w60xqpw.cloudfront.net/sites/all/files/pdf/unwto_highlights14_en.pdf

TABLE 3.2 Distributions of Scores on Two Tests

	Test A	Test B
Score (% correct)	Frequency of Scores	Frequency of Scores
97	14	22
91	14	3
90	14	4
86	14	22
77	14	3
60	14	22
55	14	22
	$N = 98$	98

TABLE 3.3 A Distribution of Test Scores

Score (% correct)	Frequency
93	8
68	3
67	4
66	2
62	7
	$N = 24$

modes—55, 60, 86, and 97—and it is unlikely that reporting all four would convey any useful information about central tendency in the distribution.

Second, the modal score of ordinal or interval-ratio variables may not be central to the distribution as a whole. That is, *most common* does not necessarily identify the center of the distribution. For example, consider the unusual but, again, not impossible distribution of test scores in Table 3.3. The mode of the distribution is 93. Is this score close to the majority of the scores? If the instructor summarized this distribution by reporting only the modal score, would he or she be conveying an accurate picture of the distribution as a whole? *(For practice in finding and interpreting the mode, see problems 3.1 to 3.7.)*

STATISTICS
IN EVERYDAY LIFE

The Changing American Religious Profile

The modal (or most common) religion in the United States, by far, is Protestant (see Tables 2.2 to 2.4). However, the nation is changing rapidly in this area of life, as in so many others. For example, since 1990, the percentage of Americans who identify as "Protestant" has dropped from 61% to 54%, while the percentage of Americans who profess no religious affiliation ("None") has almost doubled, from a little more than 8% to almost 16%. Also, the "Other" category has increased from 1.5% to almost 3%. Protestantism remains the modal religion but has lost much of its numerical predominance as the United States grows increasingly diverse.

Source: U.S. Bureau of the Census. 2012. *Statistical Abstract of the United States:* 2012. p. 61. Retrieved from http://www.census.gov/prod/2011pubs/12statab/pop.pdf

The Median

Unlike the mode, the **median (*Md*)** is always at the exact center of a distribution of scores. The median is the score of the case that is at the middle of a distribution: Half the cases have scores higher than the median and half the cases have scores lower than the median. Thus, if the median family income for a community is $52,000, half the families earn more than $52,000 and half earn less.

Before finding the median, you must place the cases in order from the highest to the lowest score—or from the lowest to the highest. Then you can determine the median by locating the case that divides the distribution into two equal halves. The median is the score of the middle case. If five students received grades of 93, 87, 80, 75, and 61 on a test, the median would be 80, the score that splits the distribution into two equal halves.

When the number of cases (N) is odd, the value of the median is unambiguous because there will always be a middle case. With an even number of cases, however, there will be two middle cases and, in this situation, the median is defined as the score exactly halfway between the scores of the two middle cases.

To illustrate, assume that seven students were asked to indicate their level of support for intercollegiate athletics at their universities on a scale ranging from 10 (great support) to 0 (no support). After arranging their responses from high to low, you can find the median by locating the case that divides the distribution into two equal halves. With a total of seven cases, the middle case would be the fourth case, since there will be three cases above and three cases below this case. Table 3.4 lists the cases in order and identifies the median. With seven cases, the median is the score of the fourth case.

Now suppose we add another student, whose support for athletics was measured as a 1. This would make N an even number (8) and we would no longer have a single middle case. Table 3.5 presents the new distribution of scores, and, as you can see, any value between 7 and 5 would technically satisfy the definition of a median (that is, it would split the distribution into two equal halves of four cases each). We resolve this ambiguity by defining the median as the average of the scores of the two middle cases. In this example, the median would be defined as $(7 + 5)/2$, or 6.

TABLE 3.4 Finding the Median with Seven Cases (*N* is odd)

Case	Score	
A	10	
B	10	
C	8	
D	7	← Md
E	5	
F	4	
G	2	

TABLE 3.5 Finding the Median with Eight Cases (*N* is even)

Case	Score	
A	10	
B	10	
C	8	
D	7	
		← Md = (7 + 5)/2 = 6
E	5	
F	4	
G	2	
H	1	

Remember that we follow different procedures to find the median, depending on whether *N* is odd or even. The procedures are stated in general terms in the "One Step at a Time" box.

The median cannot be calculated for variables measured at the nominal level because it requires that scores be ranked from high to low, and nominal-level variables cannot be ordered or ranked. The scores of nominal-level variables are different from each other but do not form a mathematical scale of any sort. The median can be found for either ordinal or interval-ratio data but is generally more appropriate for the former. *(The median may be found for any problem at the end of this chapter.)*

ONE STEP AT A TIME Finding the Median

Step ***Operation***

1. Array the scores in order from high score to low score.
2. Count the number of scores to determine whether *N* is odd or even.

IF *N* is ODD
⇓

3. The median will be the score of the middle case.
4. To find the middle case, add 1 to *N* and divide by 2.
5. The value you calculated in step 4 is the number of the middle case. The median is the score of this case. For example, if *N* = 13, the median will be the score of the (13 + 1)/2, or seventh, case.

IF *N* is EVEN
⇓

3. The median is halfway between the scores of the two middle cases.
4. To find the first middle case, divide *N* by 2.
5. To find the second middle case, increase the value you computed in step 4 by 1.
6. Add the scores of the two middle cases together and divide by 2. The result is the median. For example, if *N* = 14, the median is the score halfway between the scores of the seventh and eighth cases. If the middle cases have the same score, that score is defined as the median.

The Mean

The **mean** ($\bar{X}$—read this as "ex-bar"),[1] or arithmetic average, is by far the most commonly used measure of central tendency. It reports the average score of a distribution, and it is calculated by dividing the sum of the scores by the number of scores (N).

To illustrate: A birth control clinic administered a 20-item test of general knowledge about contraception to 10 clients. The numbers of correct responses were 2, 10, 15, 11, 9, 16, 18, 10, 11, and 7. To find the mean of this distribution, add the scores (total = 109) and divide by the number of scores (10). The result (10.9) is the average score on the test.

The mathematical formula for the mean is

FORMULA 3.1

$$\bar{X} = \frac{\Sigma(X_i)}{N}$$

Where $\bar{X}$ = the mean
$\Sigma(X_i)$ = the summation of the scores
N = the number of cases

Let's take a moment to consider the new symbols introduced in this formula. First, the symbol Σ (uppercase Greek letter sigma) is a mathematical operator just like the plus sign (+) or divide sign (÷). It stands for "the summation of" and directs us to add whatever quantities are stated immediately following it.[2]

The second new symbol is X_i ("*X*-sub-*i*"), which refers to any single score—the "*i*th" score. If we wish to refer to the score of a particular case, we replace the subscript with the number of the case. Thus, X_1 would refer to the score of the first case, X_2 to the score of the second case, X_{26} to the 26th case, and so forth.

The operation of adding all the scores is symbolized as $\Sigma(X_i)$. This combination of symbols directs us to sum the scores, beginning with the first score and ending with the last score in the distribution. Thus, Formula 3.1 states in symbols what has already been stated in words (to calculate the mean, add the scores and divide by the number of scores), but in a succinct and precise way.

Because computation of the mean requires addition and division, it should be used with variables measured at the interval-ratio level. However, researchers do calculate the mean for variables measured at the ordinal level, because the mean is much more flexible than the median and is a central feature of many interesting and powerful advanced statistical techniques. Thus, if the researcher plans to do any more than merely describe his or her data, the mean will probably be the preferred measure of central tendency even for ordinal-level variables. *(For practice in computing the mean, see any of the problems at the end of this chapter.)*

[1]This is the symbol for the mean of a sample. The symbol for the mean of a population is Greek letter mu (μ pronounced "mew").
[2]See the Prologue (Basic Review of Math) for further information on the summation sign and on summation operations.

ONE STEP AT A TIME **Finding the Mean**

Step	*Operation*
1.	Add up the scores $\Sigma(X_i)$.
2.	Divide the quantity you found in step 1 by N.

Three Characteristics of the Mean

The mean is the most commonly used measure of central tendency, and we will consider its mathematical and statistical characteristics in some detail.

1. **The Mean Balances All the Scores.** First, the mean is an excellent measure of central tendency because it acts like a fulcrum that "balances" all the scores: The mean is the point around which all of the scores cancel out. We can express this property symbolically:

$$\Sigma(X_i - \bar{X}) = 0$$

This expression says that if we subtract the mean from each score (X_i) in a distribution and then sum the differences, the result will always be 0.

To illustrate, consider the test scores presented in Table 3.6. The mean of these five scores is 390/5, or 78. The difference between each score and the mean is listed in the right-hand column, and the sum of these differences is 0. The total of the negative differences (−19) is exactly equal to the total of the positive differences (+19), as will always be the case. Thus, the mean "balances" the scores and is at the center of the distribution.

2. **The Mean Minimizes the Variation of the Scores.** A second characteristic of the mean is called the "least-squares" principle, a characteristic that is expressed in the statement

$$\Sigma(X_i - \bar{X})^2 = \text{minimum}$$

or: The mean is the point in a distribution around which the variation of the scores (as indicated by the squared differences) is minimized. If the differences

TABLE 3.6 A Demonstration Showing That All Scores Cancel Out Around the Mean

X_i	$X_i - \bar{X}$
65	65 − 78 = −13
73	73 − 78 = −5
77	77 − 78 = −1
85	85 − 78 = 7
90	90 − 78 = 12
$\Sigma X_i = 390$	$\Sigma(X_i - \bar{X}) = 0$
$\bar{X} = 390/5 = 78$	

TABLE 3.7 A Demonstration Showing That the Mean Is the Point of Minimized Variation

1	2	3	4
X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$(X_i - 77)^2$
65	$65 - 78 = -13$	$(-13)^2 = 169$	$65 - 77 = (-12)^2 = 144$
73	$73 - 78 = -5$	$(-5)^2 = 25$	$73 - 77 = (-4)^2 = 16$
77	$77 - 78 = -1$	$(-1)^2 = 1$	$77 - 77 = (0)^2 = 0$
85	$85 - 78 = 7$	$(7)^2 = 49$	$85 - 77 = (8)^2 = 64$
90	$90 - 78 = 12$	$(12)^2 = 144$	$90 - 77 = (13)^2 = 169$
$\Sigma(X_i) = 390$	$\Sigma(X_i - \bar{X}) = 0$	$\Sigma(X_i - \bar{X})^2 = 388$	$\Sigma(X_i - 77)^2 = 393$

between the scores and the mean are squared and then added, the resultant sum will be less than the sum of the squared differences between the scores and any other point in the distribution.

To illustrate this principle, consider Table 3.7. Column 1 of the table shows the same five scores displayed in Table 3.6, and column 2 displays the differences between the scores and the mean. In column 3, the differences between the scores and the mean are squared, and the sum of these differences is 388.

If we performed those same mathematical operations with any number other than the mean—say, 77—the result would be a sum greater than 388. This point is illustrated in column 4 of Table 3.7, which shows that the sum of the squared differences around 77 is 393, a value greater than 388.

This least-squares principle underlines the fact that the mean is closer to all of the scores than the other measures of central tendency. Also, this characteristic of the mean is important for the statistical techniques of correlation and regression, topics we take up toward the end of this text.

3. **The Mean Is Affected by All Scores and Can Be Misleading if the Distribution Has "Outliers."** The final important characteristic of the mean is that all the scores in a distribution are included in its calculation ("to find the mean, add up *all* the scores and divide by *N*"). In contrast, the mode reflects only the most common score, and the median deals only with the score of the middle case.

   On one hand, this characteristic is an advantage because the mean utilizes all the available information. On the other hand, when a distribution has **outliers** or some extremely high or low scores, the mean may become misleading: It may not represent the central or typical score. Distributions with outliers have a **skew**. If there are some extremely high scores, this is called a *positive skew,* and distributions with a few very low scores have a *negative skew*.

   The point to remember is that the mean will be pulled in the direction of the outlying scores relative to the median. With a positive skew, the mean will be greater in value than the median, and just the opposite will occur with a negative skew.

   Why is this problematic? Because the median uses *only* the middle cases, it will *always* reflect the center of the distribution. The mean, because it uses *all* cases (including outliers), may be much higher or lower than the bulk of the scores and give a false impression of centrality.

TABLE 3.8 A Demonstration Showing That the Mean Is Affected by Every Score

1	2	3	4	5	6
Scores	Measures of Central Tendency	Scores	Measures of Central Tendency	Scores	Measures of Central Tendency
15		15		0	
20	Mean = 25	20	Mean = 718	20	Mean = 22
25	Median = 25	25	Median = 25	25	Median = 25
30		30		30	
35		3500		35	

To illustrate, consider Table 3.8. The five scores in column 1 are symmetrical or unskewed: They are equidistant from each other, and both the high and low scores are an equal distance from the central score of 25. Because these scores have no skew, the median and mean are the same value (see column 2).

In column 3, an extremely high score (3500) has been added, and the distribution now has a positive skew. In column 4, we see that this very high score has no effect on the median; it remains at 25. This is because the median is based *only* on the score of the middle case and is not affected by changes in the scores of other cases. The mean, because it takes *all* scores into account, is very much affected. The mean changes from 25 to 718 solely because of the one extremely high score of 3500.

Finally, the scores in column 5 are the same as those in column 1, except the low score is changed to 0, thus creating a negative skew. Column 6 shows that the median stays at 25 but, in response to the lower score, the mean drops to 22.

We should make a few more points about Table 3.8:

1. The mean and median will have the same value only in distributions that are symmetrical, like the distribution in column 1.
2. Note that the mean in column 4 is very different from the five scores listed in column 3. In this case, is the mean or the median a better representation of the centrality of the scores? When a distribution has a pronounced skew, the median may be the preferred measure of central tendency even for interval-ratio variables.
3. Compare the means and medians in column 4 and column 6. In column 4, the mean is very different from the median, reflecting the fact that the "outlier" or extreme score of 3500 is very different from the other scores. The difference between the mean and median in column 6 is much less because the skew is not as extreme.

The general principle, again, is that the mean is always pulled in the direction of extreme scores relative to the median. The greater the skew (whether negative or positive), the greater will be the difference between the two measures of central tendency. The mean and median will have the same value when and only when a distribution is symmetrical. Figures 3.1 to 3.3 depict three different line charts that demonstrate these relationships.

FIGURE 3.1 A Positively Skewed Distribution (The mean is greater in value than the median.)

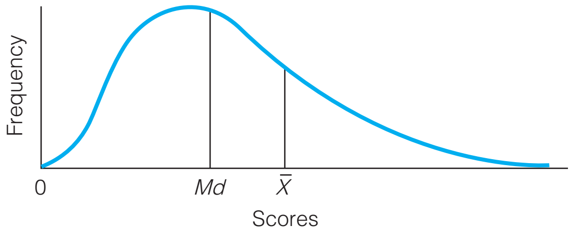


FIGURE 3.2 A Negatively Skewed Distribution (The mean is less than the median.)

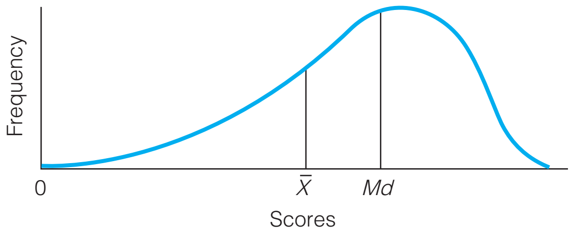
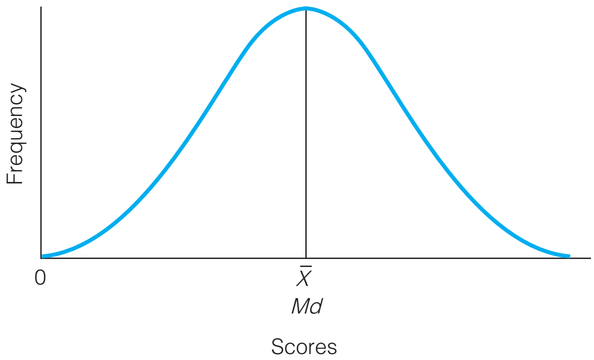


FIGURE 3.3 An Unskewed, Symmetrical Distribution (The mean and median are equal.)



These relationships between medians and means have a practical value. For one thing, a quick comparison of the median and mean will always tell you whether a distribution is skewed and the direction of the skew. If the mean is less than the median, the distribution has a negative skew. If the mean is greater than the median, the distribution has a positive skew.

Second, these characteristics of the mean and median also provide a simple and effective way to "lie" with statistics. For example, if you want to maximize the average score of a positively skewed distribution, report the mean. Income data almost always have a positive skew (there are only a few very wealthy people). If you want to impress someone with the general affluence of a mixed-income community, report the mean. If you want a lower figure, report the median.

STATISTICS
IN EVERYDAY LIFE

Means, Medians, and Baseball Salaries*

Major League Baseball publishes the salaries of every team, and this information can create some interesting comparisons. For example, compare the New York Yankees and the Boston Red Sox, perhaps the most intense rivals in baseball and one of the most intense in all of sports. The accompanying table displays some payroll statistics for each team on opening day of the 2014 season.

Note that, for both teams, the mean salary is much higher than the median salary. Also note that the median salaries are much closer together (there is a difference of about $215,000) than the means (a difference of $1,264,492).

What can you tell about the payrolls of the teams from these comparisons? Can you detect a skew and, if so, in what direction? Is the skew greater for one of the two teams? How do you know? What could cause this difference?

Here's a hint and some further information. The lowest salaries on both teams were about $500,000, the major league minimum. The range from the highest to lowest salary was about $24 million for the Yankees and about $16 million for the Red Sox.

Summary Statistics on the Opening-Day Payrolls of Two Teams

Statistic	New York Yankees	Boston Red Sox
Median	$2,000,000	$2,214,847
Mean	$6,479,339	$5,214,847

*I am grateful to Professor John Shandra for suggesting this comparison.

Which measure of central tendency is more appropriate for skewed distributions? This will depend on what point the researcher wishes to make but, as a rule, either both statistics or the median alone should be reported when the distribution has a few extreme scores. *(For practice in dealing with the effects of extreme scores on means and medians, see problems 3.4, 3.11, 3.13, and 3.15.)*

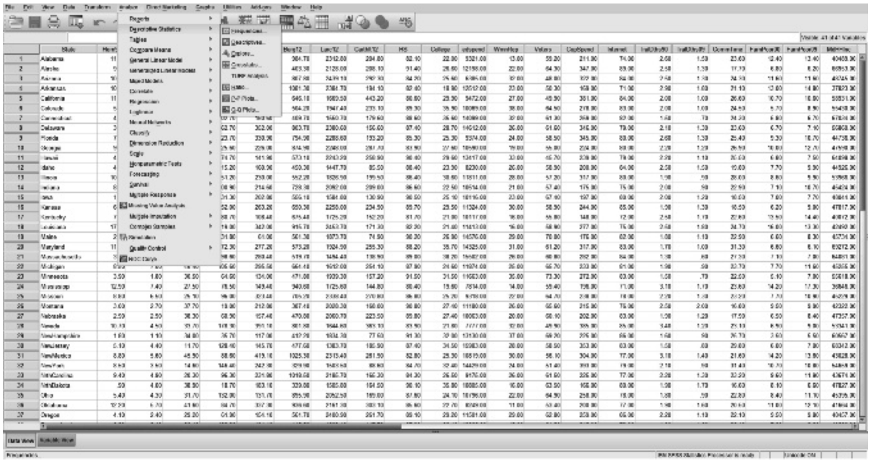
Using SPSS to Produce Measures of Central Tendency

In this installment of "Using SPSS," we will get measures of central tendency for the homicide rates of the 50 states in two different years. Our database will be the *States* file. We will continue to use **Frequencies**, the only SPSS procedure that will calculate all three measures of central tendency.

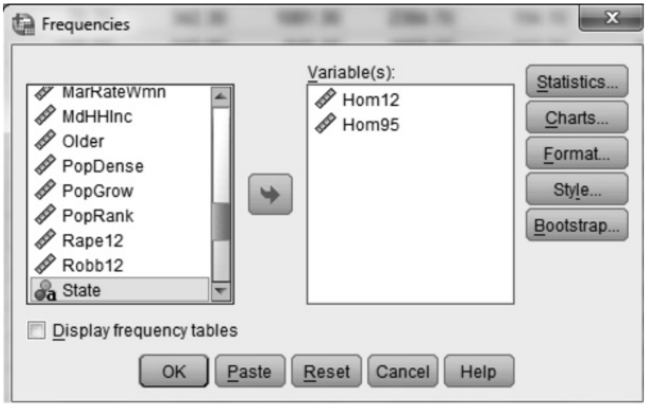
Follow these steps:

1. Click the SPSS icon on your desktop.
2. Load the *States* database.
   - **a.** Find the **File** command on the far left of the menu bar and then click **File → Open → Data.**
   - **b.** Find the *States* database. You can download this file from the website for this text if you haven't already.

3. From the menu bar across the top of the SPSS window, click **Analyze**, **Descriptive Statistics**, and **Frequencies**.



4. Find the variables *Hom95* (homicide rates in 1995) and *Hom12* (homicide rates in 2012) in the box on the left of the “Frequencies” window, and click the arrow to move the variable names to the box on the right. When you are done, the screen will look like this:



5. Make sure you click the box next to the “Display frequency tables” option under the variable list window to *not* select this option. We are interested only in the summary statistics, not in the tables.
6. Click **Statistics** and check the boxes next to Mean, Median, and Mode. Click **Continue**.
7. Click **OK**.

The output will look like this:

Statistics

		Homicide rate 2012	Homicide rate 1995
N	Valid	50	50
	Missing	0	0
Mean		4.2600	6.7120
Median		4.2000	6.2500
Mode[a]		1.80	1.80

[a]Multiple modes exist. The smallest value is shown.

The note under the output table tells us that there are multiple modes for both years, so this statistic is not very useful. For 1995, the mean (6.71) is higher than the median (6.25), indicating a positive skew in the data. If we inspected the state scores for 1995, we would see that Louisiana is the "outlier" with a much higher homicide rate (17.00) than the other states.

The homicide rate for 2012 is lower, as shown by both the mean and median. If we inspected the individual scores, we would see that the homicide rate of virtually every state was lower in 2012 than in 1995 and this accounts for the lower measures of central tendency. Comparing the mean (4.26) and median (4.20) for 2012 tells us that there is a slight positive skew in the distribution. As in 1995, Louisiana remains the outlier: its homicide rate of 10.8 was 3.2 percentage points higher than the next highest state.

Overall, we can conclude that the homicide rate fell between 1995 and 2012 and that there was a less extreme positive skew in the latter year. We will defer further discussion of these variables until we have introduced the concept of dispersion in Chapter 4.

Choosing a Measure of Central Tendency

You should consider two main criteria when choosing a measure of central tendency. First, make sure that you know the level of measurement of the variable in question. This will generally tell you whether you should report the mode, median, or mean.

Table 3.9 shows the relationship between the level of measurement and measures of central tendency. The capitalized, boldface "**YES**" identifies the most appropriate measure of central tendency for each level of measurement and the non-boldface "Yes" indicates the levels of measurement for which the measure is also permitted. An entry of "No" in the table means that the statistic cannot

TABLE 3.9 The Relationship Between Level of Measurement and Measures of Central Tendency

Measure of Central Tendency	Level of Measurement		
	Nominal	Ordinal	Interval-Ratio
Mode	**YES**	Yes	Yes
Median	No	**YES**	Yes
Mean	No	Yes (?)	**YES**

TABLE 3.10 Choosing a Measure of Central Tendency

Use the mode when:	**1**	The variable is measured at the nominal level.
	2	You want a quick and easy measure for ordinal and interval-ratio variables.
	3	You want to report the most common score.
Use the median when:	**1**	The variable is measured at the ordinal level.
	2	An interval-ratio variable is badly skewed.
	3	You want to report the central score. The median always lies at the exact center of the distribution.
Use the mean when:	**1**	The variable is measured at the interval-ratio level (except when the variable is badly skewed).
	2	You want to report the typical score. The mean is the "fulcrum that exactly balances all of the scores."
	3	You anticipate additional statistical analysis.

be computed for that level of measurement. Finally, the "Yes (?)" entry in the bottom row indicates that the mean is often used with ordinal-level variables even though, strictly speaking, this practice violates level-of-measurement guidelines.

Second, consider the definitions of the three measures of central tendency and remember that they provide different types of information. They will be the same value only under certain specific conditions (namely, for symmetrical distributions with one mode), and each has its own message to report. In many circumstances, you might want to report all three.

The guidelines in Table 3.10 stress both selection criteria and may be helpful when choosing a specific measure of central tendency.

Applying Statistics 3.1 The Mean, Mode, and Median

Ten students have been asked how many hours they spent in the college library during the past week. What is the average "library time" for these students? The hours are reported in the following list, and we will find the mode, the median, and the mean for these data.

Student	Hours Spent in the Library Last Week (X_i)
A	0
B	2
C	5
D	5
E	7
F	10
G	14
H	14
I	20
J	30
	$(\Sigma X_i) = 107$

By scanning the scores, we can see that two scores, 5 and 14, occurred twice, and no other score occurred more than once. This distribution has two modes, 5 and 14.

Because the number of cases is even, the median will be the average of the two middle cases after all cases have been ranked in order. With 10 cases, the first middle case will be the (N/2), or (10/2), or fifth case. The second middle case is the (N/2) + 1, or (10/2) + 1, or sixth case. The median will be the score halfway between the scores of the fifth and sixth cases. Counting down from the top, we find that the score of the fifth case (student E) is 7 and the score of the sixth case (student F) is 10. The median for these

(*continued*)

data is (7 + 10)/2, or (17/2), or 8.5. The median, the score that divides this distribution in half, is 8.5.

The mean is found by first adding all the scores and then dividing by the number of scores. The sum of the scores is 107, so the mean is

$$\bar{X} = \frac{\Sigma(X_i)}{N} = \frac{107}{10} = 10.7$$

These 10 students spent an average of 10.7 hours in the library during the week in question.

Note that the mean is higher than the median. This indicates a positive skew in the distribution (a few extremely high scores). By inspection, we can see that the positive skew is caused by the two students who spent many more hours (20 and 30 hours) in the library than the other students.

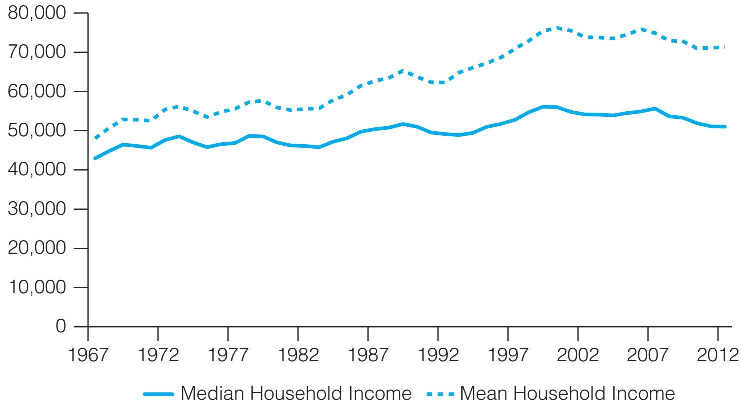
STATISTICS IN EVERYDAY LIFE

Assessing America's Financial Health

Are Americans better off financially today? One way to begin to answer this question is to look at median and mean incomes over time. Is the average (mean) income rising or falling? What about the income of the typical American (the median)?

The graph below presents some answers to these questions. Note that income is expressed in 2012 dollars to eliminate the effect of inflation. Without this adjustment, recent incomes would appear to be higher than older incomes, not because of increased buying power and well-being, but rather because of the changing value of the dollar. Also, the graph is based on total income for entire households, not individual income.

Mean and Median Household Incomes, U.S., 1967–2012 (in 2012 Dollars)



The median shows that the income of the average American household increased from a little over $40,000 in 1967 to about $50,000 in the most recent year. The line representing the median rises and falls over the years, reaching its highest value in the late 1990s and falling since the recession of 2007. In spite of the recent decline, the median indicates that the standard of living increased for the average American household over this time period.

The mean income rises and falls in almost exactly the same way as the median. Notice, however, that the mean is always higher than the median, a reflection of the characteristic positive skew of income data. Also notice that the size of the gap between the mean and the median increases over time, indicating that the degree of skew is increasing. This is because the income of the affluent is increasing relative to the income of the typical American household.

These patterns show that the typical American household is financially better off in 2012 than in 1967 but also suggests that gains in income are going disproportionately to the more affluent. We will return to this topic in Chapter 4.

SUMMARY

1. The three measures of central tendency presented in this chapter share a common purpose. Each reports some information about the most typical or representative value in a distribution. Appropriate use of these statistics permits the researcher to report important information about an entire distribution of scores in a single, easily understood number.
2. The mode reports the most common score and is used most appropriately with nominal-level variables.
3. The median (*Md*) reports the score that is the exact center of the distribution. It is most appropriately used with variables measured at the ordinal level and with interval-ratio variables when the distribution is skewed.
4. The mean, the most frequently used of the three measures, reports the most typical score. It is used most appropriately with variables measured at the interval-ratio level (except when the distribution is highly skewed).
5. The mean has a number of important mathematical characteristics. First, it is the point in a distribution of scores around which all other scores cancel out. Second, the mean is the point of minimized variation (this is the "least-squares" principle). Last, the mean is affected by every score in the distribution and is therefore pulled in the direction of extreme scores.

SUMMARY OF FORMULAS

FORMULA 3.1 Mean $\bar{X} = \frac{\Sigma(X_i)}{N}$

GLOSSARY

Mean. The arithmetic average of the scores. $\bar{X}$ represents the mean of a sample, and μ, the mean of a population.

Measures of central tendency. Statistics that summarize a distribution of scores by reporting the most typical or representative value of the distribution.

Median (*Md*). The point in a distribution of scores above and below which exactly half of the cases fall.

Mode. The most common score in a distribution.

Outliers. Scores that are very high or very low compared to most scores.

Σ (uppercase Greek letter sigma). "The summation of."

Skew. The extent to which a distribution of scores has a few scores that are extremely high (positive skew) or extremely low (negative skew).

X_i ("*X*-sub-*i*"). Any score in a distribution.

PROBLEMS

3.1 SOC A variety of information has been gathered from a sample of college freshmen and seniors, including

- Their region of birth;
- The extent to which they support legalization of marijuana (measured on a scale on which 7 = strong support, 4 = neutral, and 1 = strong opposition);
- The amount of money they spend each week out-of-pocket for food, drinks, and entertainment;
- How many movies they watched in their dorm rooms last week;
- Their opinion of cafeteria food (10 = excellent, 0 = very bad); and
- Their religious affiliation.

Some results are presented here. Find the *most appropriate* measure of central tendency for each variable for freshmen and then for seniors. Report both the measure you selected and its value for each variable (e.g., "Mode = 3" or "Median = 3.5"). (*HINT: Determine the level of measurement for each variable first. In general, this will tell you which measure of central tendency is appropriate. See the section "Choosing a Measure of Central Tendency" to review the relationship between measure of central tendency and level of measurement. Also, remember that the mode is the most common score, and especially remember to array scores from high to low before finding the median.*)

FRESHMEN

Student	Region of Birth	Legalization	Out-of- Pocket Expenses	Movies	Cafeteria Food	Religion
A	North	7	43	0	10	Protestant
B	North	4	49	14	7	Protestant
C	South	3	55	10	2	Catholic
D	Midwest	2	57	7	1	None
E	North	3	72	5	8	Protestant
F	North	5	58	1	6	Jew
G	South	1	62	0	10	Protestant
H	South	4	75	14	0	Other
I	Midwest	1	61	3	5	Other
J	West	2	53	4	6	Catholic

SENIORS

Student	Region of Birth	Legalization	Out-of- Pocket Expenses	Movies	Cafeteria Food	Religion
K	North	7	75	0	1	None
L	Midwest	6	72	5	2	Protestant
M	North	7	70	11	8	Protestant
N	North	5	95	3	4	Catholic
O	South	1	72	4	3	Protestant
P	South	5	67	14	6	Protestant
Q	West	6	50	0	2	Catholic
R	West	7	59	7	9	None
S	North	3	55	5	4	None
T	West	5	95	3	7	Other
U	North	4	88	5	4	None

3.2 A variety of information has been collected for all district high schools. Find the most appropriate measure of central tendency for each variable, and summarize this information in a paragraph. (*HINT: The level of measurement of the variable will generally tell you which measure of central tendency is appropriate. Remember to organize the scores from high to low before finding the median.*)

High School	Enrollment	Largest Racial/ Ethnic Group	% College Bound	Most Popular Sport	Condition of Physical Plant (10 = Excellent)
A	1400	White	25	Football	10
B	1223	White	77	Baseball	7
C	876	Black	52	Football	5
D	1567	Hispanic	29	Football	8
E	778	White	43	Basketball	4
F	1690	Black	35	Basketball	5
G	1250	White	66	Soccer	6
H	970	White	54	Football	9

3.3 *PS* You have been observing the local Democratic Party in a large city and have compiled some information about a small sample of party regulars. Find the appropriate measure of central tendency for each variable.

Respondent	Gender	Social Class	No. of Years in Party	Education	Marital Status	No. of Children
A	M	High	32	High school	Married	5
B	M	Middle	17	High school	Married	0
C	M	Working	32	High school	Single	0
D	M	Working	50	8th grade	Widowed	7
E	M	Working	25	4th grade	Married	4
F	M	Middle	25	High school	Divorced	3
G	F	High	12	College	Divorced	3
H	F	High	10	College	Separated	2
I	F	Middle	21	College	Married	1
J	F	Middle	33	College	Married	5
K	M	Working	37	High school	Single	0
L	F	Working	15	High school	Divorced	0
M	F	Working	31	8th grade	Widowed	1

3.4 *PA* The following table presents the annual person-hours of time lost due to traffic congestion for a group of cities for 2009. This statistic is a measure of traffic congestion.

City	Annual Person-Hours Lost to Traffic Congestion
Baltimore	33
Boston	28
Buffalo	11
Chicago	44
Cleveland	13
Dallas	32
Detroit	23
Houston	37
Kansas City	14
Los Angeles	40
Miami	26
Minneapolis	27
New Orleans	20
New York	24
Philadelphia	26
Phoenix	23
Pittsburgh	23
San Antonio	19
San Diego	23
San Francisco	30
Seattle	27
Washington, D.C.	41

(*continued next column*)

(*continued*)

Source: U.S. Bureau of the Census. 2012. *Statistical Abstract of the United States: 2012.* Table 1099. Available at http://www.census.gov/prod/2011pubs/12statab/trans.pdf

a. Calculate the mean and median of this distribution.
b. Compare the mean and median. Which is the higher value? Why?
c. If you removed Chicago (the case with the highest score) from this distribution and recalculated, what would happen to the mean? To the median? Why?
d. Report the mean and median as you would in a formal research report.

3.5 *SOC* Data have been gathered on four variables for 15 respondents (see the following table). Find and report the appropriate measure of central tendency for each variable. (On the "Attitude on Abortion" scale, a high score indicates strong opposition.)

Respondent	Marital Status	Racial/Ethnic Group	Age	Attitude on Abortion
A	Single	White	18	10
B	Single	Hispanic	20	9

(*continued*)

Respondent	Marital Status	Racial/Ethnic Group	Age	Attitude on Abortion
C	Widowed	White	21	8
D	Married	White	30	10
E	Married	Hispanic	25	7
F	Married	White	26	7
G	Divorced	Black	19	9
H	Widowed	White	29	6
I	Divorced	White	31	10
J	Married	Black	55	5
K	Widowed	Asian American	32	4
L	Married	Native American	28	3
M	Divorced	White	23	2
N	Married	White	24	1
O	Divorced	Black	32	9

3.6 *SOC* The following tables list the median family incomes for the 13 Canadian provinces and territories in 2000 and 2011 and for 13 states of the United States in 1999 and 2012. For the provinces and the states, compute the mean and median family income for each year and compare the two measures of central tendency. Which measure of central tendency is greater for each year? Are the distributions skewed? In which direction?

Median Income for Canadian Provinces and Territories, 2000 and 2011 (Canadian dollars)

Province or Territory	2000	2011
Newfoundland and Labrador	38,800	67,200
Prince Edward Island	44,200	66,500
Nova Scotia	44,500	66,300
New Brunswick	43,200	63,930
Quebec	47,700	68,170
Ontario	55,700	73,290
Manitoba	47,300	68,710
Saskatchewan	45,800	77,300
Alberta	55,200	89,930
British Columbia	49,100	69,150
Yukon	56,000	90,090
Northwest Territories	61,000	105,560
Nunavut	37,600	65,280
Mean =		
Median =		

Source: Statistics Canada. Retrieved from http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/famil108a-eng.htm

Median Income for Thirteen States, 1999 and 2012 (U.S. dollars)

State	1999	2012
Alabama	36,213	43,464
Alaska	51,509	63,348
Arkansas	29,762	39,018
California	43,744	57,020
Connecticut	50,798	64,247
Illinois	46,392	51,738
Kansas	37,476	50,003
Maryland	52,310	71,836
Michigan	46,238	50,015
New York	40,058	47,680
Ohio	39,617	44,375
South Dakota	35,962	49,415
Texas	38,978	51,926
Mean =		
Median =		

Sources: 1999: U.S. Bureau of the Census. *Statistical Abstract of the United States: 2001.* p. 436. Retrieved from http://www.census.gov/prod/2002pubs/01statab/income.pdf. 2012: U.S. Bureau of the Census. *American Community Survey, 2012.* Retrieved from http://www.census.gov/hhes/www/income/data/statemedian/

3.7 *SOC* Find the appropriate measure of central tendency for each variable displayed in problem 2.5. Report each statistic as you would in a formal research report.

3.8 *SOC* The accompanying table gives scores on four variables for 30 cases from the General Social Survey:

- "Age" is reported in years.

- "Happiness" consists of answers to the question "Taken all together, would you say that you are
  1. very happy
  2. pretty happy
  3. not too happy?"
- "Number of Partners" refers to sex partners over the past five years:
  - A score of 0 to 4 is the actual number of partners.
  - A score of 5 means 5–10 partners.
  - A score of 6 means 11–20 partners.
  - A score of 7 means 21–100 partners.
  - A score of 8 means more than 100 partners.
- "Religion" is self-identified religious affiliation (Protestant, Catholic, Jew, Other, or None).

For each variable, find the appropriate measure of central tendency, and write a sentence reporting this statistical information as you would in a research report.

Respondent	Age	Happiness	Number of Partners	Religion
1	20	1	2	Protestant
2	32	1	1	Protestant
3	31	1	1	Catholic
4	34	2	5	Protestant
5	34	2	3	Protestant
6	31	3	0	Jew
7	35	1	4	None
8	42	1	3	Protestant
9	48	1	1	Catholic
10	27	2	1	None
11	41	1	1	Protestant
12	42	2	0	Other
13	29	1	8	None
14	28	1	1	Jew
15	47	2	1	Protestant
16	69	2	2	Catholic
17	44	1	4	Other
18	21	3	1	Protestant
19	33	2	1	None
20	56	1	2	Protestant
21	73	2	0	Catholic
22	31	1	1	Catholic
23	53	2	3	None
24	78	1	0	Protestant
25	47	2	3	Protestant
26	88	3	0	Catholic
27	43	1	2	Protestant
28	24	1	1	None
29	24	2	3	None
30	60	1	1	Protestant

3.9 *SOC* The college administration is considering a total ban on student automobiles. You have conducted a poll on this issue of 20 fellow students and 20 of the neighbors who live around the campus and have calculated scores for your respondents. On the scale you used, a high score indicates strong opposition to the proposed ban. The scores are presented here for both groups. Calculate an appropriate measure of central tendency and compare the two groups in a sentence or two.

Students		Neighbors	
10	11	0	7
10	9	1	6
10	8	0	0
10	11	1	3
9	8	7	4
10	11	11	0
9	7	0	0
5	1	1	10
5	2	10	9
0	10	10	0

3.10 *SW* As the head of a social services agency, you believe that your staff of 20 social workers is very much overworked compared with 10 years ago. The case loads for each worker are reported below for each of the two years in question. Has the average caseload increased? Which measure of central tendency is most appropriate to answer this question? Why?

2002		2012	
52	55	42	82
50	49	75	50
57	50	69	52
49	52	65	50
45	59	58	55
65	60	64	65
60	65	69	60
55	68	60	60
42	60	50	60
50	42	60	60

3.11 *SOC* The following table lists the approximate number of cars per 1000 population for 15 nations. Compute the mean and median for these data. Which is greater in value? Is there a positive skew in the data? How do you know?

Nation	Number of Motor Vehicles per 1000 Population (2009)
United States	786
Germany	588
Canada	608

(*continued*)

Nation	Number of Motor Vehicles per 1000 Population (2009)
Japan	588
Australia	703
Israel	330
Russia	271
Mexico	278
Tunisia	130
Bolivia	87
China	69
Niger	8
Kenya	25
Thailand	172
India	18
Mean =	
Median =	

Source: World Bank, 2014. Retrieved from http://data.worldbank.org/indicator/IS.VEH.NVEH.P3

3.12 SW Compute the median and mean for both the pretest and posttest for the test scores first presented in problem 2.6 and reproduced here. Interpret these statistics.

Case	Pretest	Posttest
A	8	12
B	7	13
C	10	12
D	15	19
E	10	8
F	10	17
G	3	12
H	10	11
I	5	7
J	15	12
K	13	20
L	4	5
M	10	15
N	8	11
O	12	20

3.13 SOC A sample of 25 freshmen at a major university completed a survey that measured their degree of racial prejudice (the higher the score, the greater the prejudice).

a. Compute the median and mean scores for these data.

10	43	30	30	45
40	12	40	42	35
45	25	10	33	50
42	32	38	11	47
22	26	37	38	10

b. These same 25 students completed the same survey during their senior year. Compute the median and mean for this second set of scores, and compare them to the earlier set. What happened?

10	45	35	27	50
35	10	50	40	30
40	10	10	37	10
40	15	30	20	43
23	25	30	40	10

3.14 SOC You have compiled the following information on each of the graduates voted "most likely to succeed" by a local high school for a 10-year period. For each variable, find the appropriate measure of central tendency.

Case	Present Income ($)	Marital Status	Owns a Home?	Years of Education Post-High School
A	104,000	Divorced	Yes	8
B	68,000	Divorced	No	4
C	54,000	Married	Yes	4
D	45,000	Married	No	4
E	40,000	Single	No	4
F	85,000	Separated	Yes	8
G	30,000	Married	No	3
H	27,000	Married	No	1
I	93,000	Married	Yes	6
J	48,000	Single	Yes	4

3.15 SOC Professional athletes are threatening to strike because they claim that they are underpaid. The team owners have released a statement that says, in part, "The average salary for players was $1.2 million last year." The players counter by issuing their own statement that says, in part, "The typical player earned only $753,000 last year." Is either side necessarily lying? If you were a sports reporter and had just read Chapter 3 of this text, what questions would you ask about these statistics?

Statistical Analysis Using SPSS

3.16 SOC In this exercise, you will use SPSS to get measures of central tendency for several different variables in the *States* data set, which is available from the website for this text.

- Find and click the SPSS icon on your desktop.
- Load the *States* data set.
- From the menu bar across the top of the SPSS window, click **Analyze**, **Descriptive Statistics**, and **Frequencies**.
- Find these variables in the box on the left of the "Frequencies" window: *FamPoor09* (percentage of all families in the state below the poverty line), *ForBorn* (percentage of the population in the state that was born outside the United States in 2009), and *TeenBirthRate* (births per 1000 females age 15–19, 2011). Click the arrow to move the variable names into the box on the right.
- Click the **Statistics** button and, under "Central Tendency," select **Mean**, **Median**, and **Mode**. Click **Continue** to return to the "Frequencies" window.
- Make sure the box next to the "Display frequency tables" option is *not* checked. We are interested only in the summary statistics, not in the tables.
- Click **OK**.

a. Write a paragraph analyzing and summarizing the three variables. Are any of the distributions skewed? In what direction?

b. Select one of the skewed variables and skim the scores of the variable in the SPSS Data Editor to find the outlier(s). That is, find the column in which the scores of the variable are stored and skim the scores looking for especially high scores (if the skew is positive) or low scores (if the skew is negative). What sociological factors might account for these very high or low scores?

3.17 SOC In this exercise, you will use SPSS to get measures of central tendency for several different variables in the *GSS2012* data set, which is available from the website for this text. In this data set, the cases are a representative sample (see Chapter 1) of the U.S. population.

- Find and click the SPSS icon on your desktop.
- Load the *GSS2012* data set.
- From the menu bar across the top of the SPSS window, click **Analyze**, **Descriptive Statistics**, and **Frequencies**.
- Find these variables in the box on the left of the "Frequencies" window: *closeblk*, *hrs1*, *partnrs5,* and *region*. Consult Appendix G for information on what each variable measures and the coding scheme for the scores. Click the arrow to move the variable names into the box on the right.
- Click the **Statistics** button and, under "Central Tendency," select **Mean**, **Median**, and **Mode**. Click **Continue** to return to the "Frequencies" window.
- Make sure the box next to the "Display frequency tables" option is *not* checked. We are interested only in the summary statistics, not in the tables.
- Click **OK**.

For each variable, select the *most appropriate* measure of central tendency (consult Tables 3.9 and 3.10) and write a summary sentence about the variable.

3.18 SOC In this exercise, you will use SPSS to get measures of central tendency for several different variables in the *Intl-POP* data set, which is available from the website for this text. This data set includes information on 99 nations.

- Find and click the SPSS icon on your desktop.
- Load the *Intl-POP* data set.
- From the menu bar across the top of the SPSS window, click **Analyze**, **Descriptive Statistics**, and **Frequencies**.
- Find these variables in the box on the left of the "Frequencies" window: *Corruption*, *GNIcap*, *IncLevel, Rights*, and *Urban*. Consult *Appendix G* for information on what each variable measures and, if applicable, the coding scheme for the scores. Click the arrow to move the variable names into the box on the right.
- Click the **Statistics** button and, under "Central Tendency," select **Mean**, **Median**, and **Mode**. Click **Continue** to return to the "Frequencies" window.
- Make sure the box next to the "Display frequency tables" option is *not* checked. We are interested only in the summary statistics, not in the tables.
- Click **OK**.

For each variable, select the *most appropriate* measure of central tendency (consult Tables 3.9 and 3.10) and write a summary sentence about the variable.

YOU ARE THE RESEARCHER

The Typical American

Is there such a thing as a "typical" American? In this exercise, you will develop a profile of the average American based on measures of central tendency for ten variables, chosen by you from the 2012 General Social Survey (*GSS2012*). Choose variables that you think are the most important in defining what it means to be a member of this society and then choose an *appropriate* measure of central tendency for each variable. Use this information to write a description of the typical American. We will also use this opportunity to introduce a new SPSS program.

Step 1: Choosing Your Variables

Scroll through the list of available variables either in *Appendix G* or by using the **Utilities Variables** command in SPSS to see the online codebook. Select ten variables that, in your view, are central to defining or describing the "typical American" and list them in the table below. *Select at least one variable from each of the three levels of measurement*.

Variable	SPSS Name	Explain exactly what this variable measures	Level of Measurement
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

Step 2: Getting the Statistics

Find and click the SPSS icon on your desktop. Load the 2012 GSS (*GSS2012*) data set.

Using the "Frequencies" Procedure for the Mode and Median

The only SPSS procedure that will produce all three measures of central tendency is **Frequencies**. In this step, you will use **Frequencies** to get modes and medians for any nominal- and ordinal-level variables you selected in step 1.

Click **Analyze → Descriptive Statistics → Frequencies**. In the "Frequencies" dialog box, find the names of your nominal- and ordinal-level variables in the list on the left and click the arrow button in the middle of the screen to move the names to the "Variables" box on the right.

Click the **Statistics** button, find the "Central Tendency" box on the right, and click **Median** and **Mode**. Click **Continue**, and you will be returned to the "Frequencies" dialog box. Click (uncheck) the "Display frequency tables" box so that SPSS will *not* produce frequency distribution tables. Click **OK** and record your results in the following table. Report the mode for all nominal-level variables and the median for ordinal-level variables. Use as many lines as needed.

Variable	SPSS Name	Level of Measurement	Mode	Median
1				
2				
3				
4				
5				
6				
7				
8				
9				

Using the "Descriptives" Procedure for the Mean

The **Descriptives** command in SPSS is designed to provide summary statistics for interval-ratio-level variables. By default (i.e., unless you tell it otherwise), **Descriptives** produces the mean, the minimum and maximum scores (i.e., the lowest and highest scores, which can be used to compute the range), and the standard deviation. We will introduce the range and standard deviation in the next chapter.

To use **Descriptives**, click **Analyze, Descriptive Statistics**, and **Descriptives**. The "Descriptives" dialog box will open. This dialog box looks just like the "Frequencies" dialog box and works in the same way. Find the names of your interval-ratio-level variables in the list on the left and, once they are highlighted, click the arrow button in the middle of the screen to transfer them to the "Variables" box on the right. Click **OK** and record your results in the table below, using as many lines as necessary.

Variable	SPSS Name	Mean
1		
2		
3		
4		
5		
6		
7		
8		
9		

Step 3: Interpreting Results

Examine the two tables displaying your results, and write a summary paragraph describing the typical American. Be sure to report all ten variables and, as appropriate, describe the most common case (the mode), the typical case (the median), or the typical score (mean). Write as if you are reporting in a newspaper: Your goal should be clarity and accuracy. As an example, you might report that "The typical American is Protestant and married."

4 Measures of Dispersion

LEARNING OBJECTIVES By the end of this chapter, you will be able to:

1. Explain the purpose of measures of dispersion and the information they convey.
2. Compute and explain the range (R), the interquartile range (Q), the standard deviation (s), and the variance (s^2).
3. Select an appropriate measure of dispersion and correctly calculate and interpret the statistic.
4. Describe and explain the mathematical characteristics of the standard deviation.
5. Analyze a boxplot.
6. Use SPSS to produce the standard deviation and range.

USING STATISTICS

The statistical techniques presented in this chapter are used to describe the variability or diversity in a set of scores. They can be used to describe the

- Variations in grades. Some students get the same grade virtually all the time while others vary from high to low grades.
- Diversity of lifestyles in different social settings. Larger cities generally support a wider variety of lifestyles than small towns.
- Differences in racial and ethnic diversity from place to place. Some states (California and New York, for example) are home to many different racial, cultural, and language groups while others (Iowa and Maine) are much less diverse.
- Variations in income inequality from nation to nation or from time to time. Some nations have vast differences in wealth and income from the very richest to the poorest members while the distance is smaller in other nations.

In Chapters 2 and 3, you learned a variety of ways to describe a variable, including frequency distributions, graphs, and measures of central tendency. For a complete description of a distribution of scores, we must combine these with **measures of dispersion**, the subject of this chapter. While measures of central tendency describe the typical, average, or central score, measures of dispersion describe the variety, diversity, or heterogeneity of a distribution of scores.

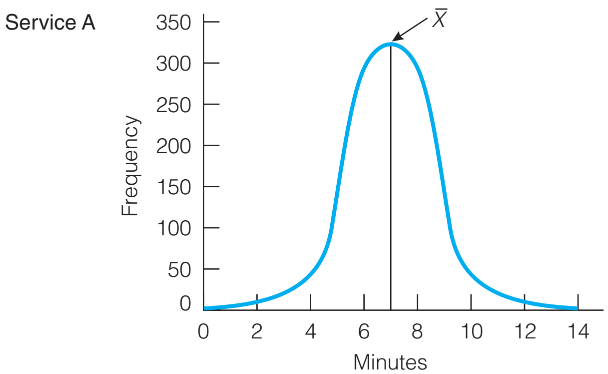
The importance of the concept of **dispersion** might be easier to grasp with the aid of an example. Suppose that the director of public safety wants to evaluate two ambulance services that have contracted with the city to provide emergency medical aid. As a part of the investigation, she has collected data showing that the mean response time to calls for assistance is 7.4 minutes for Service A and 7.6 minutes for Service B.

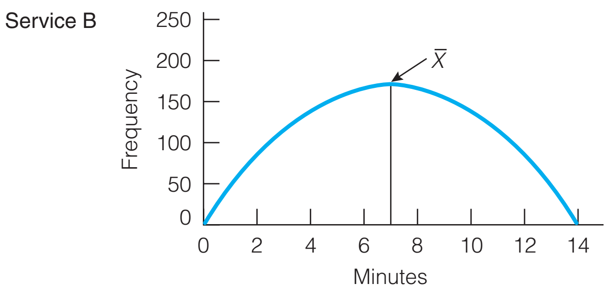
These averages are virtually identical and provide no clear basis for judging one service as more efficient than the other. Measures of dispersion, however, can reveal substantial differences between distributions even when the measures of central tendency are the same. For example, consider the graphs in Figure 4.1, which display the distribution of response times for the two services.

Note that the line chart for Service B is much flatter than that for Service A. This is because the scores for Service B are more spread out, or more diverse, than the scores for Service A. In other words, Service B was much more variable in response time and had more scores in the high *and* low ranges and fewer in the middle. Service A was more consistent in its response time, and its scores are more clustered, or grouped, around the mean. Both distributions have essentially the same *average* response time, but there is considerably more *variation*, or dispersion, in the response times for Service B.

If you were the director of public safety, would you be more likely to select an ambulance service that was always on the scene of an emergency in about the same amount of time (Service A) or one that was sometimes very slow and sometimes very quick to respond (Service B)? Note that if you had not considered dispersion,

FIGURE 4.1 Response Time for Two Ambulance Services





a possibly important difference in the performance of the two ambulance services might have gone unnoticed.

Keep the two shapes in Figure 4.1 in mind as visual representations of the concept of dispersion. The greater clustering of scores around the mean in the distribution for Service A indicates *less* dispersion, and the flatter curve of the distribution for Service B indicates *more* variety, or dispersion. Any measure of dispersion will decrease as the scores become less variable and the distribution becomes more peaked (that is, as the distribution looks more and more like Service A's) and increase as the scores become more variable and the distribution becomes flatter (that is, as the distribution looks more and more like Service B's).

These ideas and Figure 4.1 may give you a general notion of what is meant by dispersion, but the concept is not easily described in words alone: We need to consider some statistics that are designed to quantify the variation in a set of scores. In this chapter, we introduce some of the more common measures of dispersion. We begin with the range and the interquartile range but devote most of our attention to the standard deviation. We also consider a visual representation of dispersion called a boxplot.

The Range (*R*) and Interquartile Range (*Q*)

The **range (*R*)** is defined as the distance between the highest and lowest scores in a distribution:

FORMULA 4.1

$$\boldsymbol{R} = \text{High score} - \text{Low score}$$

The range is easy to calculate and is perhaps most useful as a quick and general indicator of variability. The statistic is also easy to interpret: The greater the value of the range, the greater the distance from high to low score, and the greater the dispersion in the distribution.

Unfortunately, because it is based on only the highest and lowest scores, the range has some important limitations. First, almost any sizeable distribution will contain some scores that are atypically high and/or low (outliers) compared to most of the scores (for example, see Table 3.3). Thus, *R* might exaggerate the amount of dispersion for most of the scores in the distribution. Also, *R* yields no information about the variation of the scores between the highest and lowest scores.

The **interquartile range (*Q*)** is a kind of range. It avoids some of the problems associated with *R* by considering only the middle 50% of the cases in a distribution. To find *Q*, first arrange the scores from highest to lowest and then divide the distribution into quarters (as distinct from halves when locating the median). The first quartile (Q_1) is the point below which 25% of the cases fall and above which 75% of the cases fall. The second quartile (Q_2) divides the distribution into halves (thus, Q_2 is equal to the median). The third quartile (Q_3) is the point below which 75% of the cases fall and above which 25% of the cases fall.

The interquartile range is defined as the distance from the third to the first quartile:

FORMULA 4.2

$$\boldsymbol{Q} = Q_3 - Q_1$$

The interquartile range essentially extracts the middle 50% of the distribution and, like *R*, is based on only two scores. *Q* is interpreted in the same way as *R*: The greater its value, the greater the dispersion. Unlike the range, *Q* avoids the problem of being based on the most extreme scores, but it also fails to yield any information about the variation of the scores other than the two on which it is based.

Computing the Range and Interquartile Range

Table 4.1 presents birth rates (number of births per 1000 population) for a group of 40 nations. What are the range and interquartile range of these data? Note that the scores have already been ordered from high to low. This makes the range easy to calculate and is necessary for finding the interquartile range. Of these 40 nations, Niger ranked the highest with a birth rate of 50 and Germany and Japan ranked the lowest with a birth rate of 8. The range is 50 − 8 or 42 ($R = 42$).

To find *Q*, we must locate the first and third quartiles (Q_1 and Q_3). We define these points in terms of the scores associated with certain cases, as we did when finding the median. Q_1 is determined by multiplying *N* by 0.25. Because $(40) \times (0.25) = 10$, Q_1 is the score associated with the tenth case, counting up from the lowest score. The tenth case is Iceland, with a score of 14. So $Q_1 = 14$.

TABLE 4.1 Birth Rates (Number of Births per 1000 Population) for 40 Nations, 2013

Rank	Nation	Birth Rate	Rank	Nation	Birth Rate
40 (High)	Niger	50	20	Libya	22
39	Angola	47	19	India	22
38	Uganda	45	18	Venezuela	21
37	Mozambique	44	17	Mexico	19
36	Nigeria	42	16	Colombia	19
35	Malawi	40	15	Kuwait	19
34	Tanzania	40	14	Australia	18
33	Guinea	38	13	Vietnam	17
32	Senegal	38	12	Ireland	16
31	Togo	37	11	Chile	15
30	Kenya	36	10	Iceland	14
29	Mauritania	35	9	United States	13
28	Ethiopia	34	8	Russia	13
27	Ghana	33	7	France	13
26	Guatemala	32	6	China	12
25	Pakistan	30	5	Canada	11
24	Haiti	26	4	Spain	10
23	Cambodia	25	3	Italy	9
22	Egypt	25	2	Japan	8
21	Syria	25	1 (Low)	Germany	8

The case that lies at the third quartile (Q_3) is given by multiplying N by 0.75, and $(40) \times (0.75) = 30$th case. The 30th case, again counting up from the lowest score, is Kenya, with a score of 36 ($Q_3 = 36$). Therefore,

$$Q = Q_3 - Q_1$$
$$Q = 36 - 14$$
$$Q = 22$$

(For practice in finding and interpreting R and Q, see problems 4.12 and 4.13. The range may be found for any of the problems at the end of this chapter.)

ONE STEP AT A TIME Finding the Interquartile Range (Q)

Step	***Operation***
1.	Array the cases in order.
2.	Find the case that lies at the first quartile (Q_1) by multiplying N by 0.25. If the result is not a whole number, round off to the nearest whole number. This is the number of the case that marks the first quartile. Note the *score* of this case.
3.	Find the case that lies at the third quartile (Q_3) by multiplying N by 0.75. If the result is not a whole number, round off to the nearest whole number. This is the number of the case that marks the third quartile. Note the *score* of this case.
4.	Subtract the score of the case at the first quartile (Q_1) from the score of the case at the third quartile (Q_3). The result is the interquartile range or Q.

STATISTICS IN EVERYDAY LIFE

Increasing Racial Diversity

In Chapter 3, we noted that religious diversity is increasing in the United States. A similar transformation is occurring in terms of race, ethnicity, culture, and language. Variables like race and ethnicity are nominal in level of measurement, so we can't express their dispersion with statistics like the range. However, the increasing diversity can be quickly appreciated from the table below. By the middle of this century, the United States is projected to become "majority-minority," a much more diverse population in which Hispanic and Asian Americans and Pacific Islanders are far more prominent.

Racial and Ethnic Makeup of U.S. Society, 1980 to 2050: Percentage of Total Population

Racial and Ethnic Groups	1980	2010	2050 (Projected)
Non-Hispanic whites	81%	65%	47%
Non-Hispanic blacks	11%	13%	13%
Hispanics	6%	16%	28%
Asian Americans and Pacific Islanders	2%	5%	10%
American Indians	1%	1%	1%
	101%	100%	99%

Source: U.S. Census Bureau. 2012. Retrieved from http://www.census.gov/population/projections/data/national/2008/summarytables.html

The Standard Deviation and Variance

A basic limitation of both Q and R is that they use only two scores and, thus, do not use all of the available information. Also, neither statistic provides any information on how far the scores are from each other or from some central point, such as the mean. How can we design a measure of dispersion that would avoid these limitations?

We can begin with some specifications. A good measure of dispersion should:

- Use all the scores in the distribution. The statistic should use all the information available.
- Describe the average or typical deviation of the scores. The statistic should give us an idea about how far the scores are from each other or from the center of the distribution.
- Increase in value as the scores became more diverse. This would be a very handy feature when comparing distributions because it would permit us to tell at a glance which was more variable: The higher the numerical value of the statistic, the greater the dispersion.

One way to develop a statistic to meet these criteria would be to start with the distances between each score and the mean, or the **deviations** $(X_i - \overline{X})$. The value of the deviations will increase as the differences between the scores and the mean increase. If the scores are more clustered around the mean (remember the graph for Service A in Figure 4.1), the deviations will be small. If the scores are more spread out, or more varied (like the scores for Service B in Figure 4.1), the deviations will be larger. How can we use the deviations to develop a useful statistic?

One course of action would be to use the sum of the deviations, $\Sigma(X_i - \overline{X})$ as the basis for a statistic, but, as we saw in Chapter 3, the sum of deviations will always be 0. To illustrate, consider the distribution of five scores presented in Table 4.2. If we sum the deviations of any set of scores from their mean, we would always wind up with a total of 0, regardless of the amount of variety in the scores.

TABLE 4.2 A Demonstration That the Sum of the Deviations of the Scores Around the Mean Will Always Total Zero

Scores (X_i)	Deviations ($X_i - \overline{X}$)
10	(10 − 30) = −20
20	(20 − 30) = −10
30	(30 − 30) = 0
40	(40 − 30) = 10
50	(50 − 30) = 20
$\Sigma(X_i) = 150$	$\Sigma(X_i - \overline{X}) = 0$
$\overline{X} = 150/5 = 30$	

TABLE 4.3 Computing the Standard Deviation

Scores (X_i)	Deviations ($X_i - \overline{X}$)	Deviations Squared $(X_i - \overline{X})^2$
10	(10 − 30) = −20	$(-20)^2 =$ 400
20	(20 − 30) = −10	$(-10)^2 =$ 100
30	(30 − 30) = 0	$(0)^2 =$ 0
40	(40 − 30) = 10	$(10)^2 =$ 100
50	(50 − 30) = 20	$(20)^2 =$ 400
$\Sigma(X_i) = 150$	$\Sigma(X_i - \overline{X}) = 0$	$\Sigma(X_i - \overline{X})^2 = 1000$

Still, the sum of the deviations is a logical basis for a measure of dispersion, and statisticians have developed a way around the fact that the positive deviations always equal the negative deviations. If we square each of the deviations, all values will be positive because a negative number multiplied by itself becomes positive. For example, $(-20) \times (-20) = +400$. For the scores in Table 4.2, the sum of the squared deviations would be (400 + 100 + 0 + 100 + 400), or 1000 (see Table 4.3). Thus, a statistic based on the sum of the squared deviations will have the properties we want in a good measure of dispersion.

Before we finish designing our measure of dispersion, we must deal with another problem. The sum of the squared deviations will increase with sample size: The larger the *number* of scores, the greater the value of the measure. This would make it very difficult to compare the relative variability of distributions based on samples of different sizes. We can solve this problem by dividing the sum of the squared deviations by N (sample size) and thus standardizing for samples of different sizes.

These procedures yield a statistic known as the **variance**, which is symbolized as s^2. The variance is used primarily in inferential statistics, although it is a central concept in the design of some measures of association. For purposes of describing the dispersion of a distribution, a closely related statistic, called the **standard deviation** (symbolized as s), is typically used, and this statistic is our focus for the remainder of the chapter.

The formulas for the variance (s^2) and standard deviation (s) are

FORMULA 4.3

$$s^2 = \frac{\Sigma(X_i - \overline{X})^2}{N}$$

FORMULA 4.4

$$s = \sqrt{\frac{\Sigma(X_i - \overline{X})^2}{N}}$$

Strictly speaking, Formulas 4.3 and 4.4 are for the variance and standard deviation of a population. Slightly different formulas, with $N - 1$ instead of N in the denominator, should be used when we are working with random samples rather than entire populations. This is an important point because many of the electronic calculators and statistical software packages you might be using (including SPSS) use $N - 1$ in the denominator and, thus, produce results that will be at least slightly different from those given by Formulas 4.3 and 4.4. The size of the difference will decrease as sample size increases, but many of the problems and examples in this chapter use small samples, and the differences between using N and $N - 1$ in the denominator can be considerable in such cases. Some calculators allow the user to choose either $N - 1$ or N when calculating the standard deviation. If you have such a calculator, choose "N" and your answers should match ours.

It is advisable to use a table such as Table 4.3 when computing the standard deviation. The five scores in our example are listed in the left-hand column of the table, the deviations are in the middle column, and the squared deviations are in the right-hand column.

The total of the last column in Table 4.3 is the sum of the squared deviations, which is substituted directly into Formula 4.4:

$$s = \sqrt{\frac{\Sigma(X_i - \overline{X})^2}{N}}$$

$$s = \sqrt{\frac{1000}{5}}$$

$$s = \sqrt{200}$$

$$s = 14.14$$

See the "One Step at a Time" box for detailed instructions on solving Formulas 4.3 and 4.4.

ONE STEP AT A TIME — Finding the Standard Deviation (s) and the Variance (s^2)

Step	*Operation*
1.	Construct a computing table like Table 4.3, with columns for the scores (X_i), the deviations ($X_i - \overline{X}$), and the deviations squared ($X_i - \overline{X})^2$.
2.	List the scores (X_i) in the left-hand column. Add up the scores and divide by N to find the mean.
3.	Find the deviations ($X_i - \overline{X}$) by subtracting the mean from each score, one at a time. List the deviations in the second column.
4.	Add up the deviations. The sum must equal zero (within rounding error). If the sum of the deviations does not equal zero, you have made an error and need to repeat step 3.
5.	Square each deviation and list the result in the third column.
6.	Add up column 3 (the squared deviations) and substitute that value into numerator of Formula 4.4.
7.	Divide the sum of the squared deviations from step 6 by N. To find the variance (s^2), square the value of the standard deviation (s). (Note: this value is the variance or s^2)
8.	Find the square root of the quantity you found in step 7. This is the standard deviation.

Applying Statistics 4.1 The Standard Deviation

At a local preschool, 10 children were observed for 1 hour, and the number of aggressive acts committed by each was recorded in the following list. What is the standard deviation of this distribution? We will use Formula 4.4 to compute the standard deviation.

Number of Aggressive Acts

X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	1 − 4 = −3	9
3	3 − 4 = −1	1
5	5 − 4 = 1	1
2	2 − 4 = −2	4
7	7 − 4 = 3	9
11	11 − 4 = 7	49
1	1 − 4 = −3	9
8	8 − 4 = 4	16
2	2 − 4 = −2	4
0	0 − 4 = −4	16
$\Sigma(X_i) = 40$	$\Sigma(X_i - \bar{X}) = 0$	$\Sigma(X_i - \bar{X})^2 = 118$

$$\bar{X} = \frac{\Sigma X_i}{N} = \frac{40}{10} = 4.0$$

Substituting into Formula 4.4, we have

$$s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{N}} = \sqrt{\frac{118}{10}} = \sqrt{11.8} = 3.44$$

The standard deviation for these data is 3.44.

Computing the Standard Deviation: An Additional Example

An additional example will help to clarify the procedures for computing and interpreting the standard deviation. A researcher is comparing the student bodies of two campuses. One college is located in a small town, and almost all the students reside on campus. The other is located in a large city, and the students are almost all part-time commuters. The researcher wishes to compare the ages of the students at the two campuses and has compiled the information presented in Table 4.4. Which student body is older and which is more diverse on age? (Needless to say, these very small groups are much too small for serious research and are used here only to simplify computations.)

TABLE 4.4 Computing the Standard Deviation for Two Campuses

Residential Campus		
Scores (X_i)	Deviations ($X_i - \bar{X}$)	Deviations Squared $(X_i - \bar{X})^2$
18	$(18 - 19) = -1$	$(-1)^2 = 1$
19	$(19 - 19) = 0$	$(0)^2 = 0$
20	$(20 - 19) = 1$	$(1)^2 = 1$
18	$(18 - 19) = -1$	$(-1)^2 = 1$
20	$(20 - 19) = 1$	$(1)^2 = 1$
$\Sigma(X_i) = 95$	$\Sigma(X_i - \bar{X}) = 0$	$\Sigma(X_i - \bar{X})^2 = 4$
$\bar{X} = \frac{\Sigma X_i}{N} = \frac{95}{5} = 19$		

Urban Campus		
Scores (X_i)	Deviations ($X_i - \bar{X}$)	Deviations Squared $(X_i - \bar{X})^2$
20	$(20 - 23) = -3$	$(-3)^2 = 9$
22	$(22 - 23) = -1$	$(-1)^2 = 1$
18	$(18 - 23) = -5$	$(-5)^2 = 25$
25	$(25 - 23) = 2$	$(2)^2 = 4$
30	$(30 - 23) = 7$	$(7)^2 = 49$
$\Sigma(X_i) = 115$	$\Sigma(X_i - \bar{X}) = 0$	$\Sigma(X_i - \bar{X})^2 = 88$
$\bar{X} = \frac{\Sigma X_i}{N} = \frac{115}{5} = 23$		

We see from the means that the students from the residential campus are quite a bit younger than the students from the urban campus (19 vs. 23 years of age). Which group is more diverse in age? Computing the standard deviation (Formula 4.4) will answer this question:

Residential Campus:

$$s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{N}} = \sqrt{\frac{4}{5}} = \sqrt{0.8} = 0.89$$

Urban Campus:

$$s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{N}} = \sqrt{\frac{88}{5}} = \sqrt{17.6} = 4.20$$

The higher value of the standard deviation for the urban campus means that it is more diverse with respect to age. As you can see by scanning the scores, the ages of the students at the residential college fall within a narrow age range ($R = 20 - 18 = 2$), whereas the students at the urban campus are more mixed and include students of age 25 and 30 ($R = 30 - 18 = 12$). *(For practice in computing and interpreting the standard deviation, see any of the problems at the end of this chapter. Problems with smaller data sets, such as 4.1 and 4.2, are recommended for practicing computations until you are comfortable with these procedures.)*

Visualizing Dispersion: Boxplots

A graph called a **boxplot** or a **box and whiskers plot** provides a helpful way to visualize and analyze dispersion and gives us an opportunity to apply some of our growing array of statistical tools. Boxplots use the median (see Chapter 3), range (R), and interquartile range (Q) to depict both central tendency and variability. They also display any outliers or extreme scores that might be included in the distribution.

Let's begin by examining the parts of a boxplot using Figure 4.2 as an example. This graph presents the birth rates (number of births per 1000 population) of 99 nations from a variety of levels of development and affluence. The boxplot consists of a box and two T-shaped lines, one above and one below the box. The box stretches from the 3rd quartile at the top (Q_3) to the 1st quartile at the bottom (Q_1). Thus, the height of the box reflects the value of Q or the interquartile range. The horizontal line through the box is drawn at the median (Md).

The T-shaped lines (or "whiskers") reflect the range of the scores. They extend 1.5 times the height of the box *or* to the high and low scores, whichever is closer to the box. In this case, the highest actual birth rate is 51 births per 1000 population, and this is where the upper whisker is drawn. The lower whisker is drawn at a birth rate of 8, the lowest birth rate in this group of nations.

In a boxplot, an "outlier" is defined as any score outside the "whiskers," or any score beyond 1.5 times the height of the box in either direction. Scores that are beyond three times the height of the box in either direction are called "extreme outliers." The nations displayed in Figure 4.2 have no extreme scores but we shall see some examples shortly.

Boxplots are especially useful when we want to compare the distribution of a variable across different conditions or times. Figure 4.3 presents birth rates

FIGURE 4.2 Boxplot for Birth Rates for 99 Nations, 2013

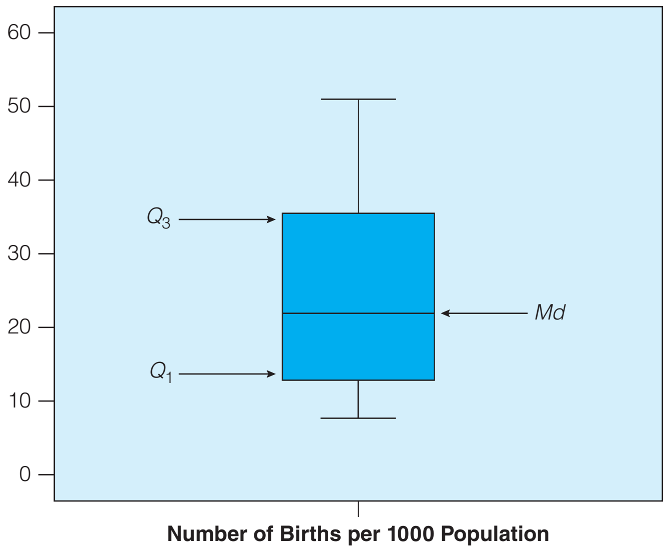


FIGURE 4.3 Boxplot for Birth Rates by Income Levels for 99 Nations, 2013

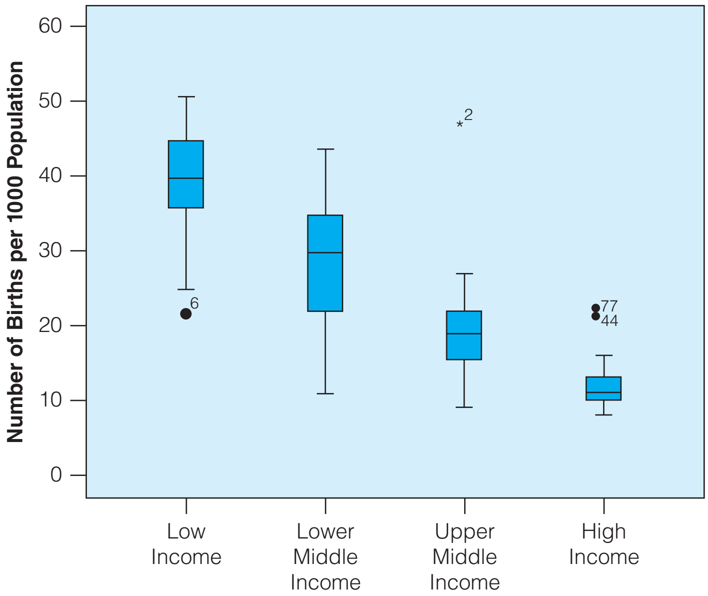


TABLE 4.5 Summary Statistics for Birth Rate by Income Levels for 99 Nations, 2013

	Income Level			
Statistics	Low Income	Lower Middle Income	Upper Middle Income	High Income
Low Score	31	11	9	8
Q_1	35.5	22	15.3	10
Median	40	30	19	11
Q_3	45	36	22	13
High Score	51	44	47	22
Q	9.5	14	6.7	3
R	20	33	38	14
s	7.5	8.6	7.4	3.4
$\overline{X}$	39.2	29.1	19.7	11.8
	N = 25	21	24	29

for the same 99 nations but, this time, the nations have been classified into income groups. Summary statistics for birth rate by income group are reported in Table 4.5.

The income groups are based on economic criteria developed by the World Bank (http://data.worldbank.org/about/country-classifications/country-and-lending-groups), and the nations in each income group share many social characteristics:

- Low-income nations have agrarian economies, which typically value large families. This category includes many African nations, Haiti, and some Asian nations.

- Lower-middle-income nations are generally in the early stages of industrialization and include Egypt, Bolivia, and India.
- Upper-middle-income nations have more advanced economies and include Argentina, Turkey, and China.
- Almost all high-income nations are highly industrialized and urbanized, and they have the highest standards of living. Children are more expensive to raise in these economies so families are smaller and birth rates lower. This category includes the United States, Canada, most Western European nations, and Japan, Australia, and New Zealand.

Looking first at central tendency, we see that the median birth rate (marked by the horizontal line in the box) decreases sharply as income level rises, as does the mean (see Table 4.5). Low-income nations have much higher average birth rates, and high-income nations have the lowest average birth rates. Lower-middle and upper-middle nations are intermediate in average birth rates.

Turning to dispersion, we see several patterns. First, upper-middle-income nations have the largest range ($R = 38$) and, by this statistic, are the most variable groups. However, the range for the nations in this category is inflated by an outlier (nation #2, which is Angola). Angola's score is an "extreme" outlier (it has a score more than 3 times the height of the box) and is marked by a* on the graph. The range, as you recall, is based on *only* the highest and lowest scores and may not be the best measure of dispersion in this case.

Second, lower-middle-income nations have the highest interquartile range (Q) and standard deviation (s) and, by these statistics, the greatest dispersion. Because R is inflated for the upper-middle-income nations by the outlier,

STATISTICS
IN EVERYDAY LIFE

Life Expectancy Around the Globe

The citizens of more affluent societies can routinely expect to live into their 70s, 80s, or even older. This is not the case, of course, for nations with lower standards of living, nutrition, and health care. The table below presents information on life expectancy for a group of 40 nations at four different income levels. Note that average life expectancy increases as income level increases, but the measures of dispersion (standard deviation and range) do just the reverse: dispersion decreases as income level increases. Why are lower-income nations more diverse on this variable? *(Hint: are people in less affluent societies more vulnerable at all ages?)*

Summary Statistics on Life Expectancy by Income Level for 40 Nations

Income Level	Mean	Standard Deviation	High	Low	Range	*N*
High	80.8	1.5	83	78	5	10
Upper-Middle	75.4	1.7	79	73	6	10
Lower-Middle	65.1	6.3	73	54	19	10
Low	58.7	6.7	70	49	21	10

Source: Population Reference Bureau. 2013. *World Population Data Sheet*. Retrieved from http://www.prb.org/Publications/Datasheets/2013/2013-world-population-data-sheet/data-sheet.aspx

Q and *s* seem to be better measures of dispersion in this situation and we would probably conclude that lower-middle-income nations are the most variable in birth rates.

Third, high-income nations are the least variable by far. This is apparent from their short boxplot and by their much lower interquartile range (3), range (14), and standard deviation (3.4). These cases are most similar to each other, in spite of the two outliers (#77 is Saudi Arabia and #44 is Israel). Why would these two nations be so different from other high-income nations in their birth rates?

In conclusion, boxplots provide useful visual and analytical information about dispersion and central tendency. Like *R*, *Q*, and *s*, boxplots are most useful when comparing variables across different conditions. *(For practice in interpreting boxplots, see problems 4.18 and 4.19.)*

Interpreting the Standard Deviation

It is very possible that the meaning of the standard deviation (i.e., why we calculate it) is not completely obvious to you at this point. You might be asking: "Once I've gone to the trouble of calculating the standard deviation, what do I have?" The meaning of this measure of dispersion can be expressed in three ways. The first and most important involves the normal curve, and we will defer this interpretation until the next chapter.

A second way of thinking about the standard deviation is as an index of dispersion that increases in value as the distribution becomes more variable. In other words, the standard deviation is higher for more diverse distributions and lower for less diverse distributions. The lowest value the standard deviation can have is 0, and this would occur for distributions with no dispersion (i.e., if every single case in the sample had exactly the same score). Thus, 0 is the lowest value possible for the standard deviation (although there is no upper limit).

A third way to get a feel for the meaning of the standard deviation is by comparing one distribution with another. We already did this when comparing the two ambulance services in Figure 4.1 and the residential and urban campuses in Table 4.4. You might also do this when comparing one group with another (e.g., men vs. women, blacks vs. whites) or the same variable at two different times. For example, suppose we found that the ages of the students on a particular campus had changed over time as indicated by the summary statistics shown in Table 4.6.

TABLE 4.6 Age on a College Campus in Two Different Years (fictitious data)

1990	2010
$\bar{X} = 21$	$\bar{X} = 25$
$s = 1$	$s = 4$

Clearly, the student body has grown older on the average, and, according to the standard deviation, it has also grown more diverse in age. The lower standard deviation for 1990 indicates that ages would be more clustered around the mean in that year (remember the distribution for Service A in Figure 4.1). In 2010, in contrast, the larger standard deviation means that the distribution is flatter and more spread out, like the distribution for Service B in Figure 4.1. In other words, compared to 2010, the students in 1990 were more similar to each other with respect to age and more clustered in a narrower age range. The standard deviation is extremely useful for making comparisons of this sort between distributions of scores.

Applying Statistics 4.2 Describing Dispersion

The homicide rates of five New England and five western states are listed in the tables below. Which group of states has the higher rate? Which is more variable?

Homicide Rate for Five New England States, 2012 (Homicides per 100,000 population)

State	Homicide Rate	Deviation	Deviation Squared
New Hampshire	1.1	−0.76	0.58
Massachusetts	1.8	−0.06	0.00
Rhode Island	3.2	1.34	1.80
Vermont	1.3	−0.56	0.31
Maine	1.9	0.04	0.00
	$\Sigma(X_i) = 9.3$	$\Sigma(X_i - X) = 0.00$	$\Sigma(X_i - \bar{X})^2 = 2.69$

$$\bar{X} = \frac{\Sigma X_i}{N} = \frac{9.3}{5} = 1.86$$

$$s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{N}} = \sqrt{\frac{2.69}{5}} = \sqrt{0.54} = 0.73$$

Homicide Rate for Five Western States, 2012 (Homicides per 100,000 population)

State	Homicide Rate	Deviation	Deviation Squared
Nevada	4.5	0.02	0.00
California	5.0	0.52	0.27
Arizona	5.5	1.02	1.04
Texas	4.4	−0.08	0.00
Washington	3.0	−1.48	2.19
	$\Sigma(X_i) = 22.4$	$\Sigma(X_i - \bar{X}) = 0.00$	$\Sigma(X_i - \bar{X})^2 = 3.51$

$$\bar{X} = \frac{\Sigma X_i}{N} = \frac{22.4}{5} = 4.48$$

$$s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{N}} = \sqrt{\frac{3.51}{5}} = \sqrt{0.70} = 0.84$$

With such small groups, you can tell by inspection that these five western states have higher homicide rates. This impression is confirmed by the mean (1.86 for the New England states and 4.48 for the western states).

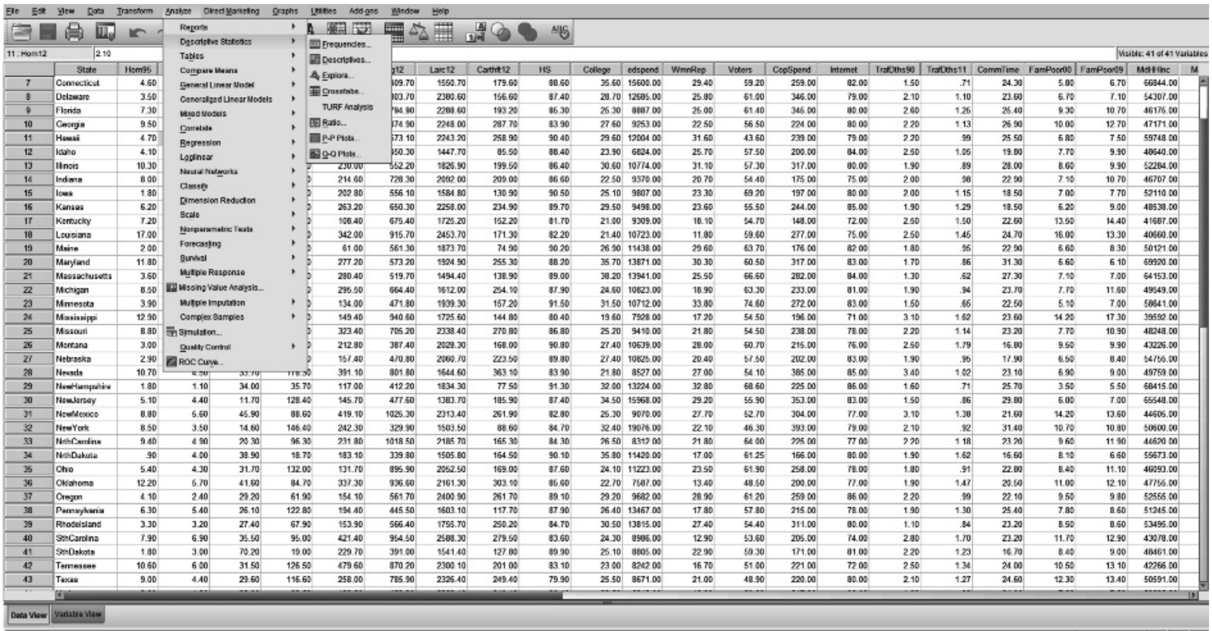
The five western states are also more variable. The range for the western states is 2.5 ($R = 5.5 - 3.0 = 2.5$), higher than the range for the New England states of 2.1 ($R = 3.2 - 1.1 = 2.1$). Similarly, the standard deviation for the western states (0.84) is greater than the standard deviation for the New England states (0.73).

Using SPSS to Produce Measures of Dispersion

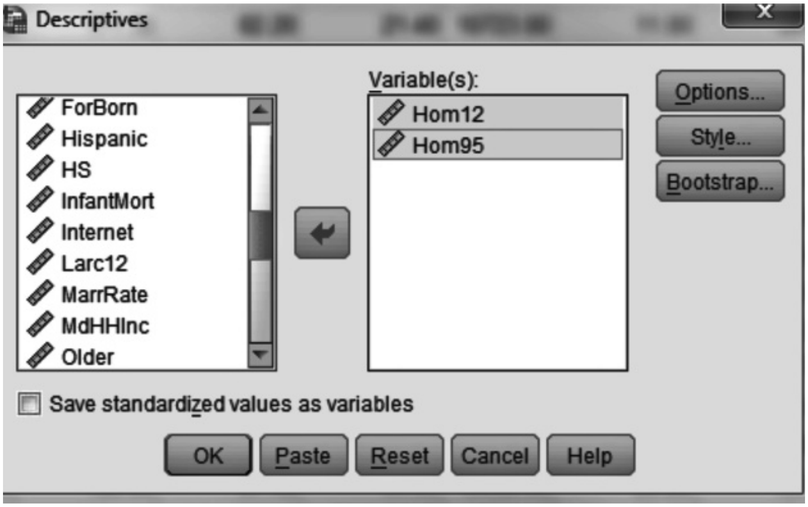
In Chapter 3, we looked at the mean and median U.S. homicide rates for 1995 and 2012. In this installment of "Using SPSS," we will expand the analysis to include measures of dispersion. We will use the **Descriptives** command, which was introduced at the end of Chapter 3.

Follow these steps:

1. Click the SPSS icon on your desktop.
2. Load the *States* database.
   a. Find the **File** command on the far left of the menu bar and then click **File → Open → Data**.
   b. Find the *States* database. You can download this file from the website for this text if you haven't already.
3. From the menu bar across the top of the SPSS window, click **Analyze**, **Descriptive Statistics**, and **Descriptives**.



4. Find the variables *Hom95* (state homicide rates in 1995) and *Hom12* (state homicide rates in 2012) in the box on the left of the "Descriptives" window and click the arrow to move the variable names to the box on the right. When you are done, the screen will look like this:



5. Click the **Options** button on the "Descriptives" window and check the box next to **Range** and **Variance** to select those statistics. Other statistics (standard deviation, mean, and the low and high scores) are selected by default.
6. Click **OK**.

The output will look like this:

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Homicide rate 2012	50	9.70	1.10	10.80	4.2600	1.96874	3.876
Homicide rate 1995	50	16.10	.90	17.00	6.7120	3.62027	13.106
Valid N (listwise)	50						

We saw in Chapter 3 that the median and mean homicide rates declined between 1995 and 2012. The measures of dispersion (range, standard deviation, and variance) also declined, indicating that the states were more similar to each other in 2012. The standard deviation, for example, fell from 3.62 in 1995 to 1.97 in 2012. As you recall from our discussion in Chapter 3, this decrease reflects the lower score of Louisiana (an extreme outlier in 1995) and the fact that the homicide rate fell in virtually every state between 1995 and 2012.

BECOMING A CRITICAL CONSUMER: Central Tendency, Dispersion, and the Shape of Distributions

In this installment of "Becoming a Critical Consumer," we will review and emphasize four of the points we have made in Chapters 2–4.

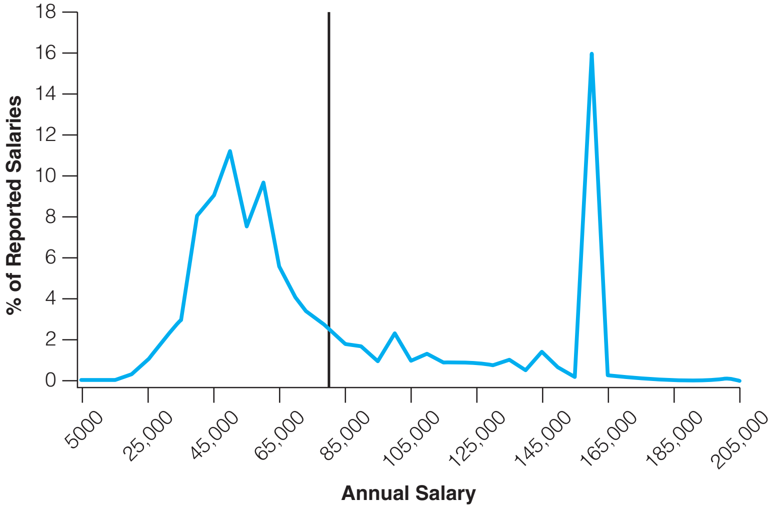
1. When considering reports regarding central tendency, pay careful attention to *which* measure is being used. Does the report say "most common" (mode), "the typical case" (median) or "the average score" (mean). As we have seen, these measures can convey very different impression of central tendency. For example, consider the report on the salaries of major league baseball players cited in Chapter 3, in which the mean and median give totally different impressions.
2. Remember that information about dispersion or variety can be just as essential as information about central tendency. For example, imagine that you were seeking the perfect place to live and prefer a city where the daily temperature is 72 degrees Fahrenheit year round. After researching the possibilities, you find three cities that meet your criterion for average temperature but have very different profiles in terms of dispersion:

City	Average Daily Temperature	Range	Standard Deviation
A	72	100	20
B	72	50	10
C	72	10	2

The temperatures in City A range from blistering to freezing. City B has less variation but still has "all four seasons." City C has very little variation in temperature and seems like your ideal residence. Note that you might have made a serious mistake if you had considered *only* central tendency in your quest for the ideal residence.

3. Expanding on the second point: If possible, you should examine the distribution as a whole, not just the summary statistics. Without this information, you might come to erroneous conclusions. For example, consider the graph displaying the starting salaries for over 20,000 lawyers in 2012, as reported by the National Association for Law Placement.*

*This analysis is based on Parikh, Ravi. 2014. "Anscombe's Quartet, and Why Summary Statistics Don't Tell the Whole Story." *Heap Analytics*. Retrieved from http://data.heapanalytics.com/anscombes-quartet-and-why-summary-statistics-dont-tell-the-whole-story/



Source: http://www.nalp.org/salarydistrib

(*continued*)

The average salary is almost $81,000, but a glance at the graph shows that there are few cases close to the average. Instead, there is a bimodal distribution in which some lawyers—those who attend the top schools and are hired by top firms—average almost $165,000 in their starting income. Most graduating lawyers make far less than this and, in fact, far less than the average salary. Looking only at central tendency would create some seriously wrong impressions about the relative affluence of legal careers.

4. Because the professional research literature focuses on the relationships between variables, univariate descriptive statistics—like means, standard deviations, and frequency distributions—probably will not be included in the final publication. However, some statistics (for example, the mean and standard deviation) serve a dual function. They not only are valuable descriptive statistics but also form the basis for many analytical techniques. Thus, they may be reported in the latter role if not in the former.

When included in research reports, measures of central tendency and dispersion will most often be presented in some summary form such as a table. A fragment of such a summary table might look like this:

Characteristics of the Sample

Variable	$\bar{X}$	s	N
Age	33.2	1.3	1078
Number of children	2.3	0.7	1077
Years married	7.8	1.5	1052
Income	$55,786	1500	987

Tables such as this describe the overall characteristics of the sample succinctly and, if you inspect the table carefully, you will have a good sense of the nature of the sample on the traits relevant to the project. Note that the number of cases varies from variable to variable. This is normal in social science research and is caused by missing data or incomplete information on some of the cases.

STATISTICS IN EVERYDAY LIFE

Changing Income Inequality in the United States

In Chapter 3, we looked at how the median and mean household incomes have changed in the United States since 1967. We saw an improvement in the financial situation of the typical American (based on the median), along with a rise in average or mean income. We also saw a positive skew (the mean was greater than the median) that grew in size over time (that is, the size of the gap between the two measures grew). What can we learn by adding a consideration of dispersion?

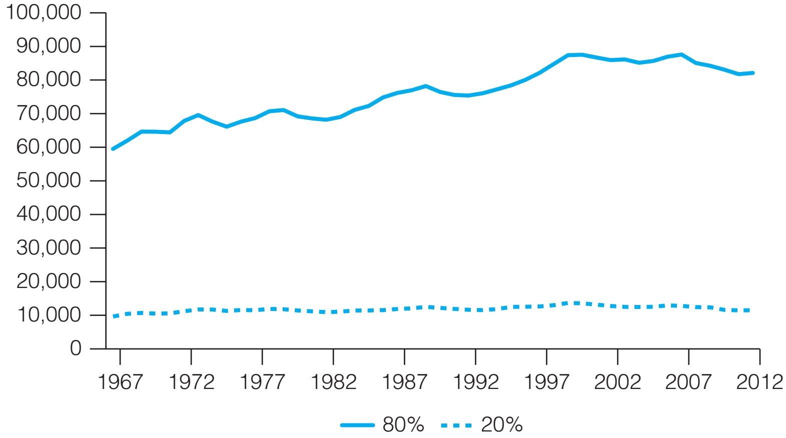
The standard deviation would be the preferred measure of dispersion for an interval-ratio-level variable such as income, but the Census Bureau does not supply this statistic and we will examine dispersion using a statistic that is a variation of the interquartile range (Q). Instead of marking the first and third quartiles, the lines in the following graph mark the 20th and 80th percentiles. The latter is the point that splits the distribution into two groups: 20% of households earned less than this value and 80% earned more. Likewise, the 80th percentile splits the distribution into two groups: 20% earned more than this income and 80% earned less. The graph displaying these values uses 2012 dollars to control for the effect of inflation.

Note that the line marking the 20th percentile stayed level throughout the time period while the 80th percentile line rose (at least until 2007). This means that the financial situation of lower-income Americans has remained virtually unchanged while the income of more affluent Americans has risen. This increasing income for the more affluent is very consistent with the increasing positive skew displayed in the related graph in Chapter 3.

(*continued*)

STATISTICS IN EVERYDAY LIFE (*continued*)

Mean Incomes for Lowest Fifth (20%) and Fourth Fifth (80%) of Households, 1967–2012 (In 2012 Dollars)



If we subtract the average income for the lowest fifth or 20% of the population from the average income for the 80% level, we have a measure of dispersion similar to the interquartile range (Q). The value of this measure increased throughout the time period. In 1967, it was about $50,000 but the gap grew to over $70,000 in 2012. (Had we measured from the 95% mark instead of the 80% mark, the gap would have grown from almost $160,000 to a little over $300,000.)

Thus, American incomes grew more variable and unequal over the time period. The increase in dispersion is solely a result of the rising incomes associated with the more affluent, not with any changes in the incomes associated with the first quintile (20%).

We saw in Chapter 3, that the income of the average American household (median) and the average income for all American households (mean) increased over this time period. Now we can see that the increase was not shared equally. It was virtually nonexistent for the less affluent (the first quintile or 20%) and quite robust for those with higher incomes. Thus, dispersion increased: Households with modest incomes continued to have modest incomes, and, consistent with the ancient folk wisdom, the rich got richer.

SUMMARY

1. Measures of dispersion summarize information about the heterogeneity, or variety, in a distribution of scores. When combined with a measure of central tendency, these statistics convey a large volume of information in just a few numbers. While measures of central tendency locate the central points of the distribution, measures of dispersion indicate the amount of diversity in the distribution.
2. The range (R) is the distance from the highest to the lowest score in the distribution. The interquartile range (Q) is the distance from the third to the first quartile (the "range" of the middle 50% of the scores). These two ranges can be used with variables measured at either the ordinal or interval-ratio level.
3. The standard deviation (s) is the most important measure of dispersion because of its central role in many more advanced statistical applications. The standard deviation has a minimum value of 0 (indicating no variation in the distribution) and increases in value as the

variability of the distribution increases. It is used most appropriately with variables measured at the interval-ratio level but is frequently computed for ordinal-level variables.

4. The variance (s^2) is used primarily in inferential statistics and in the design of some measures of association.
5. Boxplots provide a useful way of visualizing and analyzing dispersion.

SUMMARY OF FORMULAS

FORMULA 4.1	Range:	$R = \text{High score} - \text{Low score}$
FORMULA 4.2	Interquartile range:	$Q = Q_3 - Q_1$
FORMULA 4.3	Variance:	$s^2 = \dfrac{\Sigma(X_i - \overline{X})^2}{N}$
FORMULA 4.4	Standard deviation:	$s = \sqrt{\dfrac{\Sigma(X_i - \overline{X})^2}{N}}$

GLOSSARY

Box and whiskers plot. See **Boxplot**.

Boxplot. A graph that presents information about the central tendency and dispersion of a variable.

Deviation. The distance between the score and the mean.

Dispersion. The amount of variety, or heterogeneity, in a distribution of scores.

Interquartile range (*Q*). The distance from the third quartile to the first quartile.

Measures of dispersion. Statistics that indicate the amount of variety, or heterogeneity, in a distribution of scores.

Range (*R*). The highest score minus the lowest score.

Standard deviation. The statistic computed by summing the squared deviations of the scores around the mean, dividing by N, and, finally, taking the square root of the result. The most important and useful descriptive measure of dispersion; s represents the standard deviation of a sample; σ represents the standard deviation of a population.

Variance. The sum of the squared deviations of the scores around the mean, divided by N. A measure of dispersion used primarily in inferential statistics and also in correlation and regression techniques; s^2 represents the variance of a sample; σ^2 represents the variance of a population.

PROBLEMS

4.1 Compute the range and standard deviation of the following 10 scores. *(HINT: It will be helpful to organize your computations as in Table 4.4.)*

10, 12, 15, 20, 25, 30, 32, 35, 40, 50

4.2 Compute the range and standard deviation of the following 10 test scores.

66, 75, 69, 72, 84, 90, 96, 70, 55, 45

4.3 *SOC* In problem 3.1 at the end of Chapter 3, you calculated measures of central tendency for six variables for freshmen and seniors. Three of those variables are reproduced here. Calculate the mean (if necessary), the range, and the standard deviation for each variable. What information is added by the measures of dispersion? Write a paragraph summarizing the differences between freshmen and seniors.

Out-of-Pocket Expenses		Number of Movies		Rating of Cafeteria Food	
Freshmen	Seniors	Freshmen	Seniors	Freshmen	Senior
43	75	0	0	10	1
49	72	14	5	7	2
55	70	10	11	2	8
57	95	7	3	1	4
72	72	5	4	8	3
58	67	1	14	6	6
62	50	0	0	10	2
75	59	14	7	0	9
61	55	3	5	5	4
53	95	4	3	6	7
	88		5		4

4.4 SOC In problem 3.5 at the end of Chapter 3, you calculated measures of central tendency for four variables for 15 respondents. Two of those variables are reproduced here. Calculate the mean (if necessary), the range, and the standard deviation for each variable. What information is added by the measures of dispersion? Write a paragraph summarizing these statistics.

Respondent	Age	Attitude on Abortion (High score = Strong Opposition)
A	18	10
B	20	9
C	21	8
D	30	10
E	25	7
F	26	7
G	19	9
H	29	6
I	31	10
J	55	5
K	32	4
L	28	3
M	23	2
N	24	1
O	32	9

4.5 SOC In problem 3.6, you computed mean and median income for the 13 Canadian provinces and territories and for 13 U.S. states in two separate years. Now compute the standard deviation and range for each year, and, taking account of the two measures of central tendency and the two measures of dispersion, write a paragraph summarizing the distributions. What do the measures of dispersion add to what you already knew about central tendency? Did the median income of the provinces and states become more or less variable over the period? The scores are reproduced here.

Median Income for Canadian Provinces and Territories, 2000 and 2011 (Canadian dollars)

Province or Territory	2000	2011
Newfoundland and Labrador	38,800	67,200
Prince Edward Island	44,200	66,500
Nova Scotia	44,500	66,300
New Brunswick	43,200	63,930
Quebec	47,700	68,170
Ontario	55,700	73,290
Manitoba	47,300	68,710
Saskatchewan	45,800	77,300
Alberta	55,200	89,930
British Columbia	49,100	69,150
Yukon	56,000	90,090
Northwest Territories	61,000	105,560
Nunavut	37,600	65,280
Mean =		
Median =		
Range =		
Standard Deviation =		

Source: Statistics Canada. Retrieved from http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/famil108a-eng.htm

Median Income for Thirteen States, 1999 and 2012 (U.S. dollars)

State	1999	2012
Alabama	36,213	43,464
Alaska	51,509	63,348
Arkansas	29,762	39,018
California	43,744	57,020
Connecticut	50,798	64,247
Illinois	46,392	51,738
Kansas	37,476	50,003
Maryland	52,310	71,836
Michigan	46,238	50,015
New York	40,058	47,680
Ohio	39,617	44,375
South Dakota	35,962	49,415
Texas	38,978	51,926
Mean =		
Median =		
Range =		
Standard Deviation =		

Sources: 1999: U.S. Bureau of the Census. *Statistical Abstract of the United States: 2001*. p. 436. Retrieved from http://www.census.gov/prod/2002pubs/01statab/income.pdf. 2012: U.S. Bureau of the Census. *American Community Survey, 2012*. Retrieved from http://www.census.gov/hhes/www/income/data/statemedian/

4.6 Compute the standard deviation for the pretest and posttest scores that were used in problems 2.6 and 3.12. The scores are reproduced here. Taking into account all of the information you have on these variables, write a paragraph describing how the sample changed from test to test. What does the standard deviation add to the information you already had?

Case	Pretest	Posttest
A	8	12
B	7	13
C	10	12
D	15	19
E	10	8
F	10	17
G	3	12
H	10	11
I	5	7
J	15	12
K	13	20
L	4	5
M	10	15
N	8	11
O	12	20

4.7 SOC Labor force participation rates (percent employed), percent high school graduates, and mean income for males and females in ten states are reported here. Calculate the mean and the standard deviation for both groups for each variable and describe the differences. Are males and females unequal on any of these variables? How great is the gender inequality?

	Labor Force Participation Rate		Percentage HS Graduates		Mean Income	
State	Male	Female	Male	Female	Male	Female
A	65.8	54.3	81.0	81.9	55,623	50,012
B	76.7	63.0	88.4	89.7	52,345	51,556
C	71.8	57.2	82.4	84.6	55,789	48,231
D	76.1	66.6	89.5	90.9	48,907	46,289
E	75.1	63.1	86.9	88.7	62,023	58,034
F	69.9	61.1	86.3	86.4	55,000	53,897
G	73.6	59.6	87.1	87.6	49,145	47,148
H	70.5	60.3	87.0	87.6	51,897	50,659
I	66.3	55.2	81.7	84.1	51,238	45,289
J	74.5	67.1	89.1	91.6	60,746	56,489

4.8 SOC Data on several variables measuring overall health and well-being for 11 nations are reported here for 2010, with projections to 2020. Are nations becoming more or less diverse on these variables? Calculate the mean, range, and standard deviation for each year for each variable. Summarize the results in a paragraph.

Nation	Life Expectancy (years)		Infant Mortality Rate*		Fertility Rate#	
	2010	2020	2010	2020	2010	2020
Canada	81	82	5.0	4.4	1.6	1.6
China	75	76	16.5	12.6	1.5	1.5
Egypt	72	75	26.2	17.9	3.0	2.7
Germany	79	81	4.0	3.6	1.4	1.5
Japan	82	83	2.8	2.7	1.2	1.3
Mali	52	57	113.7	91.9	6.5	5.5
Mexico	76	78	17.8	13.2	2.3	2.1
Peru	71	74	27.7	20.2	2.3	2.0
Ukraine	69	70	8.7	7.3	1.3	1.4
U.S.	78	80	6.1	5.4	2.1	2.1
Zambia	52	54	68.4	50.6	6.0	5.3

Source: U.S. Bureau of the Census, 2012. *Statistical Abstract of the United States*: 2012. p. 842.

Notes:
*Number of deaths of children under one year of age per 1000 live births.
#Average number of children per female.

4.9 SOC The price of a gallon of regular gas for 20 nations is reported below. Compute the mean, median, range, interquartile range, and standard deviation for this variable, and write a paragraph summarizing these statistics.

	Nation	Cost of a Gallon of Gasoline (U.S. dollars) in 2013
1	Norway	10.08
2	Italy	8.61
3	France	8.13
4	Finland	8.05
5	Ireland	7.79
6	United Kingdom	7.75
7	South Korea	6.44
8	Japan	5.90
9	Argentina	5.38
10	South Africa	4.94
11	India	4.74
12	Canada	4.67
13	China	4.67
14	Colombia	4.52
15	United States	3.66
16	Mexico	3.43
17	Russia	3.39
18	Malaysia	2.31
19	Nigeria	2.27
20	Egypt	1.00

Source: http://www.bloomberg.com/visual-data/gas-prices/20133:South%20Africa:USD:g

4.10 SOC In problem 3.11, you computed measures of central tendency for the number of cars per 1000 population for 15 nations. The scores are reproduced here. Compute the standard deviation for this variable, and write a paragraph summarizing the mean, median, and standard deviation.

Nation	Number of Motor Vehicles per 1000 Population (2009)
United States	786
Germany	588
Canada	608
Japan	588
Australia	703
Israel	330
Russia	271
Mexico	278
Tunisia	130
Bolivia	87
China	69
Niger	8
Kenya	25
Thailand	172
India	18
Mean =	
Median =	
Standard Deviation =	

Source: World Bank, 2014. Retrieved from http://data.worldbank.org/indicator/IS.VEH.NVEH.P3

4.11 PA Per capita expenditures for public transportation for 20 cities are reported here for 2000 and 2010.

Compute the mean and standard deviation for each year, and describe the differences in expenditures for the ten-year period.

City	2000	2010
A	52	197
B	87	124
C	101	131
D	52	197
E	115	119
F	88	87
G	100	150
H	101	209
I	95	110
J	78	140
K	103	178
L	107	94
M	85	125
N	117	200
O	167	225
P	180	210
Q	120	201
R	78	141
S	55	248
T	92	131

4.12 **a.** *PA* Compute the range, interquartile range, and standard deviation for the data presented in problem 3.4. The data are reproduced here. (*NOTE: To simplify computations for Q, round off the case numbers that mark the first and third quartiles to the nearest whole number.*)

b. What would happen to the value of the standard deviation if you removed Chicago (the city with the highest score) from this distribution and recalculated? Why?

City	Annual Person-Hours Lost to Traffic Congestion
Baltimore	33
Boston	28
Buffalo	11
Chicago	44
Cleveland	13
Dallas	32
Detroit	23
Houston	37
Kansas City	14
Los Angeles	40
Miami	26
Minneapolis	27
New Orleans	20
New York	24
Philadelphia	26
Phoenix	23
Pittsburgh	23
San Antonio	19
San Diego	23
San Francisco	30
Seattle	27
Washington, D.C.	41

Source: U.S. Bureau of the Census. 2012. *Statistical Abstract of the United States: 2012*. Table 1099. Available at http://www.census.gov/prod/2011pubs/12statab/trans.pdf

4.13 *SOC* Listed here are the rates of abortion per 100,000 women for 20 states in 1973 and 1975. Compute the median, mean, standard deviation, range, and interquartile range for both years. Describe what happened to these distributions over the two-year period. Did the average rate increase or decrease? What happened to the dispersion of this distribution? Did the measures of dispersion increase or decrease? What happened between 1973 and 1975 that might explain these changes in central tendency and dispersion? *(HINT: It was a Supreme Court decision.)*

	State	1973	1975
1	Mississippi	0.2	0.6
2	Arkansas	2.9	6.3
3	Montana	3.1	9.9
4	Maine	3.5	9.5
5	South Carolina	3.8	10.3
6	Tennessee	4.2	19.2
7	Texas	6.8	19.1
8	Arizona	6.9	15.8
9	Ohio	7.3	17.9
10	Nebraska	7.3	14.3
11	Virginia	7.8	18.0
12	Iowa	8.8	14.7
13	Massachusetts	10.0	25.7
14	Pennsylvania	12.1	18.5
15	Colorado	14.4	24.6
16	Florida	15.8	30.5
17	Michigan	18.7	20.3
18	Hawaii	26.3	31.6
19	California	30.8	33.6
20	New York	53.5	40.7

Source: U.S. Bureau of the Census. 1977. *Statistical Abstract of the United States: 1977.* Washington, DC: Government Printing Office, 1977.

4.14 SW One of your goals as the new chief administrator of a large social service bureau is to equalize workloads within the various divisions of the agency. You have gathered data on caseloads per worker within each division. Which division comes closest to the ideal of an equalized workload? Which is farthest away?

A	B	C	D
50	60	60	75
51	59	61	80
55	58	58	74
60	55	59	70
68	56	59	69
59	61	60	82
60	62	61	85
57	63	60	83
50	60	59	65
55	59	58	60

4.15 You are the governor of the state and must decide which of four metropolitan police departments will win an award for efficiency. The performance of each department is summarized in monthly arrest statistics below (number of arrests per 100,000 population per month for the past five years). Which department will win the award? Why?

	Department			
	A	B	C	D
Mean =	601.30	633.17	592.70	599.99
s =	2.30	27.32	40.17	60.23

4.16 At St. Algebra College, the math department ran some special sections of the freshman math course using a variety of innovative teaching techniques. Students were randomly assigned to either the traditional sections or the experimental sections, and all students were given the same final exam. The results of the final are summarized here. What was the effect of the experimental course?

Traditional	Experimental
$\bar{X} = 77.8$	$\bar{X} = 76.8$
$s = 12.3$	$s = 6.2$
$N = 478$	$N = 465$

4.17 Compute the standard deviation for both sets of data presented in problem 3.13 and reproduced here. Compare the standard deviation computed for freshmen with the standard deviation computed for seniors. What happened? Why? Does this change relate to what happened to the mean over the four-year period? How? What happened to the shapes of the underlying distributions?

Freshmen				
10	43	30	30	45
40	12	40	42	35
45	25	10	33	50
42	32	38	11	47
22	26	37	38	10

Seniors				
10	45	35	27	50
35	10	50	40	30
40	10	10	37	10
40	15	30	20	43
23	25	30	40	10

Statistical Analysis Using SPSS

4.18 In this exercise, you will use SPSS to get measures of central tendency and dispersion for the rate of traffic fatalities in two different years (1990 and 2009) from the *States* data set. As an option, you can also get boxplots for these variables.

- Find and click the SPSS icon on your desktop.
- Load the *States* data set.
- From the menu bar across the top of the SPSS window, click **Analyze**, **Descriptive Statistics**, and **Frequencies**.
- In the list of variables in the "Frequencies" window, find *TrafDths11* (the rate of traffic fatalities in 2011) and *TrafDths90* (the rate of traffic fatalities in 1990). Click the arrow to move the variable names into the box on the right.
- Click the **Statistics** tab and check the appropriate boxes to select the following statistics: Quartiles, Mean, Median, Std. deviation, and Range. Click **Continue** and then click **OK**.
- *OPTIONAL*: To get boxplots for these variables:
  - Click **Graphs**, **Legacy Dialogs**, and **Boxplot**.
  - On the "Boxplot" window, check "Summaries of separate variables" and then click **Define**.
  - Find *TrafDths11* and *TrafDths90* in the list of variables and click the arrow to move the variable names into the "Boxes Represent" box.
  - Click **OK**.

a. Compute Q for each variable by subtracting the value for the 25th percentile from the value for the 75th percentile.

b. Write a paragraph describing the changes in the rate of traffic fatalities between 1990 and 2011. Be sure to discuss central tendency and dispersion separately, and clearly identify each statistic. *(OPTIONAL: Include an analysis of the boxplots.)*

It may be helpful to model your paragraph on one of the discussions presented earlier in this chapter, including the discussion of changing homicide rates in the "Using SPSS" section. What factors might explain the changing rate of traffic fatalities? What additional information would you need to test your explanations?

4.19 In this exercise, you will once again use the *States* data set, this time to analyze changes in the level of poverty between 2000 and 2009. We will use the **Descriptives** command to generate output.

- Find and click the SPSS icon on your desktop.
- Load the *States* data set.
- From the menu bar across the top of the SPSS window, click **Analyze**, **Descriptive Statistics**, and **Descriptives**.
- In the list of variables, find *FamPoor00* (the percentage of families below the poverty line in each state in 2000) and *FamPoor09* (the percentage of families below the poverty line in 2009). Click the arrow to move the variable names into the box on the right.
- Click the **Options** tab and check the box to select the Range. Click **Continue** and then click **OK**.
- *OPTIONAL:* To get boxplots for these variables:
  - Click **Graphs**, **Legacy Dialogs**, and **Boxplot**.
  - On the "Boxplot" window, check "Summaries of separate variables" and then click **Define**.
  - Find *Fampoor00* and *Fampoor09* in the list of variables, and click the arrow to move the variable names into the "Boxes Represent" box.
  - Click **OK**.

Write a paragraph describing the changes in family poverty between 2000 and 2009. Be sure to discuss central tendency and dispersion separately, and clearly identify each statistic. (*OPTIONAL: Include an analysis of the boxplots.*) It may be helpful to model your paragraph on one of the discussions presented earlier in this chapter.

YOU ARE THE RESEARCHER

Following Up on Earlier Projects

Two projects are presented below. The first follows up on the project presented at the end of Chapter 3, and the second follows up on the project presented at the end of Chapter 2. In the second project, you will be introduced to a new SPSS command. You are urged to complete both projects.

Project 1: The Typical American (Revisited)

In Chapter 3, you described the typical American using ten variables selected from the 2012 General Social Survey (GSS). Now you will examine variation or dispersion on some of the variables you selected.

Step 1: Choosing the Variables

Select at least five of the ordinal and interval-ratio variables you used in Chapter 3. Add more variables if you had fewer than five variables at this level of measurement.

Step 2: Getting the Statistics

Use **Descriptives** to find the range and standard deviation for each of your selected variables. With the 2012 GSS loaded, click **Analyze, Descriptive Statistics**, and **Descriptives** from the main menu of SPSS. The "Descriptives" dialog box will open. Find the names of your variables in the list on the left, and click the right arrow button to transfer them to the "Variables" box. Click **OK**, and SPSS will produce the same output you analyzed in Chapter 3. Now, however, we will consider dispersion rather than central tendency.

The standard deviation for each variable is reported in the column labeled "Std. Deviation," and the range can be computed from the values given in the Minimum and Maximum columns. At this point, the range is probably easier to understand and interpret than the standard deviation. As we have seen, the latter is more meaningful when we have a point of comparison. For example, suppose we were interested in the variable *tvhours* and how television viewing habits have changed over the years. The "Descriptives" output for 2012 shows that people watched an average of 3.09 hours a day with a standard deviation of 2.90. Suppose that data from 1975 showed an average of 3.70 hours of television viewing a day with a standard deviation of 1.1. You could conclude that television watching had, on the average, decreased over the 35-year period but that Americans had also become much more diverse in their viewing habits.

Record your results in this table, adding more rows as necessary:

SPSS Name of Variable	Mean	Range	Standard Deviation
1.			
2.			
3.			
4.			
5.			

Step 3: Interpreting Results

Start with the descriptions of the variables you wrote in Chapter 3 and add information about dispersion, referring to both the range and the standard deviation. Your job now is both to describe the "typical" American *and* to suggest the amount of variation around this central point.

Project 2: The Culture Wars (Revisited)

In Chapter 2, you examined some of the dimensions of the "culture wars" in U.S. society. In this project, you will reexamine the topic and learn how to use a new SPSS command to combine existing variables into a new summary variable. This computed variable can be used to summarize feelings and attitudes "in general" and to explore new dimensions of the issue.

Step 1: Creating a Scale to Summarize Attitudes Toward Abortion

One of the most controversial issues in the American culture wars is abortion: Under what conditions, if any, should abortion be legal? The 2012 GSS data set supplied with this text includes two variables that measure support for legal abortion. The variables differ in context: One asks specifically whether an abortion should be available if the mother is poor and wants no more children (*abpoor*). The other asks whether an abortion should be available "for any reason" (*abany*). Because these two situations are distinct, each item should be analyzed in its own right. Suppose, however, that you wanted to create a summary scale that indicated a person's *overall* feeling about abortion.

One way to do this would be to add the scores on the two variables together. This would create a new variable, which we will call *abscale,* with three possible scores:

- A score of 2 means that the respondent was consistently "pro-abortion" and answered "yes" (coded as "1") to both items.
- A score of 3 means that the respondent answered "yes" (1) to one item and "no" (2) to the other. This might be labeled an "intermediate" or "moderate" position.
- A score of 4 means that the respondent answered "no" (coded as "2") to both items. This would be a consistent "anti-abortion" position.

The table below summarizes the scoring possibilities.

If response on abany is	and response on abpoor is	score on abscale will be
1 (Yes)	1 (Yes)	2 (Pro-abortion)
1 (Yes)	2 (No)	3 (Moderate)
2 (No)	1 (Yes)	3 (Moderate)
2 (No)	2 (No)	4 (Anti-abortion)

The new variable, *abscale,* summarizes each respondent's overall position on the issue. Once created, *abscale* could be analyzed, transformed, and manipulated exactly like a variable actually recorded in the data file.

Using Compute

We will create *abscale* with the **Compute** command. To begin, click **Transform** and then "Compute Variable" from the main menu. The "Compute Variable" window will appear. Find the "Target Variable" box in the upper left corner of this window. With the *GSS2012* data set loaded, follow these steps to compute *abscale*:

- First, assign a name to the new variable we are about to compute: Type *abscale* in the "Target Variable" box.
- Next, we need to tell SPSS how to compute the new variable. In this case, *abscale* will be computed by adding the scores of *abany* and *abpoor.*
  - Find *abany* in the variable list on the left, and click the arrow button in the middle of the screen to transfer the variable name to the "Numeric Expression" box.
  - Click the plus sign (+) on the calculator pad under the "Numeric Expression" box, and the sign will appear next to *abany*.
  - Finally, highlight *abpoor* in the variable list, and click the arrow button to transfer the variable name to the "Numeric Expression" box.
  - The expression in the "Numeric Expression" box should now read *abany* + *abpoor*
  - Click **OK,** and *abscale* will be created and added to the data set.

If you want to keep this new variable permanently, click **Save** from the **File** menu, and the updated data set that includes *abscale* will be saved. (If you are using the student version of SPSS, remember that your data set is limited to 50 variables and you may not be allowed to save the data set with *abscale* added.)

Examining the Variables

We now have three variables that measure attitudes toward abortion—two items referring to specific situations, and a more general, summary item. It is always a good idea to check the frequency distribution for computed variables to make sure that the computations were carried out as we intended. Use the **Frequencies** procedure (click **Analyze, Descriptive Statistics,** and **Frequencies**) to get tables for *abany, abpoor,* and *abscale*. Your output will look like this:

Abortion if Woman Wants for Any Reason					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	YES	408	28.0	45.1	45.1
	NO	496	34.0	54.9	100.0
	Total	904	62.0	100.0	
Missing	IAP	509	34.9		
	DK	32	2.2		
	NA	12	.8		
	Total	553	38.0		
Total		1457	100.0		

Low Income—Can't Afford More Children					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	YES	405	27.8	45.0	45.0
	NO	495	34.0	55.0	100.0
	Total	900	61.8	100.0	
Missing	IAP	509	34.9		
	DK	35	2.4		
	NA	13	.9		
	Total	557	38.2		
Total		1457	100.0		

abscale					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2.00	350	24.0	39.8	39.8
	3.00	101	6.9	11.5	51.3
	4.00	428	29.4	48.7	100.0
	Total	879	60.3	100.0	
Missing	System	578	39.7		
Total		1457	100.0		

Missing Cases

Note that about 550 respondents have "missing" scores on *abany* and *abpoor*. Remember that no respondent is given the entire GSS, and the vast majority of the "missing cases" received a form of the GSS that did not include these two items. These cases are coded as "IAP" in the tables above.

Now look at *abscale* and note that even more cases (578) are “missing.” When SPSS executes a **Compute** statement, it ignores any cases that are missing scores on any of the constituent items. If these cases were not eliminated from the computations, a variety of errors and misclassifications could result. For example, if cases with missing scores were included, a person who scored a 2 (“anti-abortion”) on *abany* and then didn’t respond to *abpoor* would have a total score of 2 on *abscale.* Thus, this case would be treated as “pro-abortion” when the only information we have indicates that this respondent is “anti-abortion.” Cases with missing scores are ignored to avoid this type of error.

Step 2: Interpreting the Results

Compare the distributions of these three variables with each other and write a report in which you answer the following questions:

1. How does the level of approval for abortion differ from one specific situation to another?
2. What percent of the respondents approved and disapproved of abortion in both circumstances? What percentage approved in one situation but not the other? (*Use the distribution of* abscale *to answer this question.*)
3. What do these patterns reveal about value consensus in American society? Are Americans generally in agreement about the issue of abortion?

5 The Normal Curve

LEARNING OBJECTIVES By the end of this chapter, you will be able to:

1. Define and explain the concept of the normal curve.
2. Convert empirical scores to *Z* scores and use *Z* scores and the normal curve table (Appendix A) to find areas above, below, and between points on the curve.
3. Express areas under the curve in terms of probabilities.

USING STATISTICS

The statistical techniques presented in this chapter can be used to:

- Describe the position of scores on certain tests: "John scored higher than 75% of the students who took the test" or "Mary is in the 98th percentile in math ability."
- Estimate the probability that certain events will occur. For example: "The probability that a randomly selected student will have a test score over 100 on this test is 23 out of 100."

The **normal curve** is a concept of great importance in statistics. In combination with the mean and standard deviation, it is used to make precise descriptive statements about empirical distributions. Also, the normal curve is central to the theory that underlies inferential statistics. Thus, this chapter concludes our treatment of Part I (Descriptive Statistics) and lays important groundwork for Part II (Inferential Statistics).

The normal curve is sometimes used to assign grades on tests, and you may be familiar with this application. In this guise, the normal curve is often called the "bell curve," and grading "on the curve" means that the instructor wants the scores to follow a specific pattern: The modal grade is a C and there will be equal numbers of As and Fs, Bs and Ds. In other words, the distribution of grades should look like a bell or a mound.

Properties of the Normal Curve

The normal curve is a theoretical model, a line chart (see Chapter 2) that is unimodal (i.e., has a single mode, or peak), perfectly smooth, and symmetrical (unskewed), so its mean, median, and mode are all exactly the same value. It is bell-shaped, and its tails extend infinitely in both directions. Of course, no empirical distribution matches this ideal model perfectly, but some variables (e.g., test results from large classes, standardized test scores such as the GRE) are close enough to permit the assumption of normality. In turn, this assumption makes possible one of the most important uses of the normal curve—the description of empirical distributions based on our knowledge of the theoretical normal curve.

The crucial point about the normal curve is that distances along the horizontal axis, when measured in standard deviations from the mean, always encompass the same proportion of the total area under the curve. In other words, the distance from any point to the mean—when measured in standard deviations—will cut off exactly the same proportional part of the area under the curve.

To illustrate, Figures 5.1 and 5.2 present two hypothetical distributions of IQ scores, both normally distributed, for fictional groups of males and females, such that

Males	Females
$\overline{X} = 100$	$\overline{X} = 100$
$s = 20$	$s = 10$
$N = 1000$	$N = 1000$

Figures 5.1 and 5.2 are drawn with two scales on the horizontal axis of the graph. The upper scale is stated in "IQ units" and the lower scale in standard deviations from the mean. These scales are interchangeable and we can easily shift from one to the other. For example, for the males, an IQ score of 120 is one standard deviation (remember that, for the male group, $s = 20$) above the mean and an IQ of 140 is two standard deviations above (to the right of) the mean. IQ scores to the left of the mean are marked as negative values on the standard deviation scale because they are less than the mean. An IQ of 80 is one standard deviation below the mean, an IQ score of 60 is two standard deviations less than the mean, and so forth.

Figure 5.2 is marked in a similar way, except that, because its standard deviation is a different value ($s = 10$), the markings occur at different points. For

FIGURE 5.1 IQ Scores for a Group of Males

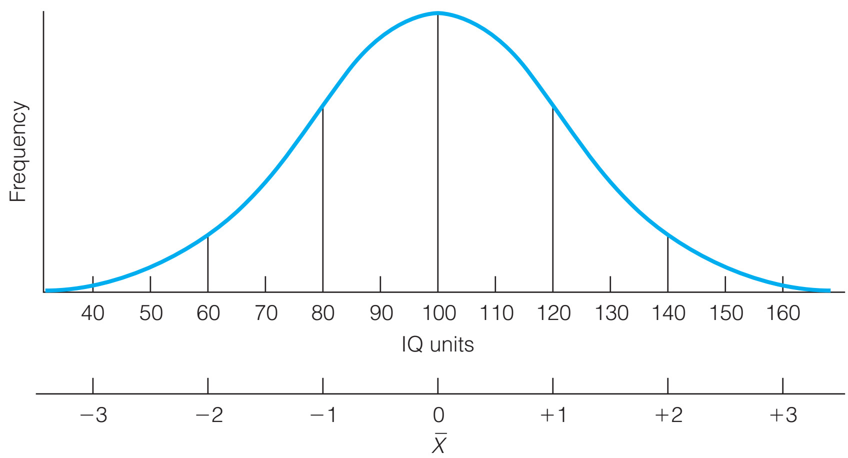
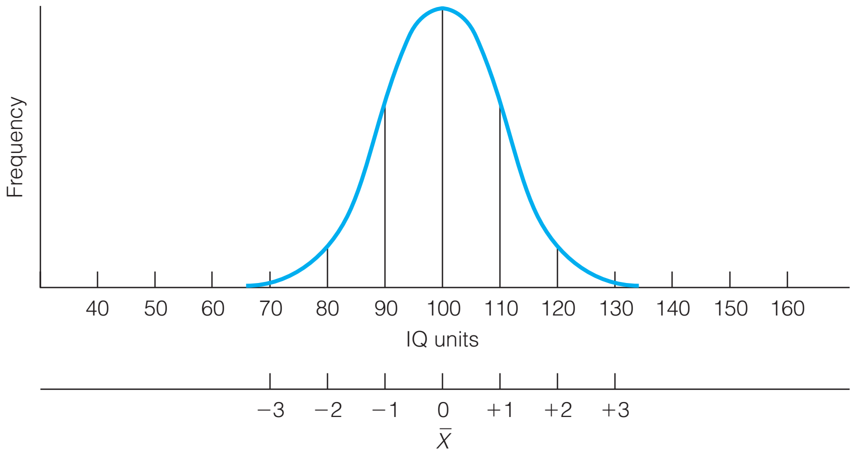


FIGURE 5.2 IQ Scores for a Group of Females



the female sample, one standard deviation above the mean is an IQ of 110, one standard deviation below the mean is an IQ of 90, and so forth.

Recall that, on any normal curve, distances along the horizontal axis, when measured in standard deviations, always encompass exactly the same proportion of the total area under the curve. Specifically, the distance between one standard deviation above the mean and one standard deviation below the mean (or ± 1 standard deviation) encompasses exactly 68.26% of the total area under the curve. This means that in Figure 5.1, 68.26% of the total area lies between the score of 80 (-1 standard deviation) and 120 ($+1$ standard deviation). The standard deviation for females is 10, so the same percentage of the area (68.26%) lies between the scores of 90 and 110. On any normal distribution, 68.26% of the total area will always fall between ± 1 standard deviation, regardless of the trait being measured and the numerical values of the mean and standard deviation.

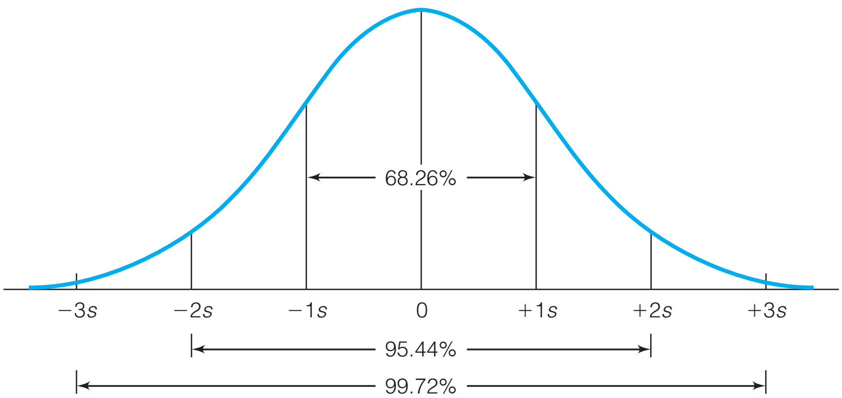
STATISTICS
IN EVERYDAY LIFE

Tests of Intelligence

Tests of intelligence are designed to generate distributions of scores that are approximately normal. That is, the tests mix easier and harder questions so that the mean, mode, and median will be about 100 and the final distribution of scores will be bell-shaped.

Intelligence testing can be controversial and, although we cannot go into detail here, you should be aware that there is considerable debate about the meaning of IQ scores. Tests of intelligence may have cultural and other biases, and there is considerable disagreement about whether or not they actually measure native mental ability. In particular, you should be aware that the more or less normal distribution of IQ scores does not necessarily mean that the underlying and elusive quality called intelligence—however that might be defined—is also normally distributed.

FIGURE 5.3 Areas Under the Normal Curve



It will be useful to familiarize yourself with the following relationships between distances from the mean and areas under the normal curve:

Between	Lies
±1 standard deviation	68.26% of the area
±2 standard deviations	95.44% of the area
±3 standard deviations	99.72% of the area

These relationships are displayed graphically in Figure 5.3.

We can describe empirical distributions that are at least approximately normal using these relationships between distance from the mean and area. The position of individual scores can be described with respect to the mean, the distribution as a whole, or any other score in the distribution.

The areas between scores can also be expressed in numbers of cases rather than percentage of total area. For example, a normal distribution of 1000 cases will contain about 683 cases (68.26% of 1000 cases) between ±1 standard deviation of the mean, about 954 between ±2 standard deviations, and about 997 between ±3 standard deviations. Thus, for any normal distribution, only a few cases will be farther away from the mean than ±3 standard deviations.

Using the Normal Curve

We have seen that we can find areas under the normal curve for scores that are exactly 1, 2, or 3 standard deviations above or below the mean. To work with values that are not exact multiples of the standard deviation, we must express the original scores in units of the standard deviation or convert them into ***Z* scores**. The original scores could be in any unit of measurement (feet, IQ, dollars), but *Z* scores always have the same values for their mean (0) and standard deviation (1).

Computing *Z* Scores

Think of converting the original scores into *Z* scores as a process of changing value scales—similar to changing from meters to yards, kilometers to miles, or gallons to liters. These units are different but equally valid ways of expressing distance, length, or volume. For example, a mile is equal to 1.61 kilometers, so two towns that are 10 miles apart are also 16.1 kilometers apart and a "5k" race covers about 3.10 miles. Although you may be more familiar with miles than kilometers, either unit works perfectly well as a way of expressing distance.

In the same way, the original (or "raw") scores and *Z* scores are two equally valid but different ways of measuring distances under the normal curve. In Figure 5.1, for example, we could describe a particular score in terms of IQ units ("John's score was 120") or standard deviations ("John scored one standard deviation above the mean").

When we compute *Z* scores, we convert the original units of measurement (IQ scores, inches, dollars, etc.) to *Z* scores and, thus, "standardize" the normal curve to a distribution that has a mean of 0 and a standard deviation of 1. The mean of the empirical normal distribution will be converted to 0, its standard deviation to 1, and all values will be expressed in *Z*-score form. The formula for computing *Z* scores is

FORMULA 5.1

$$Z = \frac{X_i - \overline{X}}{s}$$

This formula will convert any score (X_i) from an empirical normal distribution into the equivalent *Z* score. To illustrate with the men's IQ data (Figure 5.1), the *Z*-score equivalent of a raw score of 120 would be

$$Z = \frac{120 - 100}{20} = \frac{20}{20} = +1.00$$

The *Z* score of positive 1.00 indicates that the original score lies one standard deviation unit above (to the right of) the mean. A negative score would fall below (to the left of) the mean. *(For practice in computing Z scores, see any of the problems at the end of this chapter.)*

ONE STEP AT A TIME **Finding *Z* Scores**

Step	***Operation***
1.	Subtract the value of the mean ($\overline{X}$) from the value of the score (X_i).
2.	Divide the quantity found in step 1 by the value of the standard deviation (s). The result is the *Z*-score equivalent for this raw score.

The Normal Curve Table (Appendix A)

The theoretical normal curve has been very thoroughly described by statisticians. The areas related to any *Z* score have been precisely determined and organized into a table format. This **normal curve table**, or *Z*-score table, is presented as Appendix A in this text; for purposes of illustration, a small portion of it is reproduced here as Table 5.1.

The normal curve table consists of three columns, with *Z* scores in the left-hand column (column a), areas between the *Z* score and the mean in the middle (column b), and areas beyond the *Z* score in the right-hand column (column c). To find the area between any *Z* score and the mean, go down the *Z*-score column until you find the score. For example, go down column a, either in Appendix A or in Table 5.1, until you find a *Z* score of +1.00. The entry in column b shows that the "Area Between Mean and *Z*" is 0.3413.

The table presents areas in the form of proportions, but we can easily translate these into percentages by multiplying them by 100 (see Chapter 2). We could say either "a proportion of 0.3413 of the total area under the curve lies between a *Z* score of 1.00 and the mean" or "34.13% of the total area lies between a *Z* score of 1.00 and the mean."

To illustrate further, find the *Z* score of 1.50 either in the *Z*-score column of Appendix A or the abbreviated table presented in Table 5.1. This score is 1½

TABLE 5.1 An Illustration of How to Find Areas Under the Normal Curve Using Appendix A

(a) *Z*	(b) Area Between Mean and *Z*	(c) Area Beyond *Z*
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
⋮	⋮	⋮
1.00	0.3413	0.1587
1.01	0.3438	0.1562
1.02	0.3461	0.1539
1.03	0.3485	0.1515
⋮	⋮	⋮
1.50	0.4332	0.0668
1.51	0.4345	0.0655
1.52	0.4357	0.0643
1.53	0.4370	0.0630
⋮	⋮	⋮

standard deviations to the right of the mean and corresponds to an IQ of 130 for the men's IQ distribution. The area in column b for this score is 0.4332. This means that a proportion of 0.4332—or a percentage of 43.32%—of all the area under the curve lies between this score and the mean.

The third column in the table presents "Areas Beyond *Z*." These are areas above positive scores or below negative scores. The use of this column will be explained later in this chapter.

To conserve space, the normal curve table in Appendix A includes only positive *Z* scores. Because the normal curve is perfectly symmetrical, however, the area between the score and the mean—column b—for a negative score will be exactly the same as for a positive score of the same numerical value. For example, the area between a *Z* score of −1.00 and the mean will also be 34.13%, exactly the same as the area we found previously for a score of +1.00. As will be repeatedly demonstrated later, however, the sign of the *Z* score is extremely important and should be carefully noted.

For practice in using Appendix A to describe areas under an empirical normal curve, verify that the *Z* scores and areas in Table 5.2 are correct for the men's IQ distribution. For each IQ score, we compute the *Z* score with Formula 5.1, and then consult Appendix A to find areas between the score and the mean.

As shown in Table 5.3, the same procedures apply when the *Z*-score equivalent of an actual score happens to be a negative value (that is, when the raw score lies below the mean).

Remember that the areas in Appendix A will be the same for *Z* scores of the same numerical value regardless of sign. For the men's IQ distribution, the area between the score of 138 ($Z = +1.90$) and the mean is the same as the area between 62 ($Z = -1.90$) and the mean. *(For practice in using the normal curve table, see any of the problems at the end of this chapter.)*

TABLE 5.2 Finding Positive *Z* Scores with $\bar{X} = 100$ and $s = 20$

IQ Score	*Z* Score	Area Between *Z* and Mean
110	+0.50	19.15%
125	+1.25	39.44%
133	+1.65	45.05%
138	+1.90	47.13%

TABLE 5.3 Finding Negative *Z* Scores with $\bar{X} = 100$ and $s = 20$

IQ Score	*Z* Score	Area Between *Z* and Mean
93	−0.35	13.68%
85	−0.75	27.34%
67	−1.65	45.05%
62	−1.90	47.13%

Finding the Total Area Above and Below a Score

So far, we have seen how the normal curve table can be used to find areas between a Z score and the mean. Appendix A can also be used to find other kinds of areas in empirical distributions that are at least approximately normal in shape. For example, suppose you need to determine the total area below the scores of two male subjects in the distribution described in Figure 5.1. The first subject has a score of 117 ($X_1 = 117$), which is equivalent to a Z score of $+0.85$:

$$Z = \frac{X_i - \bar{X}}{s} = \frac{117 - 100}{20} = \frac{17}{20} = +0.85$$

The plus sign of the Z score indicates that the score should be placed above (to the right of) the mean. To find the area below a positive Z score, the area between the score and the mean (see column b) must be added to the area below the mean. As we noted earlier, the normal curve is symmetrical (unskewed) and its mean will be equal to its median. Therefore, the area below the mean (just like the median) will be 50%. Study Figure 5.4 carefully. We are interested in the shaded area.

By consulting the normal curve table, we find that the area between the score and the mean (see column b) is 30.23% of the total area. The area below a Z score of $+0.85$ is therefore 80.23% (50.00% + 30.23%). This subject scored higher than 80.23% of the persons tested.

The second subject has an IQ score of 73 ($X_2 = 73$), which is equivalent to a Z score of -1.35:

$$Z = \frac{X_i - \bar{X}}{s} = \frac{73 - 100}{20} = -\frac{27}{20} = -1.35$$

To find the area below a negative score, we use the right-hand column or "Area Beyond Z." The area in which we are interested is depicted in Figure 5.5, and we must determine the size of the shaded area. The "area beyond" (see column c) a score of -1.35 is 0.0885, which we can express as 8.85%. The second subject ($X_2 = 73$) scored higher than 8.85% of the tested group.

FIGURE 5.4 Finding the Area Below a Positive *Z* Score

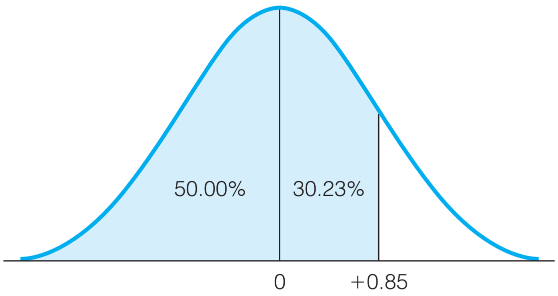
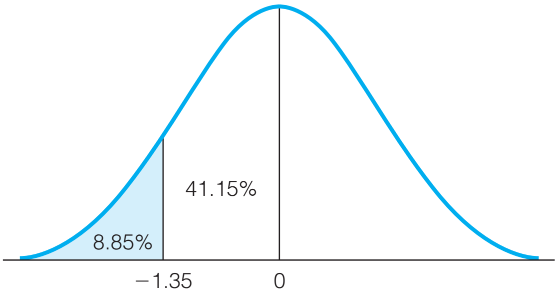


FIGURE 5.5 Finding the Area Below a Negative *Z* Score



In the foregoing examples, we found the area below a score. Essentially the same techniques are used to find the area above a score. If we need to determine the area above an IQ score of 108, for example, we would first convert to a *Z* score,

$$Z = \frac{X_i - \overline{X}}{s} = \frac{108 - 100}{20} = \frac{8}{20} = +0.40$$

and then proceed to Appendix A. The shaded area in Figure 5.6 represents the area in which we are interested. The area above a positive score is found in the "Area Beyond *Z*" column, and, in this case, the area is 0.3446, or 34.46%.

These procedures are summarized in Table 5.4 and in the "One Step at a Time" box. They might be confusing at first, and you should *always* draw the curve and shade in the areas in which you are interested. *(For practice in finding areas above or below Z scores, see problems 5.1 to 5.7.)*

FIGURE 5.6 Finding the Area Above a Positive *Z* Score

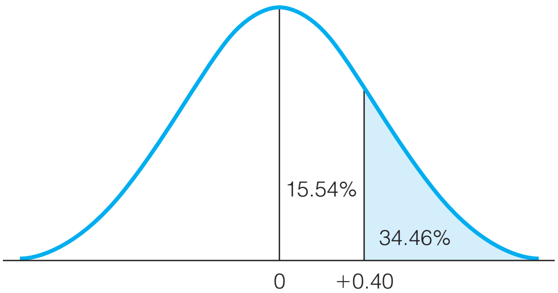


TABLE 5.4 Finding Areas Above and Below Positive and Negative Scores

	When the *Z* Score Is	
To Find Area	Positive	Negative
Above Z	Look in column c	Add column b area to 0.5000
Below Z	Add column b area to 0.5000	Look in column c

ONE STEP AT A TIME	Finding Areas Above and Below Positive and Negative *Z* Scores

Step ***Operation***

1. Compute the *Z* score. Note whether the score is positive or negative.
2. Find the *Z* score in column a of the normal curve table (Appendix A).
3. To find the total area:

- **Below a Positive Z Score:**
  Add the column b area to 0.5000. To express the area as a percentage, multiply the column b area by 100 and add it to 50.00%.
- **Above a Positive Z Score:**
  Look in column c. This value is the area above the score expressed as a proportion. To express the area as a percentage, multiply the column c area by 100.
- **Below a Negative Z Score:**
  Look in column c. This value is the area below the score expressed as a proportion. To express the area as a percentage, multiply the column c area by 100.
- **Above a Negative Z Score:**
  Add the column b area to 0.5000. To express the area as a percentage, multiply the column b area by 100 and add it to 50.00%.

Finding Areas Between Two Scores

On occasion, you will need to determine the area between two scores. When the scores are on opposite sides of the mean, the area between them can be found by adding the areas between each score and the mean. Using the men's IQ distribution as an example, if we wished to know the area between the IQ scores of 93 and 112, we would convert both scores to *Z* scores, find the area between each score and the mean in Appendix A, and add these two areas together. The first IQ score of 93 converts to a *Z* score of −0.35:

$$Z = \frac{X_i - \overline{X}}{s} = \frac{93 - 100}{20} = -\frac{7}{20} = -0.35$$

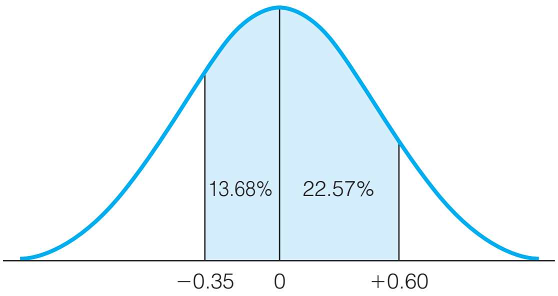
The second IQ score (112) converts to +0.60:

$$Z = \frac{X_i - \overline{X}}{s} = \frac{112 - 100}{20} = \frac{12}{20} = +0.60$$

Both scores are placed on Figure 5.7. We are interested in the total shaded area. The total area between these two scores is 13.68% + 22.57%, or 36.25%. Therefore, 36.25% of the total area (or about 363 of the 1000 cases) lies between the IQ scores of 93 and 112.

When the scores of interest are on the same side of the mean, a different procedure must be followed to determine the area between them. For example, if we

FIGURE 5.7 Finding the Area Between Two Scores



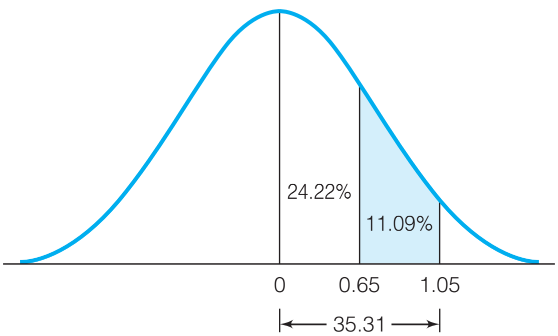
were interested in the area between the scores of 113 and 121, we would begin by converting these scores into Z scores:

$$Z = \frac{X_i - \overline{X}}{s} = \frac{113 - 100}{20} = \frac{13}{20} = +0.65$$

$$Z = \frac{X_i - \overline{X}}{s} = \frac{121 - 100}{20} = \frac{21}{20} = +1.05$$

The scores are noted in Figure 5.8; we are interested in the shaded area. To find the area between two scores on the same side of the mean, find the area between each score and the mean, and then subtract the smaller area from the larger. Using column b of Appendix A, we see that 24.22% of the total area lies between the Z score of +0.65 and the mean, and that 35.31% of the area lies between the Z score of +1.05 and the mean. Therefore, the area between these two scores is 35.31% − 24.22%, or 11.09% of the total area. The same technique would be followed if both scores had been below the mean. The procedures for finding areas between two scores are summarized in Table 5.5 and in the "One Step at a Time" box. *(For practice in finding areas between two scores, see problems 5.3, 5.4, and 5.6 to 5.9.)*

FIGURE 5.8 Finding the Area Between Two Scores



ONE STEP AT A TIME Finding Areas Between Z Scores

Step	*Operation*
1.	Compute the Z scores for both raw scores. Note whether the scores are positive or negative.
2.	Find the areas between each score and the mean (see column b).
If the scores are on the same side of the mean:	
3.	Subtract the smaller area from the larger area. Multiply this value by 100 to express it as a percentage.
If the scores are on opposite sides of the mean:	
4.	Add the two areas together to get the total area between the scores. Multiply this value by 100 to express it as a percentage.

Applying Statistics 5.1 Finding *Z* Scores and Areas Below

You have just received your score on a standardized test. Your score was 78, and you know that the mean score on the test was 67 with a standard deviation of 5. How does your score compare with the distribution of all test scores?

If you can assume that the test scores are normally distributed, you can compute a *Z* score and find the area below or above your score. The *Z*-score equivalent of your raw score would be:

$$Z = \frac{X_i - \overline{X}}{s} = \frac{78 - 67}{5} = \frac{11}{5} = +2.20$$

Turning to Appendix A, we find that the "Area Between Mean and *Z*" for this score is 0.4861, which could also be expressed as 48.61%. Because this is a positive *Z* score, we need to add this area to 50.00% to find the total area below. Your score is higher than 48.61 +50.00, or 98.61%, of all the test scores. You did pretty well!

Applying Statistics 5.2 Finding *Z* Scores and Areas Between

All sections of Biology 101 at a large university were given the same final exam. Test scores were distributed normally, with a mean of 72 and a standard deviation of 8. What percentage of students scored between 60 and 69 (a grade of D) and what percentage scored between 70 and 79 (a grade of C)?

The two scores of the D grades are both below the mean. Using Table 5.5 as a guide, we must first compute *Z* scores, find areas between each score and the mean, and then subtract the smaller area from the larger:

$$Z = \frac{X_i - \overline{X}}{s} = \frac{60 - 72}{8} = -\frac{12}{8} = -1.50$$

$$Z = \frac{X_i - \overline{X}}{s} = \frac{69 - 72}{8} = -\frac{3}{8} = -0.38$$

The area between $Z = -1.50$ and the mean is 0.4332 and the area between $Z = -0.38$ and the mean is 0.1480. Subtracting the smaller from the larger (0.4332 − 0.1480) gives 0.2852. Changing to percentage format, we can say that 28.52% of the students earned a D on the test.

To find the percentage of students who earned a C, we must add column b areas together, since the scores (70 and 79) are on opposite sides of the mean (see Table 5.5):

$$Z = \frac{X_i - \overline{X}}{s} = \frac{70 - 72}{8} = -\frac{2}{8} = -0.25$$

$$Z = \frac{X_i - \overline{X}}{s} = \frac{79 - 72}{8} = \frac{7}{8} = +0.88$$

The area between $Z = -0.25$ and the mean is 0.0987 and the area between $Z = +0.88$ and the mean is 0.3106. Therefore, the total area between these two scores is 0.0987 + 0.3106, or 0.4093. Translating to percentages again, we can say that 40.93% of the students earned a C on this test.

TABLE 5.5 Finding Areas Between Scores

Situation	Procedure
Scores are on the SAME side of the mean	Find the areas between each score and the mean (see column b). Subtract the smaller area from the larger area.
Scores are on OPPOSITE sides of the mean	Find the areas between each score and the mean (see column b). Add the areas together.

Using the Normal Curve to Estimate Probabilities

To this point, we have thought of the theoretical normal curve as a way of describing areas above, below, and between scores. We have also seen that these areas can be converted into the number of cases above, below, and between scores.

In this section, we will explore an additional use of the normal curve: It can be used to estimate probabilities or the chances that certain events will occur. We will find these probabilities using the same techniques used to find areas earlier in this chapter. The only new idea introduced in this section is that the areas under the theoretical normal curve (or in Appendix A) can be thought of as probabilities. Before we consider these mechanics, however, let us examine what is meant by "probability."

Estimating Probabilities

Although we are rarely systematic or rigorous about it, we all attempt to deal with probabilities every day, and, indeed, we base our behavior on our estimates of the likelihood that certain events will occur. We constantly ask (and answer)

STATISTICS IN EVERYDAY LIFE

How Common Is the Normal Curve?

How normal is the normal curve? How often does it appear in everyday life? Mathematically speaking, the scores of a variable are most likely to be normal in shape if they represent random deviations around a mean. For example, imagine a manufacturing process designed to produce parts of a uniform size. Any errors or mistakes will be equally likely to be too large or too small, and the number of errors will decrease as we move further away from the desired size. In other words, the errors will be distributed around the mean in a bell-shaped curve: minor errors will be common and larger errors will be rare.

In the social world, the variables that are of most concern to us are either known *not* to be normal (e.g., income, the distribution of which is positively skewed) or they have shapes that are unknown (e.g., the degree of support for capital punishment). In other words, we usually do not find normal distributions when we examine the variables in our research projects. As you will see, the importance of the normal curve relates more to the concepts and logic that underlie hypothesis testing—the subject of the next part of this text—than to descriptions of our variables.

questions such as: What is the probability of rain? Of drawing to an inside straight in poker? Of the worn-out tires on my car going flat? Of passing a test if I don't study?

To estimate the probability of an event, we must first be able to define what would constitute a "success." The preceding examples contain several different definitions of a success (rain, drawing a certain card, flat tires, and passing grades). To determine a probability, a fraction must be established, with the numerator equaling the number of events that would constitute a success and the denominator equaling the total number of possible events where a success could theoretically occur:

$$\text{Probability} = \frac{\#\ \text{successes}}{\#\ \text{events}}$$

To illustrate, assume that we wish to know the probability of selecting a specific card—say, the king of hearts—in one draw from a well-shuffled deck of cards. Our definition of a success is specific (drawing the king of hearts); and with the information given, we can establish a fraction. Only one card satisfies our definition of success, so the number of events that would constitute a success is 1; this value will be the numerator of the fraction. There are 52 possible events (that is, 52 cards in the deck), so the denominator will be 52. The fraction is thus 1/52, which represents the probability of selecting the king of hearts on one draw from a well-shuffled deck of cards. Our probability of success is 1 out of 52.

We can leave this fraction as it is, or we can express it in several other ways. For example, we can express it as an odds ratio by inverting the fraction, showing that the odds of selecting the king of hearts on a single draw are 52:1 (or a probability of fifty-two to one). We can express the fraction as a proportion by dividing the numerator by the denominator. For our example, the corresponding proportion is 0.0192, which is the proportion of all possible events that would satisfy our definition of a success. In the social sciences, probabilities are usually expressed as proportions, and we will follow this convention throughout the remainder of this section. Using p to represent "probability," the probability of drawing the king of hearts (or any specific card) can be expressed as

$$p(\text{king of hearts}) = \frac{\#\ \text{successes}}{\#\ \text{events}} = \frac{1}{52} = 0.0192$$

As conceptualized here, probabilities have an exact meaning: Over the long run, the events that we define as successes will bear a certain proportional relationship to the total number of events. The probability of 0.0192 for selecting the king of hearts in a single draw really means that, over thousands of selections of one card at a time from a well-shuffled deck of 52 cards, the proportion of successful draws would be 0.0192. Or, for every 10,000 draws, 192 would be the king of hearts, and the remaining 9808 selections would be other cards.

Thus, when we say that the probability of drawing the king of hearts in one draw is 0.0192, we are essentially applying to a single draw our knowledge of what would happen over thousands of draws.

Like proportions, probabilities range from 0.00 (meaning that the event has absolutely no chance of occurrence) to 1.00 (a certainty). As the value of the probability increases, the likelihood that the defined event will occur also increases. A probability of 0.0192 is close to 0, and this means that the event (drawing the king of hearts) is unlikely, or improbable.

These techniques can be used to establish simple probabilities in any situation in which we can specify the number of successes and the total number of events. For example, a single die has six sides, or faces, each with a different value, ranging from 1 to 6. The probability of getting any specific number (say, a 4) in a single roll of a die is therefore

$$p(\text{rolling a 4}) = \frac{1}{6} = 0.1667$$

Probability and the Normal Curve

Combining this way of thinking about probability with our knowledge of the theoretical normal curve allows us to estimate the likelihood of selecting a case that has a score within a certain range. For example, suppose we wished to estimate the probability that a randomly chosen subject from the distribution of men's IQ scores would have an IQ score between 95 and the mean score of 100. Our definition of a success here would be the selection of any subject with a score in the specified range. Normally, we would next establish a fraction with the numerator equal to the number of subjects with scores in the defined range and the denominator equal to the total number of subjects. However, if the empirical distribution is normal in form, we can skip this step, because the probabilities, in proportion form, are already stated in Appendix A. That is, the areas in Appendix A can be interpreted as probabilities.

To determine the probability that a randomly selected case will have a score between 95 and the mean, we would convert the original score to a Z score:

$$Z = \frac{X_i - \bar{X}}{s} = \frac{95 - 100}{20} = -\frac{5}{20} = -0.25$$

Using Appendix A, we see that the area between this score and the mean is 0.0987. This is the probability we are seeking. The probability that a randomly selected case will have a score between 95 and 100 is 0.0987 (or, rounded off, 0.1, or 1 out of 10). In the same fashion, the probability of selecting a subject from any range of scores can be estimated. Note that the techniques for estimating probabilities are exactly the same as those for finding areas.

To consider an additional example, what is the probability that a randomly selected male will have an IQ less than 123? We will find probabilities in exactly the same way we found areas. The score (X_i) is above the mean and, following the directions in Table 5.4, we will find the probability we are seeking by adding the area in column b to 0.5000. First, we find the Z score:

$$Z = \frac{X_i - \overline{X}}{s} = \frac{123 - 100}{20} = \frac{23}{20} = +1.15$$

Next, look in column b of Appendix A to find the area between this score and the mean. Then add the area (0.3749) to 0.5000. The probability of selecting a male with an IQ of less than 123 is 0.3749 + 0.5000, or 0.8749. Rounding this value to 0.88, we can say that the odds are 0.88 (very high) that we will select a male with an IQ score in this range. Technically, remember that this probability expresses what would happen over the long run: For every 100 males selected from this group over an infinite number of trials, 88 would have IQ scores less than 123 and 12 would not.

Let me close by stressing a very important point about probabilities and the normal curve. The probability is very high that any case randomly selected from a normal distribution will have a score close in value to the mean. The shape of the normal curve is such that most cases are clustered around the mean and decline in frequency as we move farther away—either to the right or to the left—from the mean value. In fact, given what we know about the normal curve, the probability that a randomly selected case will have a score within ± 1 standard deviations of the mean is 0.6826. Rounding off, we can say that 68 out of 100 cases—or a little more than two-thirds of all cases—selected over the long run will have a score between ± 1 standard deviations, or Z scores, of the mean. The probabilities are high that any randomly selected case will have a score close in value to the mean.

In contrast, the probability of the case having a score beyond three standard deviations from the mean is very low. Look in column c ("Area Beyond Z") for a Z score of 3.00 and you will find the value 0.0014. Adding the area in the upper tail (beyond +3.00) to the area in the lower tail (beyond −3.00) gives us 0.0014 + 0.0014, for a total of 0.0028. The probability of selecting a case with a very high score or a very low score is 0.0028. If we randomly select cases from a normal distribution, we would select cases with Z scores beyond ± 3.00 only 28 times out of every 10,000 trials.

The general point to remember is that cases with scores close to the mean are common, and cases with scores that are far above or below the mean are rare. This relationship is central for an understanding of inferential statistics in Part II. *(For practice in using the normal curve table to find probabilities, see problems 5.8 to 5.10 and 5.13.)*

ONE STEP AT A TIME | **Finding Probabilities**

Step	*Operation*
1.	Compute the Z score (or scores). Note whether the score is positive or negative.
2.	Find the Z score (or scores) in the normal curve table (Appendix A).
3.	Find the area above or below the score (or between the scores) as you would normally (see the previous "One Step at a Time" boxes in this chapter) and express the result as a proportion. Typically, probabilities are expressed as a value between 0.00 and 1.00 rounded to two digits beyond the decimal point.

Applying Statistics 5.3 Finding Probabilities

The distribution of scores on a biology final exam used in Applying Statistics 5.2 had a mean of 72 and a standard deviation of 8. What is the probability that a student selected at random will have a score less than 61? More than 80? Less than 98? To answer these questions, we must first calculate *Z* scores and then consult Appendix A. We are looking for probabilities, so we will leave the areas in proportion form.

The *Z* score for a score of 61 is

$$Z = \frac{X_i - \overline{X}}{s} = \frac{61 - 72}{8} = -\frac{11}{8} = -1.38$$

This score is a negative value (below, or to the left of, the mean), and we are looking for the area below. Using Table 5.4 as a guide, we will find the area below a negative score in column c. This area is 0.0838. Rounding off, we can say that the odds of selecting a student with a score less than 61 are only 8 out of 100. This low value tells us this would be an unlikely event.

The *Z* score for the score of 80 is

$$Z = \frac{X_i - \overline{X}}{s} = \frac{80 - 72}{8} = \frac{8}{8} = +1.00$$

The *Z* score is positive, and to find the area above (greater than) 80, we look in column c (see Table 5.4). This value is 0.1587. The odds of selecting a student with a score greater than 80 is roughly 16 out of 100, about twice as likely as selecting a student with a score of less than 61.

The *Z* score for the score of 98 is

$$Z = \frac{X_i - \overline{X}}{s} = \frac{98 - 72}{8} = \frac{26}{8} = +3.25$$

To find the area below a positive *Z* score, we add the area between the score and the mean (column b) to 0.5000 (see Table 5.4). This value is 0.4994 + 0.5000, or 0.9994. It is extremely likely that a randomly selected student will have a score less than 98. Remember that scores more than ±3 standard deviations from the mean are very rare.

STATISTICS IN EVERYDAY LIFE

Probability Theory in Action

If you enjoy games of chance—bingo, card games, the state lottery, or board games like Parcheesi—you can use your knowledge of the laws of probability to improve your performance and your likelihood of winning. This point was illustrated in the best-seller *Bringing Down the House*, later made into a movie called *21*. The book describes the adventures of a group of Massachusetts Institute of Technology students who used their understanding of probability to develop a system that almost guaranteed they could beat the house in the game of blackjack. They won hundreds of thousands of dollars before they were finally stopped by casino security.

SUMMARY

1. The normal curve, in combination with the mean and standard deviation, can be used to construct precise descriptive statements about empirical distributions that are normally distributed. This chapter also lays some important groundwork for Part II.
2. To work with the theoretical normal curve, we must transform raw scores into their equivalent Z scores. Z scores allow us to find areas under the theoretical normal curve (Appendix A).
3. We considered three uses of the theoretical normal curve: finding total areas above and below a score, finding areas between two scores, and expressing these areas as probabilities. This last use of the normal curve is especially germane because inferential statistics are centrally concerned with estimating probabilities in a fashion very similar to the process introduced in this chapter.

SUMMARY OF FORMULAS

FORMULA 5.1 *Z* Scores: $Z = \frac{X_i - \overline{X}}{s}$

GLOSSARY

Normal curve. A theoretical distribution of scores that is symmetrical, unimodal, and bell-shaped. The standard normal curve always has a mean of 0 and a standard deviation of 1.

Normal curve table. Appendix A; a detailed description of the area between a *Z* score and the mean of any standardized normal distribution.

***Z* scores.** Standard scores; the way scores are expressed after they have been standardized to the theoretical normal curve.

PROBLEMS

5.1 Scores on a quiz were normally distributed and had a mean of 10 and a standard deviation of 3. For each of the following scores, find the *Z* score and the percentage of area above and below the score.

X_i	*Z* Score	% Area Above	% Area Below
5			
6			
7			
8			
9			
11			
12			
14			
15			
16			
18			

5.2 Assume that the distribution of scores on a college entrance exam is normal, with a mean of 500 and a standard deviation of 100. For each of the following scores, find the equivalent *Z* score, the percentage of the area above the score, and the percentage of the area below the score.

X_i	*Z* Score	% Area Above	% Area Below
650			
400			
375			
586			
437			
526			
621			
498			
517			
398			

5.3 The senior class has been given a comprehensive examination to assess educational experience. The mean score on the test was 74 and the standard deviation was 10. What percentage of the students had scores

a. between 75 and 85? ____________
b. between 80 and 85? ____________
c. above 80? ____________
d. above 83? ____________
e. between 80 and 70? ____________
f. between 75 and 70? ____________
g. below 75? ____________
h. below 77? ____________
i. below 80? ____________
j. below 85? ____________

5.4 For a normal distribution where the mean is 50 and the standard deviation is 10, what percentage of the area is

a. between the scores of 40 and 47? ____________
b. above a score of 47? ____________
c. below a score of 53? ____________
d. between the scores of 35 and 65? ____________
e. above a score of 72?____________
f. below a score of 31 and above a score of 69? ____________
g. between the scores of 55 and 62? ____________
h. between the scores of 32 and 47? ____________

5.5 At St. Algebra College, the 200 freshmen enrolled in introductory biology took a final exam on which their mean score was 72 and their standard deviation was 6. The following table presents the grades of 10 students. Convert each into a *Z* score and determine the *number of people* who scored higher or lower than each of the 10 students. *(HINT: Multiply the appropriate proportion by N and round the result.)*

X_i	Z Score	Number of Students Above	Number of Students Below
60			
57			
55			
67			
70			
72			
78			
82			
90			
95			

5.6 If a distribution of test scores is normal, with a mean of 78 and a standard deviation of 11, what percentage of the area lies

a. below 60? ____________
b. below 70? ____________
c. below 80? ____________
d. below 90? ____________
e. between 60 and 65? ____________
f. between 65 and 79? ____________
g. between 70 and 95? ____________
h. between 80 and 90? ____________
i. above 99? ____________
j. above 89? ____________
k. above 75? ____________
l. above 65? ____________

5.7 *SOC* A scale measuring prejudice has been administered to a large sample of respondents. The distribution of scores is approximately normal, with a mean of 31 and a standard deviation of 5. What percentage of the sample had scores

a. below 20? ____________
b. below 40? ____________
c. between 30 and 40? ____________
d. between 35 and 45? ____________
e. above 25? ____________
f. above 35? ____________

5.8 *SOC* On the scale mentioned in problem 5.7, if a score of 40 or more is considered "highly prejudiced," what is the probability that a person selected at random will have a score in that range?

5.9 For a math test on which the mean score was 59 and the standard deviation was 4, what is the probability that a student randomly selected from this class will have a score

a. between 55 and 65? ____________
b. between 60 and 65? ____________
c. above 65? ____________
d. between 60 and 50? ____________
e. between 55 and 50? ____________
f. below 55? ____________

5.10 *SOC* The average number of dropouts for a school district has been 305 per year with a standard deviation of 50. What is the probability that the number of dropouts next year will be

a. less than 250? ____________
b. less than 300? ____________
c. more than 350? ____________
d. more than 400? ____________
e. between 250 and 350? ____________
f. between 300 and 350? ____________
g. between 350 and 375? ____________

5.11 *CJ* The local police force gives all applicants an entrance exam and accepts only those applicants who score in the top 15% on this test. If the mean score this year is 87 and the standard deviation is 8, would an individual with a score of 110 be accepted?

5.12 To be accepted into an honor society, a student must have a grade-point average (GPA) in the top 10% of the school. If the mean GPA at the school is 2.78 and the standard deviation is 0.33, which of the following GPAs would qualify?

3.20, 3.21, 3.25, 3.30, 3.35

5.13 In a distribution of scores with a mean of 35 and a standard deviation of 4, which event is more likely: that a randomly selected score will be between 29 and 31 or that a randomly selected score will be between 40 and 42?

5.14 After taking the state merit examinations for the position of school counselor and social worker, you receive the following information on the tests and on your performance. On which test did you do better?

School Counselor	Social Worker
$\bar{X} = 118$	$\bar{X} = 27$
$s = 17$	$s = 3$
Your score = 127	Your score = 29

5.15 A large university administers entrance exams to all entering students. The test is administered again as an exit exam for all graduates. The test results for one group of students are given in the table.

Freshman Year	Senior Year
$\bar{X} = 53$	$\bar{X} = 92$
$s = 7$	$s = 4$

a. The scores for five students on both tests are listed below. For each student, calculate *Z* scores and determine whether the student performed better as a freshman or as a senior.

Student	Score, Freshman Year	Score, Senior Year
A	57	97
B	51	94
C	45	82
D	73	101
E	62	98

b. Determine the probabilities that randomly selected students will have scores in each of these ranges:

Freshman Year	
Score	Probability
Less than 52	
Less than 57	
Between 40 and 50	
More than 51	
More than 62	

Senior Year	
Score	Probability
Less than 88	
Less than 98	
Between 80 and 100	
More than 97	
More than 85	

Part II Inferential Statistics

The five chapters in this part cover the techniques and concepts of inferential statistics. Generally speaking, these applications allow us to learn about large groups (populations) from smaller, carefully selected subgroups (samples). These statistical techniques are powerful and extremely useful. They are used to poll public opinion, to research the potential market for new consumer products, to project the winners of elections, to test scientific hypotheses and theories, and in hundreds of other ways both inside and outside the social sciences.

Chapter 6 includes a brief description of sampling, but the most important part of this chapter concerns the sampling distribution—the single most-important concept in inferential statistics. The sampling distribution is normal in shape, and it is the key link between populations and samples. This chapter also covers estimation, the first of the two main applications in inferential statistics. In this section, you will learn how to use means and proportions computed from samples to estimate the characteristics of a population. This technique is most commonly used in public opinion polling and election projection.

Chapters 7 through 10 cover the second application of inferential statistics: hypothesis testing. Most of the relevant concepts are introduced in Chapter 7, and each chapter covers a different application of hypothesis testing. For example, Chapter 8 presents the techniques used when we are comparing information from two different samples (e.g., men vs. women), while Chapter 9 covers applications involving more than two samples (e.g., Republicans vs. Democrats vs. Independents).

Hypothesis testing is one of the more challenging aspects of statistics for beginning students, and I have included an abundance of learning aids to ease the chore of assimilating this material. Hypothesis testing is also one of the most common and important statistical applications to be found in social science research. Mastery of this material is essential for developing your ability to read the professional literature.

6 Introduction to Inferential Statistics, the Sampling Distribution, and Estimation

LEARNING OBJECTIVES

By the end of this chapter, you will be able to:

1. Explain the purpose of inferential statistics in terms of generalizing from a sample to a population.
2. Explain the basic techniques of random sampling and these key concepts: population, sample, parameter, statistic, representative, EPSEM.
3. Differentiate among the sampling distribution, the sample, and the population.
4. Explain the two theorems presented.
5. Explain the logic of estimation and the role of the sample, sampling distribution, and the population.
6. Define and explain the concepts of bias and efficiency.
7. Construct and interpret confidence intervals using sample means and sample proportions.
8. Use SPSS to get sample statistics to use in the construction of confidence intervals.

USING STATISTICS

The statistical concepts presented in this chapter are fundamental to all applications of inferential statistics, including the estimation of the characteristics of large populations from small samples. Estimation can be used to assess:

- Changes in values and attitudes (e.g., support for capital punishment or opinions about gay marriage) in the United States over the years.
- Levels of happiness and well-being in nations throughout the world.
- The effectiveness of drugs or other therapies for the treatment of disease.
- The appeal of candidates for political office in various segments of the voting population (e.g., among women or Catholics or Southerners).
- Immediate public reactions to controversial issues, such as new laws on immigration or health care.

One of the goals of social science research is to test our theories using many different people, groups, societies, and historical eras. Obviously, we can have the greatest confidence in theories that have stood up to testing against the greatest variety of cases and social settings. However, a major problem in social science research is that the most appropriate populations for our tests are very large. For example, a theory concerning political party preference among U.S. citizens would be most suitably tested using the entire electorate, but it is impossible to interview all 235 million adult Americans. Indeed, even for theories that reasonably could be tested with smaller populations—such as a local community or the student body at a university—the logistics of gathering data from every single case in the population are staggering to contemplate.

How can we test our theories if we can't access entire populations? To deal with this problem, social scientists select samples, or subsets of cases, from the populations of interest. Our goal in inferential statistics is to learn about the characteristics of a population (often called **parameters**) based on what we can learn from the sample. Two applications of inferential statistics are covered in this text. In estimation procedures, covered in this chapter, an "educated guess" of the population parameter is made, based on what is known about the sample. In hypothesis testing, covered in Chapters 7 through 10, the validity of a hypothesis about the population is tested against sample outcomes. Before we address these applications, we need to consider sampling (the techniques for selecting cases for a sample) and a key concept in inferential statistics: the sampling distribution.

Probability Sampling: Basic Concepts

Social scientists have developed a variety of sampling techniques. In this section, we will review the basic procedures for selecting probability samples, the only type of sample that fully supports the use of inferential statistics to generalize to populations. These samples are often described as *random*, and you may be more familiar with this terminology. Because of its greater familiarity, I will often use the term *random sample* in the following chapters. The term *probability sample* is preferred, however, because, in everyday language, *random* is often used to mean "by coincidence" or to give a connotation of unpredictability. As you will see, probability samples are selected carefully and methodically, in ways that leave no room for haphazardness. Interviewing the people you happen to meet in a mall one afternoon may be "random" in some sense of the word, but this technique will not result in a sample that could support inferential statistics.

Before considering probability sampling, let me point out that social scientists often use **nonprobability samples**. For example, social scientists studying small-group dynamics or the structure of attitudes or personal values might use the students enrolled in their classes as subjects. These "convenience" samples are very useful for a number of purposes (e.g., exploring ideas or pretesting survey forms before embarking on a more ambitious project) and are typically cheaper and easier to assemble. Their major limitation is that results cannot be generalized beyond the group being tested. If a theory of prejudice, for example, has been tested only on the students in a particular section of introductory sociology at a particular university, we cannot conclude that results would apply to other types of people. Therefore, we cannot place a lot of confidence in theories tested on nonprobability samples, even when they produce very strong evidence.

Selecting Representative Samples

When constructing a probability sample, our goal is to select cases so that the final sample is **representative** of the population from which it was drawn. A sample is representative if it reproduces the important characteristics of the population. For example, if the population consists of 60% females and 40% males, the sample should have about the same proportional makeup. In other words, a representative sample is very much like the population—only smaller.

It is crucial for inferential statistics that samples be representative. How can we assure ourselves that our samples are representative? Unfortunately, it is not possible to guarantee representativeness. However, we can maximize the chance of obtaining a representative sample by following the principle of EPSEM (the "**E**qual **P**robability of **SE**lection **M**ethod"), the fundamental principle of probability sampling. To follow EPSEM and maximize the probability that our sample will be representative, we select the sample so that every element or case in the population has an equal probability of being selected. Just as a die has an equal probability of showing any value from 1 to 6 when thrown, every case in the population must have the same chance of being included in the sample.

Remember that EPSEM and representativeness are two different things. In other words, the fact that a sample is selected according to EPSEM does not guarantee that it will be an exact representation of the population. The probability is very high that an EPSEM sample will be representative; but, just as a perfectly honest coin will sometimes show 10 heads in a row when flipped, an EPSEM sample will occasionally present an inaccurate picture of the population. One of the great strengths of inferential statistics is that they allow the researcher to estimate the probability of this type of error and interpret results accordingly.

Simple Random Sampling

The most basic EPSEM sampling technique produces a **simple random sample**. There are numerous variations and refinements on this technique, but in this text, we will consider only the simplest application.

To draw a simple random sample, we need a list of all cases in the population—each with its own unique identifier, usually a number—and a system of selection that guarantees that every case has an equal chance of being chosen for the sample. The selection process could involve flipping coins, throwing dice, or drawing numbers from a hat but, in social science research, we use computer programs to produce a list of random numbers. An example of a table of random numbers is available at the website for this text.

A case is selected for the sample when its identification number appears on the list. If the list of numbers is truly random, any specific ID number is just as likely as any other ID number, and the selection process will produce an EPSEM sample. Stop selecting cases when you have reached your desired sample size; if an identification number is selected more than once, ignore the repeats.[1]

[1]Ignoring identification numbers when they repeat is called "sampling without replacement" and, technically, this practice compromises the randomness of the selection process. However, if the sample is a small fraction of the population, we will be unlikely to select the same case twice and ignoring repeats will not bias our conclusions.

STATISTICS IN EVERYDAY LIFE

Sampling and Telephone Soliciting
The sampling techniques presented in this chapter are also used for commercial purposes and for mass marketing. The next time you are annoyed by a telemarketer calling during dinner, remember that your phone number may have been selected by a sophisticated, complex application of random sampling. This may not lessen your annoyance, but it will remind you that sampling technology is a part of daily life and in many areas outside of the social sciences.

Remember that the purpose of inferential statistics is to acquire knowledge of populations based on information gathered from samples of that population. Each of the applications of inferential statistics to be presented in this text requires that samples be selected according to the EPSEM. While even the most painstaking and sophisticated sampling technique will not guarantee representativeness, the probability is high that EPSEM samples will be representative of the populations from which they were selected.

The Sampling Distribution

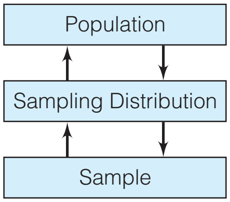
Once we have selected a probability sample according to some EPSEM procedure, what do we know? On one hand, we can gather a great deal of information from the cases in the sample. On the other hand, we know nothing about the population. Indeed, if we had information about the population, we probably wouldn't need the sample. Remember that we use inferential statistics to learn more about populations, and information from the sample is important primarily insofar as it allows us to generalize to the population.

When we use inferential statistics, we generally measure some variable (e.g., age, political party preference, or opinions about abortion) in the sample and then use the information from the sample to learn more about that variable in the population. In Part I of this text, you learned that three types of information are necessary to describe a variable adequately:

1. The shape of its distribution,
2. A measure of central tendency, and
3. A measure of dispersion.

Clearly, all three kinds of information can be gathered (or computed) on a variable from the cases in the sample. Just as clearly, none of the information is available for the population. Except in rare situations (for example, IQ tests are designed so that the scores of a population will be approximately normal in distribution, and the distribution of income is almost always positively skewed), nothing can be known about the exact shape of the distribution of a variable in the population. The means and standard deviations of variables in the population are also unknown. If we had this information for the population, inferential statistics would be unnecessary.

FIGURE 6.1 The Relationships Between the Sample, the Sampling Distribution, and the Population



In statistics, we link information from the sample to the population with the **sampling distribution**: *the theoretical, probabilistic distribution of a statistic for all possible samples of a certain size (N).* That is, the sampling distribution is the distribution of a statistic (e.g., a mean or a proportion) based on every conceivable combination of cases from the population. A crucial point about the sampling distribution is that its characteristics are based on the laws of probability, not on empirical information, and are very well known. In fact, the sampling distribution is the central concept in inferential statistics, and a prolonged examination of its characteristics is certainly in order.

The Three Distributions Used in Inferential Statistics. As illustrated by Figure 6.1, all applications of inferential statistics move between the sample and the population via the sampling distribution. Thus, three separate and distinct distributions of a variable are involved in every application of inferential statistics:

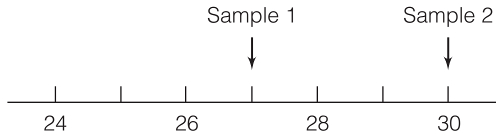
1. The *population distribution* of the variable, which, while empirical, is unknown. Amassing information about or making inferences to the population is the sole purpose of inferential statistics.
2. The *sampling distribution* of the variable, which is nonempirical, or theoretical. Because of the laws of probability, a great deal is known about this distribution. Specifically, the shape, central tendency, and dispersion of the distribution can be deduced and, therefore, the distribution can be adequately characterized.
3. The *sample distribution* of the variable, which is empirical (i.e., it exists in reality) and is known. The shape, central tendency, and dispersion of the variable can be ascertained for the sample. Remember, however, that information from the sample is important primarily insofar as it allows the researcher to learn about the population.

The utility of the sampling distribution is implied by its definition. Because it includes the statistics from all possible sample outcomes, the sampling distribution enables us to estimate the probability of any particular sample outcome, a process that will occupy our attention for this and the next four chapters.

Constructing a Sampling Distribution. The sampling distribution is theoretical, which means that it is never actually constructed. However, to understand better the structure and function of the distribution, let's consider an example of how one might be constructed. Suppose we wanted to gather some information about the ages of a particular community of 10,000 individuals. We draw an EPSEM sample of 100 residents, ask all 100 respondents their age, and use those individual scores to compute a mean age of 27. This score is noted on the graph in Figure 6.2. Note that this sample is one of countless possible combinations of 100 people taken from this population of 10,000, and the statistic (the mean of 27) is one of millions of possible sample outcomes.

Now replace the first 100 respondents, draw another sample of the same size ($N = 100$), and again compute the average age. Assume that the mean for the second sample is 30 and note this sample outcome on Figure 6.2. This second sample

FIGURE 6.2 Constructing a Sampling Distribution



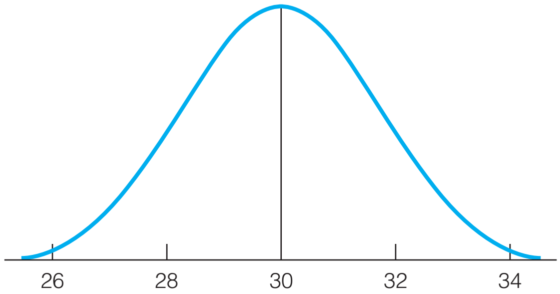
is another of the countless combinations of 100 people taken from this population of 10,000, and the sample mean of 30 is another of the millions of possible sample statistics. Replace these respondents and draw still another sample, calculate and note the mean, replace this third sample, and draw a fourth sample, continuing these operations an infinite number of times, calculating and noting the mean of each sample. Now try to imagine what Figure 6.2 would look like after tens of thousands of individual samples had been collected and the mean had been computed for each sample. What shape, mean, and standard deviation would this distribution of sample means have after we had collected all possible combinations of 100 respondents from the population of 10,000?

For one thing, we know that each sample will be at least slightly different from every other sample, since it is very unlikely that we will sample exactly the same 100 people twice. Because each sample will be unique, each sample mean will be at least slightly different in value.

We also know that not all the samples will be representative, even if we follow EPSEM closely. For example, if we continue taking samples of 100 people long enough, we will eventually choose a sample that includes only the very youngest residents. Such a sample would have a mean much lower than the true population mean. Likewise, some of our samples will include only senior citizens and will have means that are much higher than the population mean. Common sense suggests, however, that such nonrepresentative samples will be rare and that most sample means will cluster around the true population value.

To illustrate further, assume that we somehow come to know that the true mean age of the population of 10,000 individuals is 30. Because, as we have just seen, most of the sample means will also be approximately 30, the sampling distribution of these sample means should peak at 30. Some of the samples will be nonrepresentative and their means will "miss the mark," but the frequency of such misses should decline as we get farther away from the population mean of 30. That is, the distribution should slope to the base as we move away from the population value—sample means of 29 or 31 should be common; means of 20 or 40 should be rare. The samples are random, so their means should miss an equal number of times on either side of the population value, and the distribution itself should therefore be roughly symmetrical. In other words, the sampling distribution of all possible sample means should be approximately normal and will resemble the distribution presented in Figure 6.3. Recall from Chapter 5 that, on any normal curve, cases close to the mean (say, within ± 1 standard deviation) are common and cases far away from the mean (say, beyond 3 standard deviations) are rare. *(See problem 6.20 for a demonstrations of the construction of sampling distributions.)*

FIGURE 6.3 A Sampling Distribution of Sample Means



Two Theorems. These commonsense notions about the shape of the sampling distribution and other very important information about central tendency and dispersion are stated in two theorems. Before examining the theorems, we need to review some of the symbols we will use. Recall that the symbol for the mean of a sample is $\overline{X}$ and that μ is the symbol for the mean of a population. For measures of dispersion, we use s to symbolize the standard deviation of a sample and σ (lowercase Greek letter sigma) to refer to the standard deviation of a population.

The first of these theorems states:

> If repeated random samples of size N are drawn from a normal population with mean μ and standard deviation σ, then the sampling distribution of sample means will be normal, with a mean μ and a standard deviation of $\sigma/\sqrt{N}$.

To translate: If we begin with a trait that is normally distributed across a population (IQ scores, for example) and take an infinite number of equal-sized random samples from that population, then the sampling distribution of sample means will be normal. If we know that the variable is distributed normally in the population, we can assume that the sampling distribution will be normal.

The theorem tells us more than the shape of the sampling distribution, however. It also defines its mean and standard deviation. In fact, it says that the mean of the sampling distribution will be exactly the same value as the mean of the population. That is, if we know that the mean IQ of the entire population is 100, then we know that the mean of any sampling distribution of sample mean IQ scores will also be 100.

Exactly why these two means will be the same value cannot be fully explained at this level. Recall, however, that most sample means will cluster around the population value over the long run. Thus, the fact that these two values are equal should have intuitive appeal. As for dispersion, the theorem says that the standard deviation of the sampling distribution, also called the **standard error of the mean**, will be equal to the standard deviation of the population divided by the square root of N (symbolically: $\sigma/\sqrt{N}$.).

If the mean and standard deviation of a normally distributed variable in a population are known, the theorem allows us to compute the mean and standard deviation of the sampling distribution.[2] Thus, we will know exactly as much about the sampling distribution (shape, central tendency, and dispersion) as we ever knew about any empirical distribution.

The first theorem requires that the variable be normally distributed in the population. What happens when the distribution of the variable in question is unknown or is known *not* to be normal in shape (e.g., income, which always has a positive skew)? These eventualities (very common, in fact) are covered by a second theorem, called the **Central Limit Theorem**:

> If repeated random samples of size N are drawn from any population with mean μ and standard deviation σ, then, as N becomes large, the sampling distribution of sample means will approach normality, with mean μ and standard deviation $\sigma/\sqrt{N}$.

To translate: The sampling distribution of sample means will become normal in shape as sample size increases for *any* variable, even when the variable is not normally distributed across the population. When N is large, the mean of the sampling distribution will equal the population mean, and its standard deviation (or the standard error of the mean) will be equal to $\sigma/\sqrt{N}$.

The Central Limit Theorem is important because it removes the condition that the variable be normally distributed in the population. Whenever sample size is large, we can assume that the sampling distribution is normal in shape, with a mean equal to the population mean and a standard deviation equal to $\sigma/\sqrt{N}$. Thus, even if we are working with a variable that is known to have a skewed distribution (like income), we can still assume a normal sampling distribution.

The issue remaining, of course, is to define what is meant by a large sample. A good rule of thumb is that if sample size (N) is 100 or more, the Central Limit Theorem applies, and you can assume that the sampling distribution of sample statistics is normal in shape. When N is less than 100, you must have good evidence of a normal population distribution before you can assume that the sampling distribution is normal. Thus, as long as samples are larger than 100, we can assume that the sampling distribution will be normal in shape.

STATISTICS IN EVERYDAY LIFE

Using the GSS to Measure Changing American Attitudes

The General Social Survey (GSS) has been administered to representative samples of adult Americans since 1972, and this permits us to track changes in our society over time. As one illustration, data from the GSS demonstrate a considerable softening in the tendency of Americans to condemn homosexuality. In the 1970s, about 72% of respondents felt that homosexuality was “always wrong.” In the most recent administration of the GSS (2012), that percentage had fallen to about 46%.

[2]In the typical research situation, the values of the population mean and standard deviation are, of course, unknown. However, these values can be estimated from sample statistics, as we shall see in the chapters that follow.

The Sampling Distribution: An Additional Example Using the General Social Survey (GSS)

Developing an understanding of the sampling distribution—what it is and why it's important—can be a challenging task. It may be helpful to list briefly the most important points about the sampling distribution:

1. Its definition: *The sampling distribution is the distribution of a statistic (such as a mean or a proportion) for all possible sample outcomes of a certain size.*
2. Its shape: *normal* (see Chapter 5 and Appendix A).
3. Its central tendency and dispersion: *The mean of the sampling distribution is the same value as the mean of the population. The standard deviation of the sampling distribution—or the standard error—is equal to the population standard deviation divided by the square root of N (see the theorems).*
4. The role of the sampling distribution in inferential statistics: *It links the sample with the population* (see Figure 6.1).

To reinforce these points, let's consider an additional example using the General Social Survey (GSS), one of the databases used for SPSS exercises in this text. The GSS has been administered to randomly selected samples of adult Americans since 1972 and explores a broad range of characteristics and issues, including confidence in the Supreme Court, attitudes about assisted suicide, number of siblings, and level of education. The GSS has its limits, of course, but it is a very valuable resource for testing theory and for learning more about American society. Focusing on this survey, let's review the roles played by the population, the sample, and the sampling distribution when we use this database.

The GSS and the Population

We'll start with the population, or the group we want to learn more about. In the case of the GSS, the population consists of all adult (older than 18) Americans, which includes 235 million people. Clearly, we can never interview all of these people and learn what they are like or what they are thinking about abortion, gun control, sex education in the public schools, or any other issue. We should also note that this information is worth having. It could help inform public debates, provide some basis in fact for the discussion of many controversial issues (e.g., the polls show consistently that the majority of Americans favor some form of gun control), and assist people in clarifying their personal beliefs. If the information is valuable, how can we learn more about this huge population?

The GSS and the Sample

This brings us to the sample, a carefully chosen subset of the population. The General Social Survey is administered to several thousand people, each of whom is chosen by a sophisticated technology based on the principle of EPSEM. A key point to remember is that samples chosen by this method are very likely to be representative of the populations from which they were selected. Whatever is true of the sample will also be true of the population (with some limits and qualifications, of course).

The respondents are contacted at home and asked for background information (religion, gender, years of education, and so on) as well as their opinions and attitudes. When all of this information is collated, the GSS database includes information (shape, central tendency, dispersion) on hundreds of variables (age, level of prejudice, marital status) for the people in the sample. So we have a lot of information about the variables for the *sample* (the people who actually responded to the survey), but no information about these variables for the *population* (the 235 million adult Americans). How do we get from the known sample to the unknown population? This is the central question of inferential statistics; the answer, as you hopefully realize by now, is "by using the sampling distribution."

The GSS and the Sampling Distribution

Remember that, unlike the sample and the population, the sampling distribution is theoretical, and, because of the theorems presented earlier in this chapter, we know its shape, central tendency, and dispersion. For any variable from the GSS, the theorems tell us that:

- The sampling distribution will be normal in shape because the sample is "large" (N is much greater than 100). This will be true regardless of the shape of the variable in the population.
- The mean of the sampling distribution will be the same value as the mean of the population. If *all* adult Americans have completed an average of 13.5 years of schooling ($\mu = 13.5$), then the mean of the sampling distribution will also be 13.5.
- The standard deviation (or standard error) of the sampling distribution is equal to the population standard deviation (σ) divided by the square root of N.

Thus, the theorems tell us the statistical characteristics of the sampling distribution (shape, central tendency, and dispersion), and this information allows us to link the sample to the population.

How does the sampling distribution link the sample to the population? The fact that the sampling distribution will be normal when N is large is crucial. This means that more than two-thirds (68%) of all samples will be within ± 1 Z score of the mean of the sampling distribution (which is the same value as the mean of the population), about 95% are within ± 2 Z scores, and so forth. We do not (and cannot) know the actual value of the mean of the sampling distribution, but we do know that the probabilities are very high that our sample statistic is approximately equal to this parameter. Similarly, the theorems give us crucial information about the mean and standard error of the sampling distribution that we can use, as you will see later in this chapter and in the chapters that follow, to link information from the sample to the population.

To summarize, our goal is to infer information about the population (in the case of the GSS, all adult Americans). When populations are too large to test, we use information from randomly selected samples, carefully drawn from the

STATISTICS
IN EVERYDAY LIFE

Using Surveys to Compare the United States with Other Nations

The World Values Survey (http://www.worldvaluessurvey.org/) has been administered to random samples from many nations since 1981. The latest results show that the United States is more tolerant regarding homosexuality than some nations and less tolerant than others.

"Please tell me if you think homosexuality can always be justified, never be justified, or something in between."

Nation	Percentage "Never Justifiable"
Russia	54.1%
China	49.4%
United Sates	24.0%
Germany	17.8%
Japan	17.6%

Source: World Values Survey, http://www.worldvaluessurvey.org/. Results are for Wave 6 of the survey, 2010–2014.

population of interest, to estimate the characteristics of the population. In the case of the GSS, the full sample consists of several thousand adult Americans who have responded to the questions on the survey. The sampling distribution, the theoretical distribution whose characteristics are defined by the theorems, links the known sample to the unknown population.

Symbols and Terminology

In the remainder of this chapter and in the chapters to follow, we will be working with three entirely different distributions. Furthermore, we will be concerned with several different kinds of sampling distributions—including the sampling distribution of sample means and the sampling distribution of sample proportions.

To distinguish among these various distributions, we will often use symbols. The symbols for the means and standard deviations of samples and populations have already been introduced in Chapters 3 and 4. For quick reference, Table 6.1 presents the symbols that will be used for the sampling distribution. Basically, the sampling distribution is denoted with Greek letters that are subscripted according to the sample statistic of interest.

TABLE 6.1 Symbols for Means and Standard Deviations of Three Distributions

	Mean	Standard Deviation	Proportion
1. Samples	$\bar{X}$	s	P_s
2. Populations	μ	σ	P_u
3. Sampling distributions			
of means	$\mu_{\bar{X}}$	$\sigma_{\bar{X}}$	
of proportions	μ_p	σ_p	

Note that the mean and standard deviation of a sample are denoted with English letters ($\overline{X}$ and s), while the mean and standard deviation of a population are denoted with Greek-letter equivalents (μ and σ). Proportions calculated on samples are symbolized as $\boldsymbol{P_s}$ (P-sub-s; s for sample), while population proportions are denoted as $\boldsymbol{P_u}$ (P-sub-u; u for "universe" or population). The symbols for the sampling distribution are Greek letters with English-letter subscripts. The mean and standard deviation of a sampling distribution of sample means are $\mu_{\overline{X}}$ ("mu-sub-*x*-bar") and $\sigma_{\overline{X}}$ ("sigma-sub-*x*-bar"). The mean and standard deviation of a sampling distribution of sample proportions are μ_p ("mu-sub-*p*") and σ_p ("sigma-sub-*p*").

Introduction to Estimation

The object of this branch of inferential statistics is to estimate population values or parameters from statistics computed from samples. You are already familiar with public-opinion polls and election projections, the most common applications of these techniques. Polls and surveys on every conceivable issue—from the sublime to the trivial—have become a staple of mass media and popular culture. The statistical techniques you will learn in this chapter are the same as those used by the most reputable, sophisticated, and scientific pollsters.

The standard procedure for estimating population values is to construct a **confidence interval**, a mathematical statement that says that the parameter lies within a certain interval or range of values. For example, a confidence interval might say "68% ± 3%—or, between 65% and 71%—of Americans approve of capital punishment." In the media, the central value of the interval (68% in this case) is usually stressed, but it is important to realize that the population parameter (the percentage of *all* Americans who approve of capital punishment) could be anywhere in the interval between 65% and 71%.

Estimation Selection Criteria: Bias and Efficiency

Estimation procedures are based on sample statistics. Which of the many available sample statistics should be used? Estimators can be selected according to two criteria: **bias** and **efficiency**. Estimates should be based on sample statistics that are unbiased and relatively efficient. We cover each of these criteria separately.

Bias

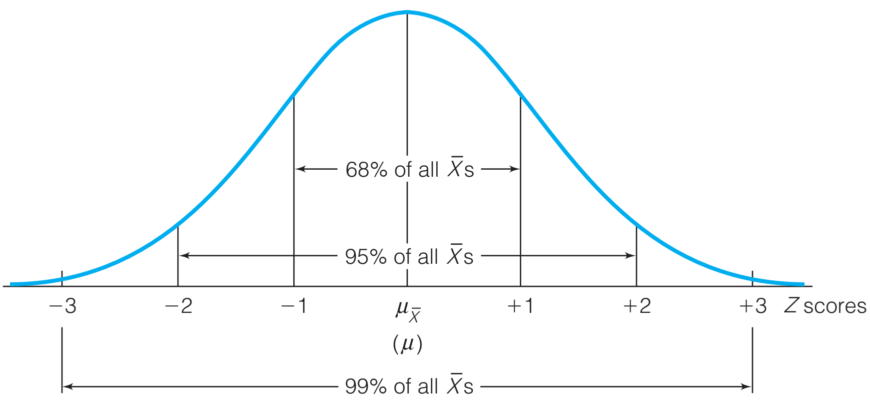
An estimator is unbiased if the mean of its sampling distribution is equal to the population value of interest. We know from the theorems presented earlier in this chapter that sample means meet this criterion. The mean of the sampling distribution of sample means (which we will note symbolically as $\mu_{\overline{X}}$) is the same as the population mean (μ).

Sample proportions (P_s) are also unbiased. If we calculate sample proportions from repeated random samples of size N, the sampling distribution of sample proportions will have a mean (μ_p) equal to the population proportion (P_u). Thus, if we are concerned with coin flips and sample honest coins 10 at a time ($N = 10$), the sampling distribution will have a mean equal to 0.5, which is the probability that an honest coin will show heads (or tails) when flipped. Of the statistics commonly used in social research, only sample means and sample proportions are unbiased (that is, have sampling distributions with means equal to the population value).

Because sample means and sample proportions are unbiased, we can determine the probability that they are within a certain distance of the population values we are trying to estimate. To illustrate, consider a specific problem. Assume that we wish to estimate the average income of a community. A random sample of 500 households is taken ($N = 500$), and a sample mean income of $45,000 is computed. In this example, the population mean (μ) is the average income of *all* households in the community, and the sample mean ($\bar{X}$) is the average income for the 500 households that happened to be selected for our sample. Note that we do not know the value of the population mean (μ)—if we did, we wouldn't need the sample—but it is μ that we are interested in. The sample mean of $45,000 is important primarily insofar as it can give us information about the population mean.

The two theorems presented earlier in this chapter give us a great deal of information about the sampling distribution of all possible sample means in this situation. Because N is large ($N > 100$), we know that the sampling distribution is normal and that it has a mean equal to the population mean ($\mu_{\bar{X}} = \mu$). We also know that all normal curves contain about 68% of the cases (the cases here are sample means) within ± 1 Z, 95% of the cases within ± 2 Zs, and more than 99% of the cases within ± 3 Zs of the mean. Remember that we are discussing the sampling distribution here—the distribution of all possible sample outcomes, or, in this instance, sample means. Thus, the probabilities are very good (approximately 68 out of 100 chances) that our sample mean of $45,000 is within ± 1 Z, excellent (95 out of 100) that it is within ± 2 Zs, and overwhelming (99 out of 100) that it is within ± 3 Zs of the mean of the sampling distribution (which is the same value as the population mean). These relationships are graphically depicted in Figure 6.4.

FIGURE 6.4 Areas Under the Sampling Distribution of Sample Means



If an estimator is unbiased, it is probably an accurate estimate of the population parameter (μ in this case). However, in less than 1% of the cases, a sample mean will be more than ± 3 *Z*s away from the mean of the sampling distribution (very inaccurate) by random chance alone. We literally have no idea whether our particular sample mean of \$45,000 is in this small minority. We do know, however, that it is very likely that our sample mean is considerably closer than ± 3 *Z*s to the mean of the sampling distribution and, thus, to the population mean.

Efficiency

The second desirable characteristic of an estimator is efficiency, which is the extent to which the sampling distribution is clustered about its mean. Efficiency, or clustering, is essentially a matter of dispersion (see Figure 4.1). The smaller the standard deviation of a sampling distribution, the greater the clustering and the higher the efficiency.

The standard deviation of the sampling distribution of sample means, or the standard error of the mean, is equal to the population standard deviation divided by the square root of *N* ($\sigma_{\bar{X}} = \sigma/\sqrt{N}$). Therefore, the standard deviation of the sampling distribution is an inverse function of *N*: As sample size increases, $\sigma_{\bar{X}}$ will decrease. We can improve the efficiency (or decrease the standard deviation of the sampling distribution) for any estimator by increasing sample size.

An example should make this clearer. Consider the two samples listed in Table 6.2. Both sample means are unbiased, but which is the more efficient estimator? For sample 1, the standard deviation of the sampling distribution of all possible means with an *N* of 100 would be \$50.00:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \frac{500}{\sqrt{100}} = \frac{500}{10} = 50.00$$

For sample 2, the standard deviation of all possible sample means with an *N* of 1000 would be a much smaller value of \$15.81:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \frac{500}{\sqrt{1000}} = \frac{500}{31.62} = 15.81$$

The sampling distribution based on sample 2 is much more clustered than the sampling distribution based on sample 1. In fact, it contains 68% of all possible sample means within ± 15.81 of μ, while the sampling distribution based on sample 1 requires a much broader interval of ± 50.00 to do the same. The estimate based on a sample with 1000 cases is much more likely to approximate

TABLE 6.2 Standard Deviation of the Sampling Distribution of Two Samples, Assuming σ = \$500

	Sample 1	Sample 2
Sample mean	$\bar{X}_1$ = \$45,000	$\bar{X}_2$ = \$45,000
Sample size	N_1 = 100	N_2 = 1000
Standard deviation of the sampling distribution (or Standard error)	$\sigma_{\bar{X}_1}$ = \$50.00	$\sigma_{\bar{X}_2}$ = \$15.81

FIGURE 6.5 A Sampling Distribution with N = 100 and $\sigma_{\bar{X}}$ = \$50.00

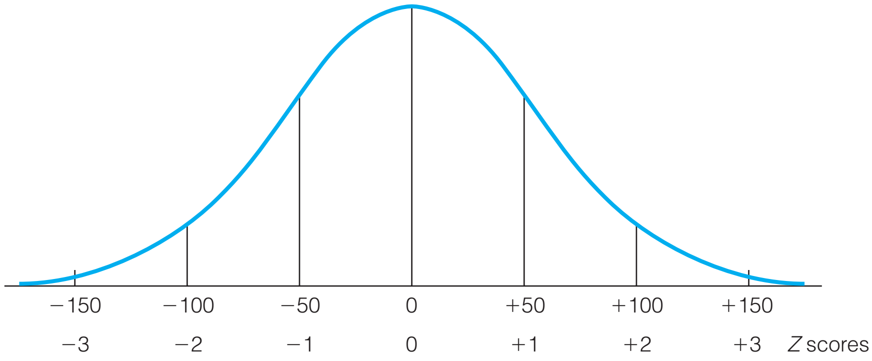
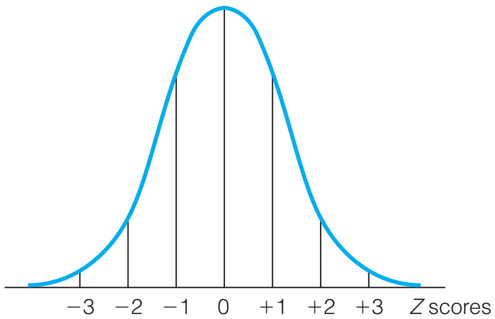


FIGURE 6.6 A Sampling Distribution with N = 1000 and $\sigma_{\bar{X}}$ = \$15.81



the population parameter than the estimate based on a sample of 100 cases. Figures 6.5 and 6.6 illustrate these relationships graphically.

The key point to remember is that the standard deviation of all sampling distributions is an inverse function of *N*: The larger the sample, the greater the clustering and the higher the efficiency. In part, these relationships between sample size and the standard deviation of the sampling distribution do nothing more than underscore our commonsense notion that more confidence can be placed in large samples than in small (as long as both have been selected according to EPSEM).

Interval Estimation Procedures

We are now ready to estimate population values based on sample statistics. We will do this by constructing confidence intervals or mathematical statements that say that the parameter is within a range of values. Confidence intervals are constructed in three steps.

The first step is to decide on the risk you are willing to take of being wrong. An interval estimate is wrong if it does *not* include the population parameter. This probability of error is called **alpha** (symbolized by Greek letter α). The exact value of alpha will depend on the nature of the research situation, but a 0.05 probability is commonly used. Setting alpha equal to 0.05—also called using the 95% **confidence level**—means that over the long run, the researcher is willing to be wrong only 5% of the time. Or, to put it another way, if an infinite number of intervals were constructed at this alpha level (and with all other things being equal), 95% of them would contain the population value and 5% would not. In reality, of course, only one interval is constructed, and, by setting the probability of error very low, we are setting the odds in our favor that the interval will include the population value.

The second step is to picture the sampling distribution, divide the probability of error equally into the upper and lower tails of the distribution, and then find the corresponding *Z* score. For example, if we decided to set alpha equal to 0.05, we would place half (0.025) of this probability in the lower tail and half in the upper tail of the distribution. The sampling distribution would thus be divided as illustrated in Figure 6.7.

We need to find the *Z* score that marks the beginnings of the shaded areas. In Chapter 5, we learned how to calculate *Z* scores and find areas under the normal curve. Here, we will reverse that process. We need to find the *Z* score beyond which lies a proportion of 0.0250 of the total area. To do this, go down column c of Appendix A until you find the proportion 0.0250. The associated *Z* score is 1.96. Because the curve is symmetrical and we are interested in both the upper and lower tails, we designate the *Z* score that corresponds to an alpha of 0.05 as ± 1.96 (see Figure 6.8).

We now know that 95% of all possible sample outcomes fall within ± 1.96 *Z* scores of the population value. In reality, of course, there is only one sample outcome, but, if we construct an interval estimate based on ± 1.96 *Z*'s, the probabilities are that 95% of all such intervals will trap the population value. Thus, we can be 95% confident that our interval contains the population value.

Besides the 95% level, there are four other commonly used confidence levels: the 90% level ($\alpha = 0.10$), the 99% level ($\alpha = 0.01$), the 99.9% level

FIGURE 6.7 The Sampling Distribution with Alpha (α) Equal to 0.05

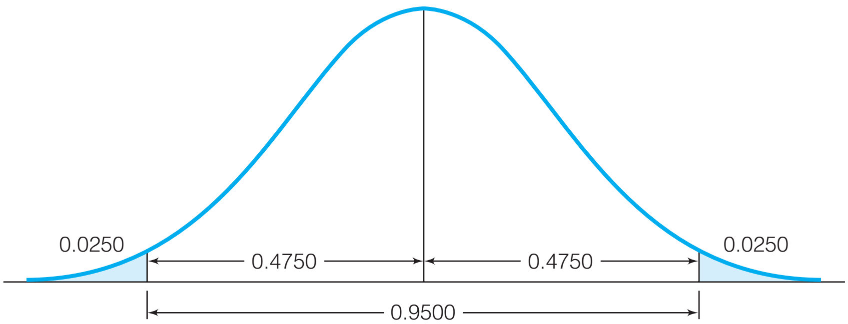
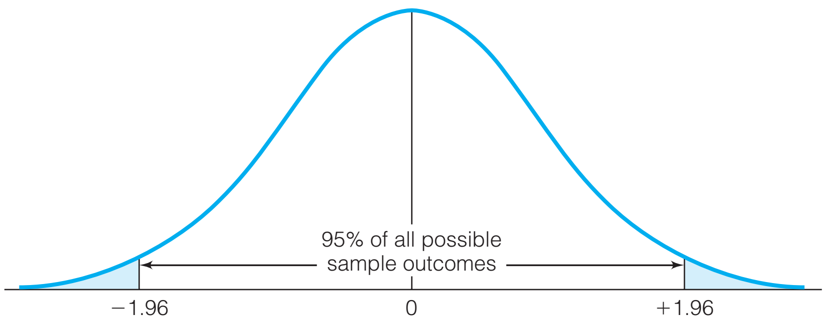


FIGURE 6.8 The Sampling Distribution with Alpha (α) Equal to 0.05



($\alpha = 0.001$), and the 99.99% level ($\alpha = 0.0001$). To find the corresponding Z scores for these levels, follow the procedures outlined earlier for an alpha of 0.05. Table 6.3 summarizes all the information you will need.

You should turn to Appendix A and confirm for yourself that the Z scores in Table 6.3 correspond to these alpha levels. As you do, note that, in the cases where alpha is set at 0.10 and 0.01, the precise areas we seek are not in the table. For example, with an alpha of 0.10, we would look in column c ("Area Beyond Z") for the area 0.0500. Instead we find an area of 0.0505 ($Z = \pm 1.64$) and an area of 0.0495 ($Z = \pm 1.65$). The Z score we are seeking is somewhere between these two other scores. When this occurs, take the larger of the two scores as Z. This will make the interval as wide as possible under the circumstances and is thus the most conservative course of action.

For the alpha of 0.01, we encounter the same problem (the exact area 0.0050 is not in the table), resolve it the same way, and take the larger score as Z. For the alpha of 0.001, we take the largest of the several scores listed for the area as our Z score. Finally, for the lowest alpha of 0.0001, the table is not detailed enough to show exact areas, and we will use ± 3.90 as our Z score. *(For practice in finding Z scores for various levels of confidence, see problem 7.3.)*

The third step is to construct the confidence interval. In the sections that follow, we construct interval estimates of population parameters, first using sample means and then using sample proportions.

TABLE 6.3 *Z* Scores for Various Levels of Alpha (α)

Confidence Level	Alpha (α)	$\alpha/2$	Z Score
90%	0.10	0.05	±1.65
95%	0.05	0.025	±1.96
99%	0.01	0.005	±2.58
99.9%	0.001	0.0005	±3.32
99.99%	0.0001	0.00005	±3.90

Interval Estimation Procedures Using Sample Means (Large Samples)

The formula for constructing a confidence interval based on sample means is given in Formula 6.1.

FORMULA 6.1

$$c.i. = \bar{X} \pm Z\left(\frac{\sigma}{\sqrt{N}}\right)$$

where $c.i.$ = confidence interval
$\bar{X}$ = the sample mean
Z = the Z score as determined by the alpha level
$\left(\frac{\sigma}{\sqrt{N}}\right)$ = the standard deviation of the sampling distribution, or the standard error of the mean

As an example, suppose you wanted to estimate the average IQ of a community and had randomly selected a sample of 200 residents, with a sample mean IQ of 105. Assume that the population standard deviation for IQ scores is about 15, so we can set σ equal to 15. If we are willing to run a 5% chance of being wrong and set alpha at 0.05, the corresponding Z score will be ± 1.96. These values can be substituted directly into Formula 6.1, and an interval can be constructed:

$$c.i. = \bar{X} \pm Z\left(\frac{\sigma}{\sqrt{N}}\right)$$

$$c.i. = 105 \pm (1.96)\left(\frac{15}{\sqrt{200}}\right)$$

$$c.i. = 105 \pm (1.96)\left(\frac{15}{14.14}\right)$$

$$c.i. = 105 \pm (1.96)(1.06)$$

$$c.i. = 105 \pm 2.08$$

That is, our estimate is that the average IQ for the population in question is somewhere between 102.92 (105 − 2.08) and 107.08 (105 + 2.08). Because 95% of all possible sample means are within ± 1.96 Z's (or 2.08 IQ units in this case) of the mean of the sampling distribution, the odds are very high that our interval will contain the population mean. In fact, even if the sample mean is as far off as ± 1.96 Zs (which is unlikely), our interval will still contain the mean of the sampling distribution and, thus, the population mean (μ). Only if our sample mean is one of the few that is more than ± 1.96 Zs from the mean of the sampling distribution will we have failed to include the population mean.

Note that the value of the population standard deviation was supplied in our example. Needless to say, it is unusual to have such information about a population. In the great majority of cases, we will not know the value of σ, but we can estimate σ with s, the sample standard deviation. Unfortunately, s is a biased estimator of σ, and the formula must be changed slightly to correct for the bias. For larger samples, the bias of s will not affect the interval very much. The revised formula for cases in which s is unknown is

FORMULA 6.2

$$c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{N-1}}\right)$$

In comparing this formula with Formula 6.1, note that there are two changes. First, σ is replaced by s, and, second, the denominator of the last term is the square root of $N - 1$ rather than the square root of N. The latter change is the correction for the fact that s is biased.

Let me stress here that the substitution of s for σ is permitted only for large samples (that is, samples with 100 or more cases). For smaller samples, when the value of the population standard deviation is unknown, the standardized normal distribution summarized in Appendix A cannot be used in the estimation process. To construct confidence intervals from sample means with sample sizes smaller than 100, we must use a different theoretical distribution, called the Student's t distribution, to find areas under the sampling distribution. We defer the presentation of the t distribution until Chapter 7 and confine our attention here to estimation procedures for large samples only.

We will close this section by working through a sample problem with Formula 6.2. Average income for a random sample of 500 residents of a particular community is \$45,000, with a standard deviation of \$200. What is the 95% interval estimate of the population mean income, μ?

Given that

$$\bar{X} = \$45{,}000$$
$$s = \$200$$
$$N = 500$$

and using an alpha of 0.05, we can construct the interval:

$$c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{N-1}}\right)$$

$$c.i. = 45{,}000 \pm (1.96)\left(\frac{200}{\sqrt{499}}\right)$$

$$c.i. = 45{,}000 \pm (1.96)\left(\frac{200}{22.34}\right)$$

$$c.i. = 45{,}000 \pm (1.96)(8.95)$$

$$c.i. = 45{,}000 \pm 17.54$$

The average income for the community as a whole is between \$44,982.46 (45,000 − 17.54) and \$45,017.54 (45,000 + 17.54). Remember that this interval has only a 5% chance of being wrong (that is, of not containing the population mean). *(For practice with confidence intervals using sample means, see problems 6.1, 6.4–6.7, 6.18, and 6.19a.)*

ONE STEP AT A TIME — Constructing Confidence Intervals for Sample Means by Using Formula 6.2

Step	***Operation***
1.	Select an alpha level and find the associated Z score in Table 6.3. If you use the alpha level of 0.05, the *Z* score is ±1.96.
2.	Substitute the sample values into Formula 6.2
	To Solve Formula 6.2:
1.	Find the square root of *N* − 1.
2.	Divide the value you found in step 1 into ***s***, the sample standard deviation.
3.	Multiply the value you found in step 2 by the value of *Z*.
4.	The value you found in step 3 is half the width of the confidence interval. To find the lower and upper limits of the interval, subtract and add this value to the sample mean.

ONE STEP AT A TIME — Interpreting Confidence Intervals Using Sample Means

Express the confidence interval in a sentence or two that identifies each of these elements:

a. The sample statistic (a mean, in this case)
b. The confidence interval
c. Sample size (*N*)
d. The population to which you are estimating
e. The confidence level (e.g., 95%)

The confidence interval we constructed in this section could be expressed as follows: "The average income for this community is estimated at \$45,000 ± \$17.54. The estimate is based on a sample of 500 respondents, and we can be 95% confident that it is correct."

Applying Statistics 6.1 Estimating a Population Mean

A study of the leisure activities of Americans was conducted on a sample of 1000 households. The respondents identified television viewing as a major form of recreation. If the sample reported an average of 3.0 hours of television viewing a day, what is the estimate of the population mean? The information from the sample is

$$\bar{X} = 3.0$$
$$s = 2.5$$
$$N = 1000$$

If we set alpha at 0.05, the corresponding Z score will be ±1.96, and the 95% confidence interval will be

$$c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{N-1}}\right)$$

$$c.i. = 3.0 \pm (1.96)\left(\frac{2.5}{\sqrt{999}}\right)$$

$$c.i. = 3.0 \pm (1.96)\left(\frac{2.5}{31.61}\right)$$

$$c.i. = 3.0 \pm (1.96)(0.08)$$

$$c.i. = 3.0 \pm 0.16$$

Based on this result, we would estimate that the population spends an average of 3.0 ± 0.16 hours per day viewing television. The lower limit of our interval estimate (3.0 − 0.16) is 2.84, and the upper limit (3.0 + 0.16) is 3.16. Thus, another way to state the interval would be

$$2.84 \le \mu \le 3.16$$

The population mean is greater than or equal to 2.84 and less than or equal to 3.16. Because alpha was set at the 0.05 level, this estimate has a 5% chance of being wrong (that is, of not containing the population mean).

Interval Estimation Procedures Using Sample Proportions (Large Samples)

Estimation procedures using sample proportions are essentially the same as those using sample means, as long as samples are large ($N > 100$). We do not cover estimation procedures for small samples in this text.

Because proportions and means are different statistics, we must use a different sampling distribution. In fact, again based on the Central Limit Theorem we know that sample proportions have sampling distributions that are normal in shape with means (σ_p) equal to the population value (P_u) and standard deviations (σ_p) equal to $\sqrt{P_u(1 - p_u)/N}$. The formula for constructing confidence intervals based on sample proportions is

FORMULA 6.3

$$c.i. = P_s \pm Z\sqrt{\frac{P_u(1 - P_u)}{N}}$$

The values for P_s and N come directly from the sample, and the value of Z is determined by the confidence level, as was the case with sample means. Only one term in the formula (P_u) is undefined—but this is the very quantity we are trying to estimate! How can we set a value for P_u?

We can resolve this dilemma by setting the value of P_u at 0.5. Because the second factor in the numerator under the radical $(1 - P_u)$ is the complement of P_u, the entire numerator $P_u(1 - P_u)$ will always have a value of 0.5×0.5, or 0.25, which is the maximum value this expression can attain. That is, if we set P_u at any value other than 0.5, the expression $P_u(1 - P_u)$ will decrease in value. If we set P_u at 0.4, for example, the second factor $(1 - P_u)$ would be 0.6, and the value of $P_u(1 - P_u)$ would decrease to 0.24. Setting P_u at 0.5 ensures that the expression $P_u(1 - P_u)$ will be at its maximum possible value and, consequently, the interval will be at maximum width. This is the most conservative solution possible to the dilemma posed by having to assign a value to P_u in the estimation equation.

To illustrate these procedures, assume that you wish to estimate the proportion of students at your university who missed at least one day of classes because of illness last semester. Out of a random sample of 200 students, 60 reported that they had been sick enough to miss at least one day of classes. The sample proportion on which we will base our estimate is thus 60/200, or 0.30. At the 95% level, the interval estimate will be

$$c.i. = P_s \pm Z\sqrt{\frac{P_u(1 - P_u)}{N}}$$

$$c.i. = 0.30 \pm (1.96)\sqrt{\frac{(0.5)(0.5)}{200}}$$

$$c.i. = 0.30 \pm (1.96)\sqrt{\frac{0.25}{200}}$$
$$c.i. = 0.30 \pm (1.96)\sqrt{0.00125}$$
$$c.i. = 0.30 \pm (1.96)(0.04)$$
$$c.i. = 0.30 \pm 0.08$$

You would estimate that the proportion of students who missed at least one day of classes because of illness was between 0.22 and 0.38. The estimate could, of course, also be phrased in percentages by reporting that between 22% and 38% of the student body was affected by illness at least once during the past semester. *(For practice with confidence intervals using sample proportions, see problems 6.2, 6.8–6.12, 6.16, 6.17, and 6.19b.)*

ONE STEP AT A TIME Constructing Confidence Intervals Using Sample Proportions

Begin by selecting an alpha level and finding the associated *Z* score in Table 6.3. If you use the alpha level of 0.05, the *Z* score is ±1.96.

Step	**Operation**
1.	Substitute the value 0.25 for the expression $P_u(1 - P_u)$ in the numerator of the fraction under the square root sign.
2.	Divide *N* into 0.25.
3.	Find the square root of the value you found in step 2.
4.	Multiply the value you found in step 3 by the value of *Z*.
5.	The value you found in step 4 is half the width of the confidence interval. To find the lower and upper limits of the interval, subtract and add this value to the sample proportion.

ONE STEP AT A TIME Interpreting Confidence Intervals Using Sample Proportions

Express the confidence interval in a sentence or two that identifies each of these elements:

a. The sample statistic (a proportion, in this case)
b. The confidence interval
c. Sample size (*N*)
d. The population to which you are estimating
e. The confidence level (e.g., 95%)

The confidence interval we constructed in this section could be expressed as follows: "On this campus, 30% ± 8% of the students were sick enough to miss class at least once during the semester. The estimate is based on a sample of 200 respondents, and we can be 95% confident that it is correct."

Applying Statistics 6.2 Estimating Population Proportions

The World Values Survey (http://www.worldvaluessurvey.org/wvs.jsp) interviews random samples from nations around the world about a variety of issues. The table below presents the percentage of respondents from Japan, Zimbabwe, and the United States who said that religion is "very important" in life.

Percentage of Samples Agreeing That Religion Is "Very Important"

Nation	Sample Size (N)	% "Very Important"	95% Confidence Interval ($\alpha = 0.05$)
Japan	2443	5.4%	5.4% ± 2.0%
Zimbabwe	1499	83.5%	83.5% ± 2.5%
United States	2232	40.4%	40.4% ± 2.2%

To construct the confidence intervals, we must convert the percentages to proportions. With an alpha of 0.05, the Z score is ±1.96 and the confidence interval for the United States is

$$c.i. = P_s \pm Z\sqrt{\frac{P_u(1 - P_u)}{N}}$$

$$c.i. = 0.404 \pm (1.96)\sqrt{\frac{(0.50)(0.50)}{2232}}$$

$$c.i. = 0.404 \pm (1.96)\sqrt{\frac{0.25}{2232}}$$

$$c.i. = 0.404 \pm (1.96)\sqrt{0.000112}$$

$$c.i. = 0.404 \pm (1.96)(0.011)$$

$$c.i. = 0.404 \pm 0.022$$

The interval can also be expressed as a percentage: 40.4% ± 2.2%. We can estimate that between 38.2% and 42.6% of the population of the United States believes that religion is "very important." This interval has a 5% chance of not containing the population value. The same formula and procedures are used to find the confidence intervals for Zimbabwe and Japan.

Why would these nations be so different in their religious beliefs? What national traditions and values could be at work?

STATISTICS IN EVERYDAY LIFE

Good Days, Bad Days

What is the happiest day of the year for Americans? According to Gallup's Daily Mood Index, that day in 2013 was Thanksgiving Day, with 70% of Americans saying they felt "happiness and enjoyment without stress and worry." The next highest score was for Memorial Day (65%) and other "happy days" included Christmas Day (64%), Easter Sunday (62%), and Mother's Day (62%).

These results are based on daily samples of about 1000 Americans, use the 95% confidence level, and are accurate to within ±3%.

Source: Sharp, Lindsey. 2014. "Americans' Moods Still Improve on Holidays and Weekends" http://www.gallup.com/poll/167060/americans-moods-improve-holidays-weekends.aspx

TABLE 6.4 Choosing Formulas for Confidence Intervals

If the Sample Statistic Is a	And the Population Standard Deviation (σ) Is	Use Formula
Mean ($\bar{X}$)	Known	6.1 $c.i. = \bar{X} \pm Z\left(\frac{\sigma}{\sqrt{N}}\right)$
Mean ($\bar{X}$)	Unknown	6.2 $c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{N-1}}\right)$
Proportion (P_s)		6.3 $c.i. = P_s \pm Z\sqrt{\frac{P_u(1-P_u)}{N}}$

Computing Confidence Intervals: A Summary

Table 6.4 presents the three formulas for confidence intervals we have covered, organized by the situations in which they are used. To compute a confidence interval for a mean, choose between the two formulas based on whether the population standard deviation (σ) is known or unknown. To compute a confidence interval for a proportion, always use Formula 6.3.

Controlling the Width of Interval Estimates

The width of a confidence interval for either a mean or a proportion can be partly controlled by manipulating two terms in the formula. First, the confidence level can be raised or lowered and, second, the interval can be widened or narrowed by gathering samples of different size.

The researcher alone determines the risk he or she is willing to take of being wrong (that is, of not including the population value in the interval estimate). The exact confidence level (or alpha level) will depend, in part, on the purpose of the research. For example, if potentially harmful drugs were being tested, the researcher would naturally demand very high levels of confidence (99.99% or even 99.999%). On the other hand, if intervals are being constructed only for loose "guesstimates," then much lower confidence levels can be tolerated (such as 90%).

Interval Width and Confidence Level

The relationship between interval size and confidence level is this: *Intervals widen as confidence levels increase*. This should make intuitive sense. Wider intervals are more likely to include the population value; hence, more confidence can be placed in them.

To illustrate this relationship, let us return to the example where we estimated the average income for a community. In this problem, we were working with a sample of 500 residents, and the average income for this sample was $45,000, with a standard deviation of $200. We constructed the 95% confidence interval and found that it extended 17.54 around the sample mean (that is, the interval was $45,000 ± $17.54).

If we had constructed the 90% confidence interval (a lower confidence level) with these sample data, the Z score in the formula would have decreased to ± 1.65, and the interval would have been narrower:

$$c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{N-1}}\right)$$
$$c.i. = 45{,}000 \pm (1.65)\left(\frac{200}{\sqrt{499}}\right)$$
$$c.i. = 45{,}000 \pm (1.65)(8.95)$$
$$c.i. = 45{,}000 \pm 14.77$$

On the other hand, if we had constructed the 99% confidence interval, the Z score would have increased to ± 2.58, and the interval would have been wider:

$$c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{N-1}}\right)$$
$$c.i. = 45{,}000 \pm (2.58)\left(\frac{200}{\sqrt{499}}\right)$$
$$c.i. = 45{,}000 \pm (2.58)(8.95)$$
$$c.i. = 45{,}000 \pm 23.09$$

At the 99.9% confidence level, the Z score would be ± 3.32, and the interval would be wider still:

$$c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{N-1}}\right)$$
$$c.i. = 45{,}000 \pm (3.32)\left(\frac{200}{\sqrt{499}}\right)$$
$$c.i. = 45{,}000 \pm (3.32)(8.95)$$
$$c.i. = 45{,}000 \pm 29.71$$

The four intervals are grouped together in Table 6.5, and we can readily see the increase in interval size as the confidence level increases. Although I have used confidence intervals for means to illustrate, exactly the same relationships apply to confidence intervals for proportions. *(To explore further the relationship between alpha and interval width, see problem 6.13.)*

TABLE 6.5 Confidence Intervals Grow Wider as Confidence Levels Increase ($\bar{X}$ = \$45,000, s = \$200, N = 500 throughout)

Alpha (α)	Confidence Level	Interval	Interval Width
0.10	90%	\$45,000 ± \$14.77	\$29.54
0.05	95%	\$45,000 ± \$17.54	\$35.08
0.01	99%	\$45,000 ± \$23.09	\$46.18
0.001	99.9%	\$45,000 ± \$29.71	\$59.42

TABLE 6.6 Confidence Intervals Grow Narrower as Sample Size Increases ($\bar{X}$ = $45,000, s = $200, α = 0.05 throughout)

N	Confidence Interval	Interval Width
100	$c.i. = 45{,}000 \pm (1.96)\left(\frac{200}{\sqrt{99}}\right) = 45{,}000 \pm 39.40$	$78.80
500	$c.i. = 45{,}000 \pm (1.96)\left(\frac{200}{\sqrt{499}}\right) = 45{,}000 \pm 17.55$	$35.10
1000	$c.i. = 45{,}000 \pm (1.96)\left(\frac{200}{\sqrt{999}}\right) = 45{,}000 \pm 12.40$	$24.80
10,000	$c.i. = 45{,}000 \pm (1.96)\left(\frac{200}{\sqrt{9999}}\right) = 45{,}000 \pm 3.92$	$7.84

Interval Width and Sample Size

Sample size bears the opposite relationship to interval width: *As sample size increases, interval width decreases*. Larger samples give more precise (narrower) estimates.

An example should make this clear. Table 6.6 presents confidence intervals for four samples of different size (N). The sample data are the same as in Table 6.5, and the confidence level is 95% throughout. The calculation of the confidence intervals is illustrated in the middle column, and the width of each interval is listed in the right-hand column. Note the decrease in interval width as sample size increases. The relationships illustrated in Table 6.6 also hold true, of course, for sample proportions.

Notice that the decrease in interval width (or increase in precision) does not bear a constant, or linear, relationship with sample size. For example, the sample with an N of 500 is five times larger than the sample with $N = 100$, but the interval is not five times as narrow. This is an important relationship because it means that N might have to be increased many times over to appreciably improve the precision of an estimate. Because the cost of a research project is a direct function of sample size, this relationship means that there is a point of diminishing returns in estimation procedures. A sample with $N = 10{,}000$ will cost about twice as much as a sample with $N = 5000$, but estimates based on the larger sample will not be twice as precise. *(To explore further the relationship between sample size and interval width, see problem 6.14.)*

Using SPSS to Construct Confidence Intervals

SPSS does not provide a program specifically for constructing confidence intervals, although some of the procedures we'll cover in future chapters do include confidence intervals as part of the output. Rather than make use of these programs, we will get sample statistics from the 2012 GSS and then construct interval estimates to the population (U.S. society as of 2012). We will construct two confidence intervals, first with a sample mean and then with a proportion.

- Click the SPSS icon on your desktop.
- Load the *GSS2012* database.
  - Find the **File** command on the far left of the menu bar and then click **File Open → Data**.
  - Find the *GSS2012* database and click the file name.

To construct a confidence interval with a sample mean

1. Click **Analyze, Descriptive Statistics**, and **Descriptives**.
2. Find *childs* in the list of variables on the left of the "Descriptives" window and click the arrow to move the variable name to the box on the right.
3. Click **OK** and the output will look like this:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
NUMBER OF CHILDREN	1455	0	8	1.84	1.67
Valid *N* (listwise)	1455				

4. We can substitute these values into Formula 6.2 to construct the 95% confidence interval:

$$c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{N-1}}\right)$$

$$c.i. = 1.84 \pm (1.96)\left(\frac{1.67}{\sqrt{1454}}\right)$$

$$c.i. = 1.84 \pm (1.96)(0.04)$$

$$c.i. = 1.84 \pm 0.08$$

Based on a sample of 1455 and at the 95% confidence level, we estimate that Americans have an average of 1.84 ± 0.08 children or between 1.76 and 1.92 children.

To construct a confidence interval with a sample proportion

1. Click **Analyze, Descriptive Statistics**, and **Frequencies**.
2. Find *marital* in the list of variables on the left and click the arrow to move the variable name to the box on the right.

3. Click **OK** and the output will look like this:

MARITAL STATUS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MARRIED	639	43.9	43.9	43.9
	WIDOWED	121	8.3	8.3	52.2
	DIVORCED	251	17.2	17.2	69.4
	SEPARATED	50	3.4	3.4	72.8
	NEVER MARRIED	396	27.2	27.2	100.0
	Total	1457	100.0	100.0	

4. We will estimate the percentage of Americans who have never married. Convert the percentage in the "Valid Percent" column (27.2%) to a proportion (0.27) and substitute into Formula 6.3:

$$c.i. = P_s \pm Z\sqrt{\frac{P_u(1 - P_u)}{N}}$$

$$c.i. = 0.27 \pm (1.96)\sqrt{\frac{(0.50)(0.50)}{1457}}$$

$$c.i. = 0.27 \pm (1.96)\sqrt{\frac{0.25}{1457}}$$

$$c.i. = 0.27 \pm (1.96)(0.01)$$

$$c.i. = 0.27 \pm 0.02$$

Converting back to percentages, we can estimate, at the 95% confidence level, that 27% ± 2% of Americans have never married. This estimate is based on a sample of 1457.

BECOMING A CRITICAL CONSUMER: Public Opinion Polls

Public opinion polls have become a part of everyday life in the United States and in many other societies, and statements such as those below are routinely found on the Internet and in the media:

- *55% of drivers have changed their driving habits as a result of the high cost of fuel.*
- *40% of voters are likely to vote for the Republican candidate for State Senator.*
- *The president's approval rating stands at 48%.*
- *17% of Americans have watched an X-rated movie in the past 6 months.*

How much credibility should be accorded to these statements? Have you ever said (or heard someone else say) "Where do they get the numbers in these surveys? They aren't talking to anyone I know"? Here are some things to keep in mind when considering survey results:

1. *Consider the Source:* Whenever you encounter an attempt to characterize how "the public" feels or thinks, you should examine the source of the statement. Generally, you can place more trust in reports that come from reputable polling firms (e.g., Gallup) or a national news source (CBS News, *USA Today*) and very little (if any) in polls conducted for partisan purposes (that is, surveys by organizations that represent a particular point of view like political parties or advocacy groups).
2. *Consider the Question:* Professional pollsters work hard to keep their questions neutral and avoid

(*continued*)

BECOMING A CRITICAL CONSUMER (*continued*)

evoking a particular response. Still, you should examine the question carefully and critically for biased wording or "trigger" phrases that might cause people to respond in a particular way. For example, assessments of support for the death penalty would almost certainly produce different results if the pollsters asked about "legalized murder by the state" as opposed to "capital punishment."

3. *Consider the Completeness of the Report:* Examine how much information is reported. Professional polling firms use interval estimates, and responsible reporting by the media will usually emphasize the estimate itself (for example, "In a survey of the American public, 47% approved of gay marriage") but also will report the width of the interval ("This estimate is accurate to within ±3%," or "Figures from this poll are subject to a sampling error of ±3%"), the alpha level (usually the confidence level of 95%), and the size of the sample ("1458 households were surveyed"). Your suspicions should be raised if any of this information is missing.
4. *Consider the Sample:* Examine the sample, not only for adequate size but also for representativeness. In particular, you should greatly discount reports based on the folksy, "man in the street" approach used in many news programs and the "comments from our readers or viewers" sometimes found in mass media outlets. These reports may be interesting and even useful but, because they are not based on EPSEM samples, the results *cannot* be generalized or used to characterize the opinions of anyone other than the actual respondents.

STATISTICS IN EVERYDAY LIFE

Elections Projections

Pollsters are very accurate in predicting the outcomes of elections, as illustrated by the 2012 race between Barack Obama and Mitt Romney. The table below presents the predictions of five different polls just before Election Day, along with the actual vote. The predictions differ from each other because pollsters use different sampling techniques and different definitions of "likely voter."

Virtually all predictions were within their respective margins of error (which varied depending mainly on sample size). The one prediction that was inaccurate was for Governor Romney: The Gallup poll was 1% over its margin of error.

Polls and Actual Results, Presidential Race, 2012*

Poll	*N*	Margin of Error	Predicted for Obama	Predicted for Romney
Washington Post/ABC	3205	±2.5%	50%	47%
Pew Research	3815	±1.8%	50%	47%
CNN	1010	±3.0%	49%	49%
Gallup	3117	±2.0%	49%	50%
NBC/Wall St. Journal	1475	±3.0%	48%	47%
ACTUAL VOTE			51%	47%

**Source:* http://www.realclearpolitics.com/epolls/latest_polls/elections/

Remember that the 95% confidence intervals used by professional pollsters are accurate only to within 2–4 percentage points, depending on sample size. This means that the polls cannot identify the likely victor in very close races. The 2012 presidential race was close and virtually all of the confidence intervals overlap. For example, the Washington Post/ABC poll predicted a 3% Obama victory but the poll had a margin of error of ±2.5%. This means that Obama's support could have been as low as 47.5% (50% − 2.5%) and Romney's support could have been as high as 49.5% (47% + 2.5%). When the confidence intervals overlap, the race is said to be "too close to call" and "a statistical dead heat."

SUMMARY

1. Because populations are almost always too large to test, a fundamental strategy of social science research is to select a sample from the defined population and then to use information from the sample to generalize to the population. This is done either by estimation or by hypothesis testing.
2. Simple random samples are created by selecting cases from a list of the population following the rule of EPSEM (each case has an equal probability of being selected). Samples selected by the rule of EPSEM have a very high probability of being representative.
3. The sampling distribution, the central concept in inferential statistics, is a theoretical distribution of all possible sample outcomes. Since its overall shape, mean, and standard deviation are known (under the conditions specified in the two theorems), the sampling distribution can be adequately characterized and utilized by researchers.
4. The two theorems that were introduced in this chapter state that when the variable of interest is normally distributed in the population or when sample size is large, the sampling distribution will be normal in shape, its mean will be equal to the population mean, and its standard deviation (or standard error) will be equal to the population standard deviation divided by the square root of *N*.
5. Population values can be estimated with sample values. With confidence intervals, which can be based on either proportions or means, we estimate that the population value falls within a certain range of values. The width of the interval is a function of the risk we are willing to take of being wrong (the alpha level) and the sample size. The interval widens as our probability of being wrong decreases and as sample size decreases.
6. Estimates based on sample statistics must be unbiased and relatively efficient. Of all the sample statistics, only means and proportions are unbiased. The means of the sampling distributions of these statistics are equal to the respective population values. Efficiency is largely a matter of sample size. The greater the sample size, the lower the value of the standard deviation of the sampling distribution, the more tightly clustered the sample outcomes will be around the mean of the sampling distribution, and the more efficient the estimate.

SUMMARY OF FORMULAS

FORMULA 6.1 Confidence interval for a sample mean (large sample, population standard deviation known):

$$c.i. = \bar{X} \pm Z\left(\frac{\sigma}{\sqrt{N}}\right)$$

FORMULA 6.2 Confidence interval for a sample mean (large sample, population standard deviation unknown):

$$c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{N-1}}\right)$$

FORMULA 6.3 Confidence interval for a sample proportion (large sample):

$$c.i. = P_s \pm Z\sqrt{\frac{P_u(1-P_u)}{N}}$$

GLOSSARY

Alpha (α). The probability of error or the probability that a confidence interval does not contain the population value. Alpha levels are usually set at 0.10, 0.05, 0.01, 0.001, or 0.0001.

Bias. A criterion used to select sample statistics as estimators. A statistic is unbiased if the mean of its sampling distribution is equal to the population value of interest.

Central Limit Theorem. A theorem that specifies the mean, standard deviation, and shape of the sampling distribution, given that the sample is large.

Confidence interval. An estimate of a population value in which a range of values is specified.

Confidence level. A frequently used alternative way of expressing alpha, the probability that an interval estimate will not contain the population value. Confidence levels of 90%, 95%, 99%, 99.9%, and 99.99% correspond to alphas of 0.10, 0.05, 0.01, 0.001, and 0.0001, respectively.

Efficiency. The extent to which the sample outcomes are clustered around the mean of the sampling distribution.

EPSEM. The **E**qual **P**robability of **SE**lection **M**ethod for selecting samples. Every element or case in the population must have an equal probability of selection for the sample.

μ. The mean of a population.

μ_p. The mean of a sampling distribution of sample proportions.

$\mu_{\bar{X}}$ The mean of a sampling distribution of sample means.

Nonprobability sample. Any sample that does not meet the EPSEM criterion. Such samples have a variety of uses in social science research but cannot be used to generalize to a population.

P_s. Any sample proportion.

P_u. Any population proportion.

Parameter. A characteristic of a population.

Representative sample. A sample that reproduces the major characteristics of the population from which it was drawn.

Sampling distribution. The distribution of a statistic for all possible sample outcomes of a certain size. Under conditions specified in two theorems, the sampling distribution will be normal in shape, with a mean equal to the population value and a standard deviation equal to the population standard deviation divided by the square root of N.

Simple random sample. A method for choosing cases from a population by which every case and every combination of cases has an equal chance of being included.

Standard error of the mean. The standard deviation of a sampling distribution of sample means.

PROBLEMS

6.1 For each set of sample outcomes below, use Formula 6.2 to construct the 95% confidence interval for estimating μ, the population mean.

a. $\bar{X} = 5.2$, $s = 0.7$, $N = 157$
b. $\bar{X} = 100$, $s = 9$, $N = 620$
c. $\bar{X} = 20$, $s = 3$, $N = 220$
d. $\bar{X} = 1020$, $s = 50$, $N = 329$
e. $\bar{X} = 7.3$, $s = 1.2$, $N = 105$
f. $\bar{X} = 33$, $s = 6$, $N = 220$

6.2 For each of the following sets of sample outcomes, use Formula 6.3 to construct the 99% confidence interval for estimating P_u.

a. $P_s = 0.40$, $N = 548$
b. $P_s = 0.37$, $N = 522$
c. $P_s = 0.79$, $N = 121$
d. $P_s = 0.14$, $N = 100$
e. $P_s = 0.43$, $N = 1049$
f. $P_s = 0.63$, $N = 300$

6.3 For each of the following confidence levels, determine the corresponding Z score.

Confidence Level	Alpha	Area Beyond Z	Z score
95%	0.05	0.0250	±1.96
94%			
92%			
97%			
98%			

6.4 SW You have developed a series of questions to measure job satisfaction of bus drivers in New York City. A random sample of 100 drivers has an average score of 10.6, with a standard deviation of 2.8. What is your estimate of the average job satisfaction score for the population of all NYC bus drivers? Use the 95% confidence level.

6.5 SOC A researcher has gathered information from a random sample of 178 households. For each of the following variables, construct confidence intervals to estimate the population mean. Use the 90% level.

a. An average of 2.3 people resides in each household. Standard deviation is 0.35.
b. There was an average of 2.1 television sets ($s = 0.10$) and 0.78 telephones ($s = 0.55$) per household.
c. The households averaged 6.0 hours of television viewing per day ($s = 3.0$).

6.6 SOC A random sample of 100 cable TV movies contained an average of 3.38 acts of physical violence per program with a standard deviation of 0.30. At the 99% level, what is your estimate of the population value?

6.7 SOC A random sample of 429 college students was interviewed about a number of matters. Use the results to construct confidence interval estimates of the population mean at the 99% level.

a. They reported that they had spent an average of $478.23 on textbooks during the previous semester, with a sample standard deviation of \$15.78.

b. They also reported that they had visited the health clinic an average of 1.5 times a semester, with a sample standard deviation of 0.3.

c. On the average, the sample had missed 2.8 days of classes per semester because of illness, with a sample standard deviation of 1.0.

d. On the average, the sample had missed 3.5 days of classes per semester for reasons other than illness, with a sample standard deviation of 1.5.

6.8 **PA** A random sample of 100 inmates at a maximum security prison shows that exactly 10 of the respondents had been the victims of violent crime during their incarceration. Estimate the proportion of victims for the population as a whole, using the 90% confidence level. *(HINT: Calculate the sample proportion P_s before using Formula 6.3. Remember that a proportion is equal to frequency divided by N.)*

6.9 **SOC** The survey mentioned in problem 6.5 found that 25 of the 178 households consisted of unmarried couples who were living together. What is your estimate of the population proportion? Use the 95% level.

6.10 **PA** A random sample of 260 workers in a high-rise office building revealed that 30% were very satisfied with the quality of elevator service. At the 99% level, what is your estimate of the population value?

6.11 A random sample of 1496 respondents of a major metropolitan area was questioned about a number of issues. Construct estimates to the population at the 95% level for each of the results reported next. Express the final confidence interval in percentages (e.g., "between 40% and 45% agreed that premarital sex was always wrong").

a. When asked to agree or disagree with the statement "Internet pornography leads to rape and other sex crimes," 823 agreed.

b. When asked to agree or disagree with the statement "Hand guns should be outlawed," 650 agreed.

c. 375 of the sample agreed that marijuana should be legalized.

d. 1023 of the sample said that they had attended a church, synagogue, mosque, or other place of worship at least once within the past month.

e. 800 agreed that public elementary schools should have sex education programs starting in the fifth grade.

6.12 **SW** A random sample of 100 patients treated in a program for alcoholism and drug dependency over the past 10 years was selected. It was determined that 53 of the patients had been readmitted to the program at least once. At the 95% level, construct an estimate of the population proportion.

6.13 For the following sample data, construct five different interval estimates of the population mean, one each for the 90%, 95%, 99%, 99.9%, and 99.99% levels. What happens to the interval width as confidence level increases? Why?

$$\bar{X} = 100$$
$$s = 10$$
$$N = 500$$

6.14 For each of the following three sample sizes, construct the 95% confidence interval for the population proportion. Use a sample proportion of 0.40 throughout. What happens to interval width as sample size increases? Why?

$$P_s = 0.40$$

Sample A: $N = 100$
Sample B: $N = 1000$
Sample C: $N = 10{,}000$

6.15 **PS** Two individuals are running for mayor of your town. You conduct an election survey a week before the election and find that 51% of the respondents prefer candidate A. Can you predict a winner? Use the 99% level. *(HINT: In a two-candidate race, what percentage of the vote would the winner need? Does the confidence interval indicate that candidate A has a sure margin of victory? Remember that while the population parameter is probably [α = 0.01] in the confidence interval, it may be anywhere in the interval.)*

$$P_s = 0.51$$
$$N = 578$$

6.16 **SOC** The World Values Survey is administered periodically to random samples from societies around the globe. Listed here are the percentage of respondents in each nation who said that they are "very happy." Construct confidence interval estimates for each nation at the 95% level.

Nation	Percent "Very Happy"	Sample Size	Confidence Interval
Germany	21.1%	2046	
Japan	32.3%	2443	
Chili	24.4%	1000	
China	15.7%	2300	
Malaysia	56.5%	1300	
Nigeria	55.6%	1759	
Russia	14.8%	2200	

6.17 SOC The fraternities and sororities at St. Algebra College have been plagued by declining membership over the past several years and want to know whether the incoming freshman class will be a fertile recruiting ground. Not having enough money to survey all 1600 freshmen, they commission you to survey the interests of a random sample. You find that 35 of your 150 respondents are "extremely" interested in social clubs. At the 95% level, what is your estimate of the number of freshmen who would be extremely interested? *(HINT: The high and low values of your final confidence interval are proportions. How can proportions be expressed as numbers?)*

6.18 SOC You are the automotive specialist for the monthly magazine *Consumer Beware*. Part of your job is to investigate the claims made by automakers in their TV commercials. You are particularly suspicious of a new economy car that the automaker claims will get 78 miles per gallon. After checking the mileage figures for a random sample of 125 owners of this car, you find an average miles per gallon of 75.5, with a standard deviation of 3.7. At the 99% level, do your results tend to confirm or refute the manufacturer's claims?

Statistical Analysis Using SPSS

6.19 In this exercise, you will use SPSS to generate sample statistics from the 2012 General Social Survey (*GSS2012*). You will then construct confidence intervals to estimate population values or parameters.

- Click the SPSS icon on your desktop.
- Load the *GSS2012* database.
- Use the **Descriptives** program to find means and standard deviations for hours worked (*hrs1*) and age (*age*).

a. Use Formula 6.2 to estimate population values for each variable. Use the 95% confidence level and express the confidence interval in a few sentences.

- Use the **Frequencies** program to get frequency distributions for religious affiliation (*relig*) and support for the death penalty (*cappun*).

b. Use Formula 6.3 to estimate population values for the percentage Catholic and the percentage that support the death penalty. Use the 95% confidence level and express the confidence intervals in a few sentences.

6.20 This exercise will give you a tangible understanding of sampling distributions and the two theorems presented in this chapter. *For this demonstration only*, the 2012 GSS database will be treated as a *population* and its characteristics will be treated as *parameters*. We will use SPSS to draw ten random samples and calculate statistics for *age* on each sample. Recall that a sampling distribution includes *all possible* samples, so these ten samples are the merest beginning of an actual sampling distribution.

After drawing each sample, you will record the mean and the standard error of the mean for *age*. Recall that the standard error is the standard deviation of the sampling distribution ($\sigma_{\bar{X}}$). The means of our ten samples should be within a standard error of the population mean, which in this artificial situation we know is 48.21 ($\mu = 48.21$).

- Find and click the SPSS icon on your desktop.
- Load the *GSS2012* data set.
- Click **Data** from the menu bar of the Data Editor and then click **Select Cases**. Next, click the button next to **Random sample of cases** and then click the **Sample** button. On the "Select Cases: Random Sample" window, we can specify the size of the sample in two different ways. We'll use the second option and select samples of size 200 ($N = 200$). Click the button next to **Exactly** and fill in the boxes so that this line reads "Exactly 200 cases from the first 1457 cases."
- Click **Continue**, and then click **OK** on the "Select Cases" window. The sample will be selected.
- To find the mean *age* for the sample, click **Analyze**, **Descriptive Statistics**, and **Descriptives**. Find *age* in the variable list and transfer it to the "Variable(s)": window.
- On the "Descriptives" menu, click the **Options** button and select **S.E. MEAN** in addition to the usual statistics. Click **Continue** and then **OK,** and the requested statistics will appear in the output window.
- Write your results in the following table. I have entered the results from our first sample.
- Before getting the next sample, we need to restore the full data set ($N = 1457$). To do this, click **Data**, **Select Cases**, and then click the **Reset** button at the bottom of the window. Click **OK** and the GSS data set will be restored to its original size.
- To produce the second sample, click **Data**, **Select Cases**, and then click the button next to **Random Sample of Cases**. Click **Sample** and, as before, specify "Exactly 200 cases from the first 1457 cases."

- Click **Continue** and **OK,** and the second sample will be selected. Use **Descriptives** to calculate statistics and record your results in the following table.
- Repeat these steps for the third sample and record your results. Make sure that you restore the full data set (click **Data**, **Select Cases**, and **Reset**) each time before selecting the next sample. Continue until you have results for ten samples.

To complete the table

- Enter the sample means and standard errors (S.E.) of the mean in the appropriate columns for each sample.
- The right-hand column shows an interval that is the sample mean plus and minus the standard error. Find the lower limit of the interval by subtracting the standard error from the sample mean and the upper limit by adding the standard error to the sample mean. The population mean (μ = 48.21) will almost always be in this interval.

	$\bar{X}$	S.E.	$\bar{X} \pm$ S.E.
1.	47.81	1.29	46.52 to 49.10
2.			_____ to _____
3.			_____ to _____
4.			_____ to _____
5.			_____ to _____
6.			_____ to _____
7.			_____ to _____
8.			_____ to _____
9.			_____ to _____
10.			_____ to _____

What can we conclude? First, if you had the patience to draw all ten samples, congratulations! Second, I must point out that this exercise would work better if you drew even more samples. Remember that the theorems about the characteristics of the sampling distribution are based on an *infinite* number of sample outcomes.

Third, the ten outcomes in the table *are* a sampling distribution, although not a very developed one. The ten outcomes in the table are just the beginning of a very long process if we were to construct an actual sampling distribution.

Remember that the sampling distribution is normal and has a mean equal to the population mean and a standard deviation equal to the standard error (S.E.). About 68% of all sample outcomes will be within ±1 standard error of the mean, 95% will be within ±2 standard errors, and so forth. What percentage of your ten sample means are in these ranges? Random chance is at work here and some of your sample means might be very different from the population mean, but most should be reasonably close (i.e., within ±1 S.E.).

We would not ordinarily know the value of the population mean, but we do know that it is the same value as the mean of the sampling distribution. Thus, the odds are in our favor that any particular sample mean will be close to the parameter. This exercise should convince you that properly drawn EPSEM samples are very likely to be representative of populations, and statistics computed on these samples should be close in value to their respective parameters.

YOU ARE THE RESEARCHER

Estimating the Characteristics of the Typical American

As we did in the "Using SPSS" demonstration, we will use SPSS to produce sample statistics from the 2012 GSS, which you can then use as the basis for interval estimates to the population.

Step 1: Choosing the Variables

As a framework for this exercise, we will return to the task of describing the "typical American" begun in Chapter 3 and continued in Chapter 4. Select four of the variables you used in the earlier exercises and add four more that you did NOT use in the earlier chapters, in the "Using SPSS" demonstration in this chapter, or in Problem 6.19. Make sure that you have at least one variable from the nominal level and at least one measured at the ordinal level with at least four categories or one at the interval-ratio level. List the variables, along with level of measurement, here.

	SPSS Variable Name	Explain Exactly What This Variable Measures	Level of Measurement
1			
2			
3			
4			
5			
6			
7			
8			

Step 2: Getting Sample Statistics

For interval-ratio variables and ordinal-level variables with four or more scores, use **Descriptives** to get sample means, standard deviations, and sample sizes. For nominal-level variables and for ordinal-level variables with three or fewer scores, use **Frequencies** to produce frequency distributions. For either procedure, click **Analyze Descriptive Statistics** and then select either **Frequencies** or **Descriptives**. Select your variables from the box on the left and click the arrow to move the variable name to the window on the right. SPSS will process all variables listed at the same time.

Step 3: Constructing Confidence Intervals

Once you have the SPSS output, use the results to construct 95% confidence intervals around the sample statistics. For nominal-level variables and ordinal-level variables with three or fewer scores, select one category (e.g., female for *sex* or Catholic for *relig*) and look in the "Valid Percent" column of the frequency distribution to get the percentage of cases in the sample in that category. Change this value to a proportion (divide by 100). This value is P_s and can be substituted directly into Formula 6.3. Remember to estimate P_u at 0.5. After you have found the confidence interval, change your proportion back to a percentage and record your results in the following table.

For ordinal-level variables with four or more scores and for interval-ratio variables, find the sample mean, standard deviation, and sample size in the output of the **Descriptives** procedure. Substitute these values into Formula 6.2 and record your results in the following table.

Step 4: Recording Results

Use the table to summarize your confidence intervals.

	SPSS Variable Name	Sample Statistic ($\bar{X}$ or P_s)	N	95% Confidence Interval
1				
2				
3				
4				
5				
6				
7				
8				

Note: Some ordinal-level variables, like *news* and *attend*, use scores to stand for broad categories (see Appendix G for examples), and sample means should be interpreted in terms of these categories, not as absolute scores. For example, if the mean of *news* is 3.7, this would indicate that the average score for this variable is somewhere between "once a week" (a score of 3) and "less than once a week" (a score of 4).

Step 5: Reporting Results

For each statistic, express the confidence interval in words, as if you were reporting results in a newspaper story. Be sure to identify the population and include each of the following: the sample statistic, sample size, the upper and lower limits of the confidence interval, and the confidence level (95%). A sample sentence might read, "I estimate that 45% ± 3% of all Americans support candidate X for president. This estimate is based on a sample of 1362 adult Americans, and I am 95% confident that it is correct."

7 Hypothesis Testing I
The One-Sample Case

LEARNING OBJECTIVES

By the end of this chapter, you will be able to:

1. Explain the logic of hypothesis testing, including the concepts of the null hypothesis, the sampling distribution, the alpha level, and the test statistic.
2. Explain what it means to "reject the null hypothesis" or "fail to reject the null hypothesis."
3. Identify and cite examples of situations in which one-sample tests of hypotheses are appropriate.
4. Test the significance of single-sample means and proportions using the five-step model, and correctly interpret the results.
5. Explain the difference between one- and two-tailed tests, and specify when each is appropriate.
6. Define and explain Type I and Type II errors, and relate each to the selection of an alpha level.
7. Use the Student's *t* distribution to test the significance of a sample mean for a small sample.

USING STATISTICS

The statistical techniques presented in this chapter are used to compare a sample to a population. Examples include:

1. A researcher working for a senior citizens advocacy group has asked a sample of 789 older residents of a particular state whether they have been victimized by crime over the past year. He wants to test his hypothesis that senior citizens, as represented by this sample, are more likely to be victimized than the population in general.
2. The board of trustees of a small college wants to expand its athletic program but is concerned about the academic progress of student athletes. They have asked the dean to investigate. Among other tests, he will compare the mean GPA of a random sample of 235 student athletes with the mean GPA of all students at the university.
3. A sociologist is assessing the effectiveness of a rehabilitation program for alcoholics in her city. The program is large and she cannot test all clients. Instead, she draws a random sample ($N = 127$) from the list of all clients

and questions them on a variety of issues. She notices that, on the average, her sample misses fewer days of work than workers in the city as a whole; she then conducts a test to see whether the workers in her sample really are more reliable than workers in the community as a whole.

In Chapter 6, you were introduced to inferential statistical techniques for estimating population parameters from sample statistics. In Chapters 7 through 10, we investigate a second application of inferential statistics, called **hypothesis testing** or **significance testing**. In this chapter, the techniques for hypothesis testing in the one-sample case are introduced. We use these techniques to compare a random sample to a population. Note that we are not interested in the sample per se but in the larger group from which it was selected: We want to know whether the group represented by the sample differs from population parameters on a specific variable.

Of course, it would be better if we could test all members of the group in which we are interested rather than these smaller samples. However, as we have seen, researchers usually do not have the resources to test large groups and must use random samples instead. In these situations, conclusions will be based on a comparison of a single sample (representing the larger group) and the population.

For example, if we found that the rate of victimization for a sample of senior citizens was higher than the rate for the state as a whole, we might conclude that "senior citizens are significantly more likely to be crime victims." The word *significantly* is a key word: It means that the difference between the sample's victimization rate and the population's rate is unlikely to be caused by random chance alone. In other words, it is likely that *all* senior citizens (not just those in the sample) have a higher victimization rate than the state as a whole. On the other hand, if we found little difference between the GPAs of a sample of athletes and the whole student body, we might conclude that athletes (*all* athletes on this campus, not just the athletes in the sample) are essentially the same as other students in terms of academic achievement.

Thus, we can use samples to represent larger groups (senior citizens or athletes) and compare and contrast the characteristics of the sample with those of the population and be extremely confident in our conclusions. Remember, however, that there is no guarantee that random samples will be representative, and there will always be a small amount of uncertainty in our conclusions. One of the great advantages of inferential statistics is that we will be able to estimate the probability of error and evaluate our decisions accordingly.

An Overview of Hypothesis Testing

We'll begin with a general overview of hypothesis testing, using the third situation mentioned in "Using Statistics" as an example. We will introduce the more technical considerations and proper terminology throughout the remainder of the chapter.

Let's examine this situation, which involved a sample of treated alcoholics, in some detail. First, the main question here is, "Are people treated in this program more reliable workers than people in the community in general?" In other words, what the researcher would really like to do is compare *all* clients (the *population* of alcoholics treated in the program) with the entire metropolitan area. If she had information for these groups, she could answer the question easily, completely, and finally.

The problem is that the researcher does not have the time or money to gather information on the thousands of people who have been treated by the program. Instead, she has drawn a random sample of 127 clients. The absentee rates for the sample and the community are

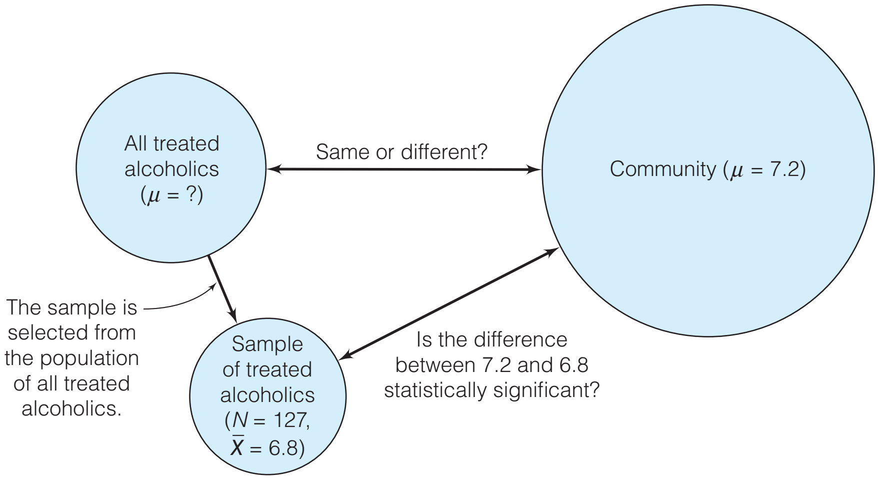
Community	Sample of Treated Alcoholics
$\mu = 7.2$ days per year	$\bar{X} = 6.8$ days per year
$\sigma = 1.43$	$N = 127$

We can see that the average rate of absenteeism for the sample is lower than the rate for the community. Although it's tempting, we can't make any conclusions yet because we are working with a random sample ($N = 127$) of the population we are interested in, not the population itself (*all* people treated in the program).

Figure 7.1 should clarify these relationships. The community is symbolized by the largest circle because it is the largest group. The population of all treated alcoholics is also symbolized by a large circle because it is a sizeable group, although only a fraction of the community as a whole. The random sample of 127, the smallest of the three groups, is symbolized by the smallest circle.

The labels on the arrows connecting the circles summarize the major questions and connections in this research situation. As we noted earlier, the main question is "Does the population of *all* treated alcoholics have different absentee rates than the community as a whole?" The population of treated alcoholics, too large to test, is represented by the random sample of 127.

FIGURE 7.1 A Test of Hypothesis for Single-Sample Means



We are interested in what caused the observed difference between the sample mean of 6.8 and the community mean of 7.2. There are two possible explanations for this difference:

1. The first explanation is that the difference between the mean for the entire community (7.2) and the sample mean (6.8) is "statistically significant," which means that it is very unlikely to have occurred by random chance alone. If this explanation is true, the population of *all* treated alcoholics is different from the community and the sample did *not* come from a population with a mean absentee rate of 7.2 days.
2. The second explanation is called the **null hypothesis** (symbolized as H_0, or H-sub-zero). It states that the observed difference between the sample and community means was caused by mere random chance: There is no important difference between treated alcoholics and the community, and the difference between the sample mean and the community mean is due to random chance. If the null hypothesis is true, treated alcoholics are just like everyone else and have a mean absentee rate of 7.2 days.

Which explanation is correct? We cannot answer this question with absolute certainty as long as we are working with a sample rather than the entire group. We can, however, set up a decision-making procedure so conservative that one of the explanations can be chosen, with the knowledge that the probability of choosing the wrong explanation is very low.

This decision-making process begins with the assumption that the second explanation—the null hypothesis—is correct. Symbolically, the assumption that the mean absentee rate for all treated alcoholics is the same as the rate for the community as a whole can be stated as

$$H_0: \mu = 7.2 \text{ days per year}$$

Remember that μ refers to the mean for *all* treated alcoholics, not just the 127 in the sample. This assumption, $\mu = 7.2$, can be tested statistically.

If the null hypothesis ("*The population of treated alcoholics is not different from the community as a whole and has a* μ *of 7.2*") is true, then we can find the probability of getting the observed sample outcome ($\bar{X} = 6.8$). Let us add an objective decision rule in advance. If the odds of getting the observed difference are less than 0.05 (5 out of 100, or 1 in 20), we will reject the null hypothesis. If this explanation were true, a difference of this size (7.2 days vs. 6.8 days) would be a very rare event, and in hypothesis testing we always bet *against* rare events.

We can estimate the probability of getting the observed sample outcome ($\bar{X} = 6.8$) if the null hypothesis is correct by using our knowledge of the sampling distribution of all possible sample outcomes. Looking back at the information we have and applying the Central Limit Theorem (see Chapter 6), we can assume that the sampling distribution is normal in shape and has a mean of 7.2 (because $\mu_{\bar{X}} = \mu$) and a standard deviation of $1.43/\sqrt{127}$ because $\sigma_{\bar{X}} = \sigma/\sqrt{N}$. We also know that the standard normal distribution can be interpreted as a distribution of probabilities (see Chapter 5) and that this sample outcome ($\bar{X} = 6.8$) is one of thousands of possible sample outcomes. The sampling distribution, with the sample outcome noted, is depicted in Figure 7.2.

FIGURE 7.2 The Sampling Distribution of All Possible Sample Means

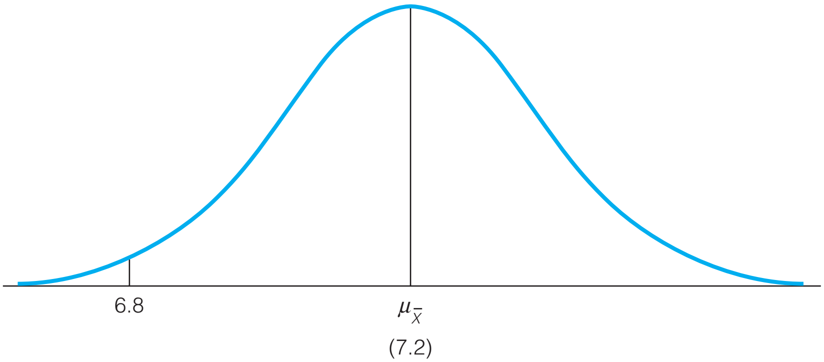
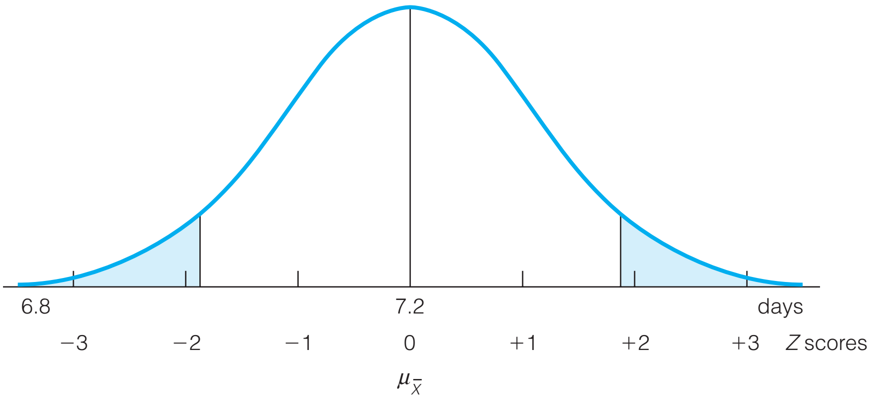


FIGURE 7.3 The Sampling Distribution of All Possible Sample Means



We can add further useful information to this sampling distribution of sample means. Specifically, we can depict our decision rule ("Any sample outcome with probability less than 0.05 will cause us to reject the null hypothesis") with *Z* scores. The probability of 0.05 can be translated into an area and divided equally into the upper and lower tails of the sampling distribution. Using Appendix A, we find that the *Z*-score equivalent of this area is ± 1.96. (To review finding *Z* scores from areas or probabilities, see Chapter 6). The areas and *Z* scores are depicted in Figure 7.3.

The decision rule can now be rephrased. Any sample outcome falling in the shaded areas depicted in Figure 7.3 has a probability of less than 0.05. Such an outcome would be rare and would cause us to reject the null hypothesis.

All that remains is to translate our sample outcome into a *Z* score to see where it falls on the curve. To do this, we use the standard formula for locating any score under a normal distribution. When we use known or empirical distributions, this formula is expressed as

$$Z = \frac{X_i - \bar{X}}{s}$$

Or, to find the equivalent *Z* score for any raw score, subtract the mean of the distribution from the raw score and divide by the standard deviation of the distribution.

Because we are now concerned with the sampling distribution of all sample means rather than an empirical distribution, the symbols in the formula will change, but the form remains exactly the same:

FORMULA 7.1

$$Z(\text{obtained}) = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

Or, to find the equivalent Z score for any sample mean, subtract the mean of the sampling distribution (which is equal to the population mean, or μ) from the sample mean and divide by the standard deviation of the sampling distribution.

Recalling the data given on this problem, we can now find the Z-score equivalent of the sample mean:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

$$Z = \frac{6.8 - 7.2}{1.43/\sqrt{127}}$$

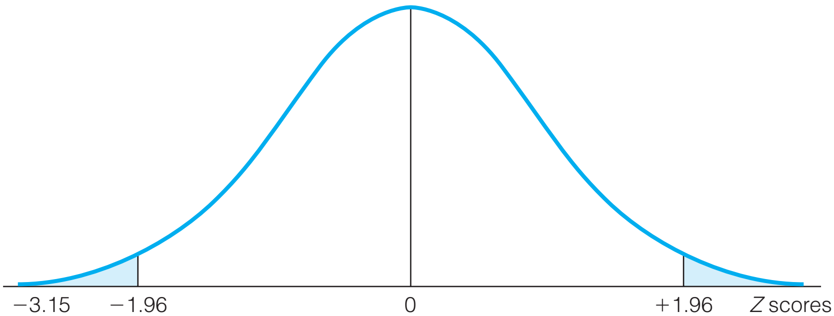
$$Z = \frac{-0.40}{1.43/11.27}$$

$$Z = \frac{-0.40}{0.127}$$

$$Z = -3.15$$

In Figure 7.4, this Z score of -3.15 is noted on the distribution of all possible sample means, and we see that the sample outcome does fall in the shaded area. If the null hypothesis is true, this sample outcome has a probability of less than 0.05. The sample outcome ($\bar{X} = 6.8$, or $Z = -3.15$) would be rare if the null hypothesis were true, and the researcher may therefore reject the null hypothesis. The sample of 127 treated alcoholics comes from a population that is significantly different from the community on absenteeism. Or, to put it another way, the sample does not come from a population that has a mean of 7.2 days of absences.

FIGURE 7.4 The Sampling Distribution of Sample Means with the Sample Outcome ($\bar{X}$ = 6.8) Noted in Z Scores



STATISTICS IN EVERYDAY LIFE

Testing Drugs
The U.S. Food and Drug Administration (FDA) estimates that it takes up to 10 years for a new drug to be approved for use in the United States. Before the manufacturer of a drug can apply for FDA approval, the drug must be tested first on animals, then on samples of healthy volunteer humans, and finally on samples of humans who have been diagnosed with the specific disease or condition that the manufacturer claims the drug was designed to treat. The results of these studies must show that the drug is safe and effective and that its benefits outweigh its risks. All the tests are based on null hypotheses that the drug is ineffective in the treatment of the target disease or condition.

Keep in mind that our decisions in significance testing are based on random samples. On rare occasions, an EPSEM sample may not be representative of the population from which it was selected. The decision-making process just outlined has a very high probability of resulting in correct decisions, but we always face an element of risk when we must work with samples rather than populations. The decision to reject the null hypothesis might be incorrect if this sample happens to be one of the few that is unrepresentative of the population of alcoholics treated in this program.

One important strength of hypothesis testing is that we can estimate the probability of making an incorrect decision. In the example at hand, the null hypothesis was rejected and the probability that this decision was incorrect is 0.05—the decision rule established at the beginning of the process. To say that the probability of rejecting the null hypothesis incorrectly is 0.05 means that, if we repeated this same test an infinite number of times, we would incorrectly reject the null hypothesis only 5 times out of every 100.

The Five-Step Model for Hypothesis Testing

The previous section introduced the concepts and procedures used in hypothesis testing. Next, we will examine a **five-step model** for organizing all tests of hypothesis.

Step 1. Making assumptions and meeting test requirements

Step 2. Stating the null hypothesis

Step 3. Selecting the sampling distribution and establishing the critical region

Step 4. Computing the test statistic

Step 5. Making a decision and interpreting the results of the test

We examine each step individually, using the problem from the previous section as an example.

Step 1. Making Assumptions and Meeting Test Requirements. Any application of statistics is based on a set of assumptions. For hypothesis tests with a sample mean, three assumptions have to be satisfied:

- The sample has been selected according to the rules of EPSEM.
- The variable being tested is interval-ratio in level of measurement.
- The sampling distribution of all possible sample means is normal in shape. This will allow us to use the standardized normal distribution to find areas under the sampling distribution. We can be sure that this assumption is satisfied by using large samples (see the Central Limit Theorem in Chapter 6).

Usually, we will state these assumptions in abbreviated form as a mathematical model for the test. For example,

Model: Random sampling
Level of measurement is interval-ratio
Sampling distribution is normal

Step 2. Stating the Null Hypothesis. The null hypothesis is always a statement of "no difference," but its exact form will vary depending on the test being conducted. In the single-sample case, the null hypothesis states that the sample comes from a population with a certain characteristic. In our specific example, the null hypothesis is that the population of treated alcoholics is "no different" from the community as a whole, that their average days of absenteeism is also 7.2, and that the difference between 7.2 and the sample mean of 6.8 is caused by random chance. As we saw previously, the null hypothesis is stated as

$$H_0: \mu = 7.2$$

where μ refers to the mean of the population of treated alcoholics. The null hypothesis is the central element in any test of hypothesis because the entire process is aimed at rejecting or failing to reject H_0.

Usually, the researcher believes there is a difference between the sample and the population and wants to reject the null hypothesis. The researcher's belief is stated in a **research hypothesis (H_1)**, a statement that directly contradicts the null hypothesis. Very often, the researcher's goal is to support the research hypothesis by rejecting the null hypothesis.

The research hypothesis can be stated in several ways. One form would state that the population from which the sample was selected did *not* have a certain characteristic or, in terms of our example, had a mean that was *not* equal to a specific value:

$$(H_1: \mu \neq 7.2)$$

where $\neq$ means "not equal to"

The research hypothesis asserts that the population of all treated alcoholics is *different from* the community as a whole. The null hypothesis, in contrast, asserts that the population of all treated alcoholics is *the same as* the community as a whole. One (and only one) of these two statements must be true.

The research hypothesis is enclosed in parentheses to emphasize that it has no formal standing in the hypothesis-testing process (except, as we shall see in the next section, in choosing between one-tailed and two-tailed tests). It serves as a reminder of what the researcher believes to be the truth.

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region. The sampling distribution is the probabilistic yardstick against which a particular sample outcome is measured. By assuming that the null hypothesis is true (and *only* by this assumption), we can attach values to the mean and standard deviation of the sampling distribution and measure the probability of any specific sample outcome. We will cover several different sampling distributions in this text, but for now we will confine our attention to the sampling distribution described by the standard normal curve, as summarized in Appendix A.

The **critical region**, or **region of rejection**, consists of the areas under the sampling distribution that include unlikely sample outcomes. In our earlier example, this area began at a Z score of ± 1.96, called **Z(critical)**, that was graphically displayed in Figure 7.3. The shaded area is the critical region. Any sample outcome for which the Z-score equivalent fell in this area (that is, below -1.96 or above $+1.96$) would have caused us to reject the null hypothesis.

By convention, the size of the critical region is reported as alpha (α), the proportion of all of the area included in the critical region. In our example, our **alpha level** was 0.05. Other commonly used alphas are 0.10, 0.01, 0.001, and 0.0001.

These decisions can be stated in abbreviated form. The critical region is noted by the Z scores that mark its beginnings.

$$\text{Sampling distribution} = Z \text{ distribution}$$

$$\alpha = 0.05$$

$$Z(\text{critical}) = \pm 1.96$$

(For practice in finding Z(critical) scores, see problem 7.1a.)

Step 4. Computing the Test Statistic. To evaluate the probability of the sample outcome, we must convert the sample value into a Z score. Finding this Z score is called computing the **test statistic**, and the resultant value will be referred to as **Z(obtained)** to differentiate it from the score that marks the beginning of the critical region. In our example, we found a Z(obtained) of -3.15. *(For practice in computing obtained Z scores for means, see problems 7.1c, 7.2 to 7.7, 7.15e and f, and 7.19d to f.)*

ONE STEP AT A TIME | **Completing Step 4 of the Five-Step Model: Computing *Z*(obtained)**

Use these procedures if the population standard deviation (σ) is known or sample size (N) is greater than 100. See the section entitled "The Student's t Distribution" for procedures when σ is unknown and N is less than 100.

Step	***Operation***
	To compute the test statistic using Formula 7.1:
1.	Find the square root of *N*.
2.	Divide the square root of *N* into the population standard deviation (σ).
3.	Subtract the population mean (μ) from the sample mean ($\bar{X}$).
4.	Divide the quantity you found in step 3 by the quantity you found in step 2. This value is *Z*(obtained).

TABLE 7.1 Making a Decision in Step 5 and Interpreting the Results of the Test

Situation	Decision	Interpretation
The test statistic is in the critical region.	Reject the null hypothesis (H_0).	The difference is statistically significant.
The test statistic is not in the critical region.	Fail to reject the null hypothesis (H_0).	The difference is not statistically significant.

Step 5. Making a Decision and Interpreting the Results of the Test. If the test statistic falls in the critical region, our decision will be to reject the null hypothesis. If the test statistic does not fall in the critical region, we fail to reject the null hypothesis. In our example, the two values were

$$Z(\text{critical}) = \pm 1.96$$

$$Z(\text{obtained}) = -3.15$$

and we saw that the Z(obtained) fell in the critical region (see Figure 7.4). Our decision was to reject the null hypothesis, which stated: "Treated alcoholics have a mean absentee rate of 7.2 days." When we reject this null hypothesis, we are saying that treated alcoholics do *not* have a mean absentee rate of 7.2 days and that there *is* a difference between them and the community. We can also say that the difference between the sample mean of 6.8 and the community mean of 7.2 is statistically significant, or unlikely to be caused by random chance alone.

Note that, in order to complete step 5, you have to do two things. First, you reject or fail to reject the null hypothesis (see Table 7.1). Second, you need to say what this decision means. In this case, we rejected the null hypothesis and concluded that there is a significant difference between the mean of the sample and the mean for the entire community. Therefore, we concluded that treated alcoholics are different from the community as a whole.

This five-step model will serve as a framework for decision making throughout the hypothesis-testing chapters. The exact nature and method of expression for our decisions will be different for different situations. However, the five-step model will provide a common frame of reference for all significance testing.

ONE STEP AT A TIME **Completing Step 5 of the Five-Step Model: Making a Decision and Interpreting Results**

Step	*Operation*
1.	Compare the Z(obtained) to the Z(critical). If Z(obtained) is *in* the critical region, *reject* the null hypothesis. If Z(obtained) is *not in* the critical region, *fail to reject* the null hypothesis.
2.	Interpret your decision in the terms of the original question. For example, our conclusion for the example problem was "Treated alcoholics are different from the community as a whole."

STATISTICS
IN EVERYDAY LIFE

Hypothesis Testing and Gambling

Think of hypothesis testing as analogous to gambling. Imagine that you have been invited to participate in a game of chance involving coin flips: heads you win, tails you lose. You would agree to participate only if you could **assume** that the coin was honest and that the probabilities of heads and tails were both 0.5.

What if your opponent flips 10 tails in a row? Think of these flips as a **test** of your original assumption. At some point, as the coin shows tail after tail, you must compare the outcomes with your original assumption of honesty and make a **decision**: either the coin is weighted toward tails or you have just witnessed a very rare series of events.

If you decide that the game is rigged, you have rejected the null hypothesis that there is no difference in the probabilities of heads and tails. This seems like a very reasonable decision, but note that there is a chance that it is wrong: it is possible (though unlikely) that the game is not rigged and the coin is not weighted toward tails.

In hypothesis testing, we also **make assumptions** (steps 1 through 3), **test** these assumptions in step 4, and make a **decision**, based on probabilities, in step 5.

Choosing a One-Tailed or Two-Tailed Test

The five-step model for hypothesis testing allows for few choices. Nonetheless, the researcher must still deal with two options. First, he or she must decide between a one-tailed and a two-tailed test. Second, an alpha level must be selected. In this section, we discuss the former decision; we discuss the latter in the next section.

Types of Significance Tests

The choice between a one- and two-tailed test is based on the researcher's expectations about the population from which the sample was selected. These expectations are reflected in the research hypothesis (H_1), which is contradictory to the null hypothesis and usually states what the researcher believes to be "the truth." In most situations, the researcher will wish to support the research hypothesis by rejecting the null hypothesis.

The format for the research hypothesis may take either of two forms, depending on the relationship between what the null hypothesis states and what the researcher believes to be the truth. In the one-sample tests considered in this chapter, the null hypothesis states that the population has a specific characteristic. In the example that we have used in this chapter, the null hypothesis stated, in symbols, "All treated alcoholics have the *same* absentee rate (7.2 days) as the community." The researcher might believe that the population of treated alcoholics actually has *less* absenteeism (their population mean is *lower than* the value stated in the null hypothesis) or *more* absenteeism (their population mean is *greater than* the value stated in the null hypothesis), or he or she might be unsure about the direction of the difference.

If the researcher is unsure about the direction, the research hypothesis would state only that the population mean is "not equal" to the value stated in the null hypothesis. The research hypothesis stated in our example ($\mu \neq 7.2$) was in this format. This is called a **two-tailed test** of significance because the researcher will be equally concerned with the possibility that the true population value is greater than *or* less than the value specified in the null hypothesis.

In other situations, the researcher might be concerned only with differences in a specific direction. If the direction of the difference can be predicted, or if the researcher is concerned only with differences in one direction, a **one-tailed test** can be used.

A one-tailed test may take one of two different forms. If the researcher believes that the true population value is greater than the value specified in the null hypothesis, the research hypothesis would use the "$>$" ("greater than") symbol. In our example, if we had predicted that treated alcoholics had *higher* absentee rates than the community (or averaged *more* than 7.2 days of absenteeism), our research hypothesis would have been

$$(H_1\text{: } \mu > 7.2)$$

Where $>$ signifies "greater than"

If we predicted that treated alcoholics had *lower* absentee rates than the community (or averaged *fewer* than 7.2 days of absenteeism), our research hypothesis would have been

$$(H_1\text{: } \mu < 7.2)$$

Where $<$ signifies "less than"

One-tailed tests are often appropriate when programs designed to solve a problem or improve a situation are being evaluated. If the program for treating alcoholics made them *less* reliable workers, for example, the program would be a failure, at least on that criterion. In this situation, the researcher may focus only on outcomes that would indicate that the program is a success (i.e., when treated alcoholics have lower rates) and conduct a one-tailed test with a research hypothesis in the form: H_1: $\mu < 7.2$.

Or consider the evaluation of a program designed to reduce unemployment. The evaluators would be concerned only with outcomes that show a decrease in the unemployment rate. If the rate shows no change or if unemployment increases, the program is a failure, and both of these outcomes might be considered equally negative by the researchers. Thus, the researchers could legitimately use a one-tailed test that stated that unemployment rates for graduates of the program would be less than ($<$) rates in the community.

One- Versus Two-Tailed Tests and the Five-Step Model

In terms of the five-step model, the choice of a one-tailed or two-tailed test determines what we do with the critical region in step 3. In a two-tailed test, we split the critical region equally into the upper and lower tails of the sampling distribution. In a one-tailed test, we place the entire critical area in either the upper or lower tail. If we believe that the population characteristic is greater than the value stated in the null hypothesis (if H_1 includes the $>$ symbol), we place the entire critical region in the upper tail. If we believe that the characteristic is less than the

TABLE 7.2 One- vs. Two-Tailed Tests, $\alpha = 0.05$

If the Research Hypothesis (H_1) Uses	The Test Is	Concern Is on	Z(critical) Is
=	Two-tailed	Both tails	±1.96
>	One-tailed	Upper tail	+1.65
<	One-tailed	Lower tail	−1.65

value stated in the null hypothesis (if H_1 includes the $<$ symbol), the entire critical region goes in the lower tail.

For example, in a two-tailed test with alpha equal to 0.05, the critical region begins at Z(critical) $= \pm 1.96$. In a one-tailed test at the same alpha level, the Z(critical) is $+1.65$ if the upper tail is specified and -1.65 if the lower tail is specified. Table 7.2 summarizes the procedures to follow in terms of the nature of the research hypothesis. The difference in placing the critical region is graphically summarized in Figure 7.5, and the critical Z scores for the most common alpha levels are given in Table 7.3 for both one- and two-tailed tests.

Note that, for a given alpha level, the critical Z values for one-tailed tests are closer to the mean of the sampling distribution. Thus, a one-tailed test is more likely to reject H_0 without changing the alpha level (assuming that we have specified the correct tail). One-tailed tests are a way of statistically both having and eating your cake and should be used whenever (1) the direction of the difference can be confidently predicted or (2) the researcher is concerned only with differences in one tail of the sampling distribution. An example should clarify these procedures.

Using a One-Tailed Test

A sociologist has noted that sociology majors seem more sophisticated, charming, and cosmopolitan than the rest of the student body. A "Sophistication Scale" test has been administered to the entire student body and to a random sample of 100 sociology majors, and these results have been obtained:

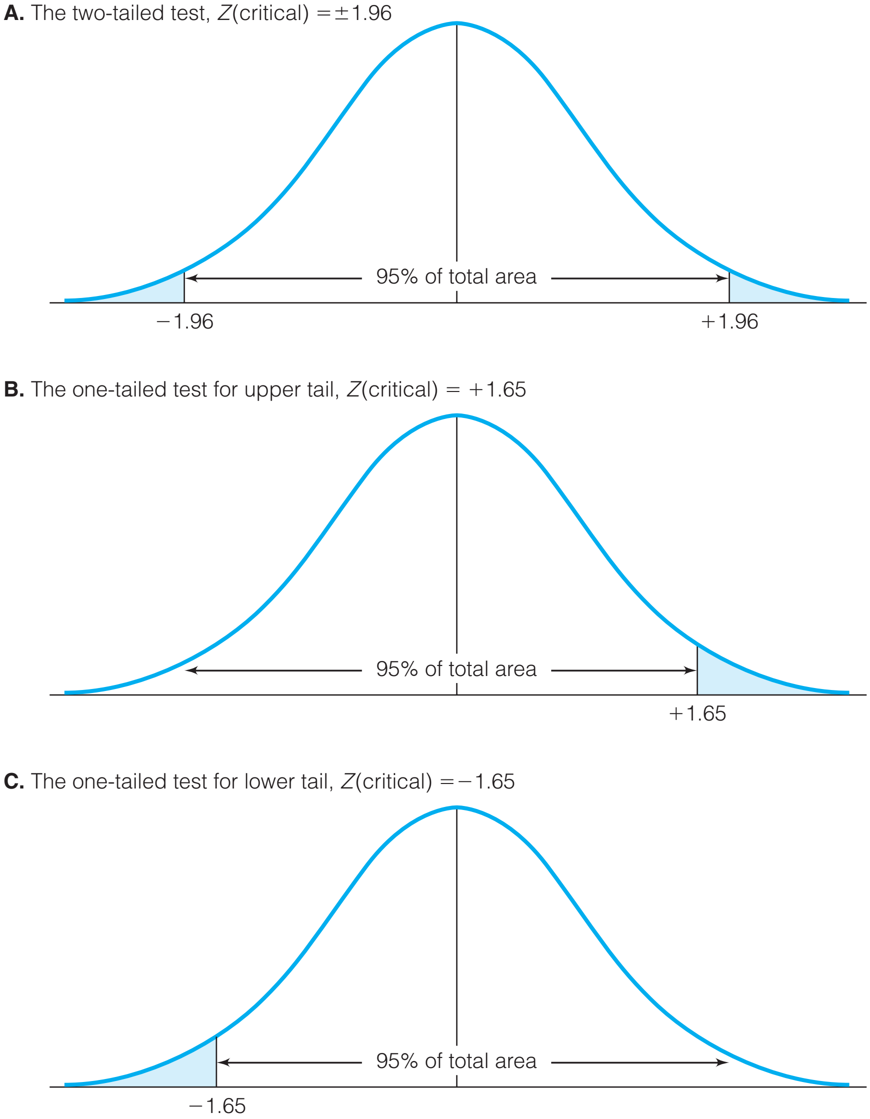
Student Body	Sociology Majors
$\mu = 17.3$	$\bar{X} = 19.2$
$\sigma = 7.4$	$N = 100$

We will use the five-step model to test the H_0 of no difference between sociology majors and the general student body.

TABLE 7.3 Finding Critical Z Scores for One-Tailed Tests

		One-Tailed Value	
Alpha	Two-Tailed Value	Upper Tail	Lower Tail
0.10	±1.65	+1.29	−1.29
0.05	±1.96	+1.65	−1.65
0.01	±2.58	+2.33	−2.33
0.001	±3.32	+3.10	−3.10
0.0001	±3.90	+3.70	−3.70

FIGURE 7.5 Establishing the Critical Region, One-Tailed Tests Versus Two-Tailed Tests (alpha = 0.05)



Step 1. Making Assumptions and Meeting Test Requirements. Because we are using a mean to summarize the sample outcome, we must assume that the Sophistication Scale generates interval-ratio-level data. With a sample size of 100, the Central Limit Theorem applies, and we can assume that the sampling distribution is normal in shape.

Model: Random sampling
Level of measurement is interval-ratio
Sampling distribution is normal

Step 2. Stating the Null Hypothesis (H_0). The null hypothesis states that there is no difference between sociology majors and the general student body. The research hypothesis (H_1) will also be stated at this point. The researcher has predicted a direction for the difference ("Sociology majors are *more* sophisticated"), so a one-tailed test is justified. The one-tailed research hypothesis asserts that sociology majors have a higher (>) score on the Sophistication Scale. The two hypotheses may be stated as

$$H_0: \mu = 17.3$$
$$(H_1: \mu > 17.3)$$

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region. We will use Appendix A to find areas under the sampling distribution. If alpha is set at 0.05, the critical region will begin at the Z score $+1.65$. That is, the researcher has predicted that sociology majors are *more* sophisticated and that this sample comes from a population that has a mean *greater than* 17.3, so he or she will be concerned only with sample outcomes in the upper tail of the sampling distribution. If sociology majors are *the same as* other students in terms of sophistication (if H_0 is true) or if they are *less* sophisticated (and come from a population with a mean less than 17.3), the theory is not supported. These decisions may be summarized as

$$\text{Sampling distribution} = Z \text{ distribution}$$
$$\alpha = 0.05$$
$$Z(\text{critical}) = +1.65$$

Step 4. Computing the Test Statistic.

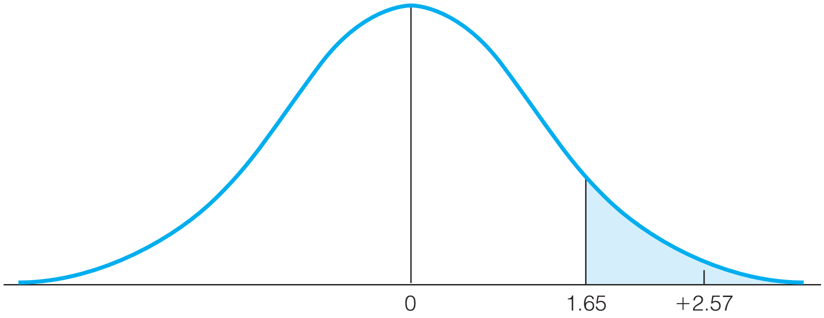
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$
$$Z = \frac{19.2 - 17.3}{7.4/\sqrt{100}}$$
$$Z = \frac{1.90}{0.74}$$
$$Z = +2.57$$

Step 5. Making a Decision and Interpreting Test Results. In this step, we compare the Z(obtained) with the Z(critical):

$$Z(\text{critical}) = +1.65$$
$$Z(\text{obtained}) = +2.57$$

The test statistic falls in the critical region and this outcome is depicted in Figure 7.6. We will reject the null hypothesis because, if H_0 were true, a difference of this size would be very unlikely. There is a significant difference between sociology

FIGURE 7.6 ***Z*(Obtained) Versus *Z*(Critical) (alpha = 0.05, One-Tailed Test)**



majors and the general student body in terms of sophistication. Since the null hypothesis has been rejected, the research hypothesis (sociology majors are more sophisticated) is supported. *(For practice in dealing with tests of significance for means that may call for one-tailed tests, see problems 7.2b, 7.3, 7.6, 7.8, and 7.17.)*

Selecting an Alpha Level

In addition to deciding between one-tailed and two-tailed tests, the researcher must select an alpha level. We have seen that the alpha level plays a crucial role in hypothesis testing. When we assign a value to alpha, we define what we mean by an "unlikely" sample outcome. If the probability of the observed sample outcome is lower than the alpha level (if the test statistic falls in the critical region), then we reject the null hypothesis.

How can we make reasonable decisions about the value of alpha? Recall that, in addition to defining what we mean by *unlikely*, the alpha level is the probability that the decision to reject the null hypothesis if the test statistic falls in the critical region will be incorrect. In hypothesis testing, the error of incorrectly rejecting the null hypothesis—or rejecting a null hypothesis that is actually true—is called **Type I error**, or **alpha error**. To minimize this type of error, use very small values for alpha.

To elaborate: When we select an alpha level, we divide the sampling distribution into two sets of possible sample outcomes. One, the critical region, includes all unlikely or rare sample outcomes. The other—the remainder of the area—consists of all sample outcomes that are not rare. A lower alpha level reduces the size of the critical region and moves the critical region further from the mean of the sampling distribution.

We can illustrate this point with the alpha levels and *Z*(critical) scores in Table 7.4 (which were also presented in Table 7.3). Note that the critical region becomes smaller and moves farther away from the mean of the sampling distribution as alpha levels go down. Thus, lower alpha levels make it less likely that we will reject the null hypothesis and—because a Type I error can occur only if we reject the null hypothesis—less likely that we will commit a Type I error.

TABLE 7.4 The Relationship Between Alpha and Z(critical) for a Two-Tailed Test

If Alpha =	The Two-Tailed Critical Region Will Begin at Z(critical) =
0.100	±1.65
0.050	±1.96
0.010	±2.58
0.001	±3.32

To minimize the probability of rejecting a null hypothesis that is in fact true, use very low alpha levels.

However, there is a complication. As the critical region decreases in size (as alpha levels decrease), the "noncritical" region becomes larger. This increases the chance that the sample outcome will not fall in the critical region and that our decision in step 5 will be to "fail to reject."

This raises the possibility of a second type of mistake, called **Type II error**, or **beta (β) error**: failing to reject a null hypothesis that is, in fact, false. As the alpha level decreases, the probability of a Type I error decreases, but the probability of a Type II error increases. The two types of error are inversely related, and it is not possible to minimize both in the same test. As the probability of one type of error decreases, the probability of the other increases, and vice versa.

Table 7.5 may clarify these relationships. The table lists the two decisions we can make in step 5 of the five-step model (reject or fail to reject the null hypothesis) along with the two possible conditions of the null hypothesis (it is either actually true or actually false). The table combines these possibilities into a total of four possible outcomes. The first two outcomes are desirable ("OK") but the other two indicate that an error has been made.

TABLE 7.5 Decision Making and the Five-Step Model

	If Our Decision Is to	And the H_0 Is Actually	The Result Is
a	Reject the H_0	False	OK
b	Fail to reject the H_0	True	OK
c	Reject the H_0	True	Type I or alpha (α) error
d	Fail to reject the H_0	False	Type II or beta (β) error

STATISTICS IN EVERYDAY LIFE

Setting the Alpha Level

In social science research, the 0.05 alpha level has become the standard indicator of a significant difference. This alpha level means that we will incorrectly reject the null hypothesis only five times out of every 100 tests. These might seem like excellent odds, but research that involves potentially harmful drugs would call for even lower alpha levels (0.001, 0.0001, or even lower) to minimize the possibility of making incorrect decisions and endangering health.

Let's consider the two desirable ("OK") outcomes first. Obviously, we always want to reject false null hypotheses and "fail to reject" those that are true. The goals of any scientific investigation are to verify the truth and to refute incorrect statements.

The remaining two combinations are errors. If we reject a null hypothesis that is actually true, we are saying that a true statement is false. Likewise, if we fail to reject a null hypothesis that is actually false, we are saying that a false statement is true. We always want to reject false statements and accept the truth when we find it. Remember, however, that hypothesis testing always carries an element of risk and that it is not possible to minimize the chances of both Type I and Type II errors simultaneously.

What all of this means, finally, is that you must think of selecting an alpha level as an attempt to balance the two types of error. Higher alpha levels will minimize the probability of Type II error (saying that false statements are true), and lower alpha levels will minimize the probability of Type I error (saying that true statements are false). Normally, in social science research, we will want to minimize Type I error, and lower alpha levels (0.05, 0.01, 0.001, or lower) will be used. The 0.05 level in particular has emerged as a generally recognized indicator of a significant result. However, the widespread use of the 0.05 level is simply a convention, and there is no reason that alpha cannot be set at virtually any sensible level (such as 0.04, 0.027, 0.083). The researcher has the responsibility of selecting the alpha level that seems most reasonable in terms of the goals of the research project.

The Student's *t* Distribution

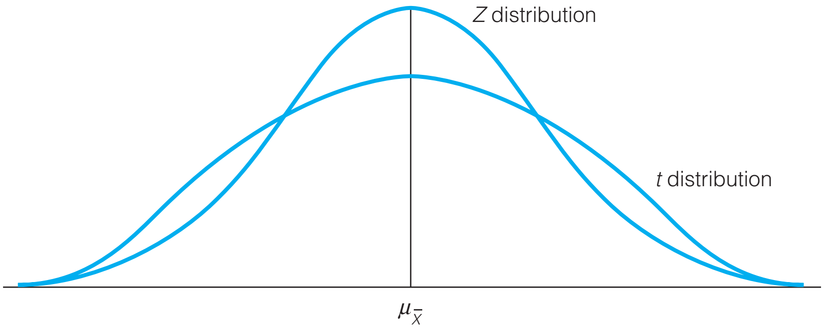
To this point, we have considered only situations involving single-sample means where the value of the population standard deviation (σ) was known. Needless to say, in most research situations, the value of σ will be unknown. However, a value for σ is required in order to compute the standard error of the mean ($\sigma/\sqrt{N}$), convert our sample outcome into a Z score, and place the Z(obtained) on the sampling distribution (step 4). How can we reasonably obtain a value for the population standard deviation?

It might seem sensible to estimate σ with s, the sample standard deviation. As we noted in Chapter 6, s is a biased estimator of σ, but the degree of bias decreases as sample size increases. For large samples (that is, samples with 100 or more cases), the sample standard deviation yields an adequate estimate of σ. Thus, for large samples, we simply substitute s for σ in the formula for Z(obtained) in step 4 and continue to use the standard normal curve to find areas under the sampling distribution.[1]

For smaller samples, however, when s is unknown, an alternative distribution called the **Student's *t* distribution** must be used to find areas under the sampling

[1]Even though its effect will be minor and will decrease with the sample size, we will always correct for the bias in s by using the term $N - 1$ rather than N in the computation for the standard deviation of the sampling distribution when σ is unknown.

FIGURE 7.7 The *t* Distribution and the *Z* Distribution



distribution and establish the critical region. The shape of the *t* distribution varies as a function of sample size. The relative shapes of the *t* and *Z* distributions are depicted in Figure 7.7. For small samples, the *t* distribution is much flatter than the *Z* distribution, but the *t* distribution comes to resemble the *Z* distribution more and more as sample size increases until the two are essentially identical for sample sizes greater than 120. As *N* increases, the sample standard deviation (s) becomes a better estimate of the population standard deviation (σ), and the *t* distribution becomes more and more like the *Z* distribution.

Applying Statistics 7.1 Testing a Sample Mean for Significance

For a random sample of 152 felony cases tried in a local court, the average prison sentence was 27.3 months. Is this significantly different from the average prison term for felons statewide ($\mu = 28.7$)? We will use the five-step model to organize the decision-making process.

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Random sampling
Level of measurement is interval-ratio
Sampling distribution is normal

This is a large sample ($N > 100$) and length of sentence is an interval-ratio variable, so we can conclude that the model assumptions are satisfied

Step 2. Stating the Null Hypothesis (H_0). The null hypothesis states that the average sentence locally (for *all* felony cases) is equal to the statewide average. In symbols:

$$H_0: \mu = 28.7$$

The research question does not specify a direction; it only asks whether the local sentences are "different from" (not higher or lower than) the statewide average. This suggests a two-tailed test:

$$H_1: \mu \neq 28.7$$

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region. This is a large sample, so we can use Appendix A to establish the critical region and state the critical scores as *Z* scores (as opposed to *t* scores).

$$\text{Sampling distribution} = Z \text{ distribution}$$
$$\alpha = 0.05$$
$$Z(\text{critical}) = \pm 1.96$$

Step 4. Computing the Test Statistic. The necessary information for conducting a test of the null hypothesis is

$$\bar{X} = 27.3 \qquad \mu = 28.7$$
$$s = 3.7$$
$$N = 152$$

(*continued*)

The test statistic, Z(obtained), is

$$Z\text{(obtained)} = \frac{\bar{X} - \mu}{s/\sqrt{N-1}}$$

$$Z\text{(obtained)} = \frac{27.3 - 28.7}{3.7/\sqrt{152-1}}$$

$$Z\text{(obtained)} = \frac{-1.40}{3.7/\sqrt{151}}$$

$$Z\text{(obtained)} = \frac{-1.40}{3.7/12.29}$$

$$Z\text{(obtained)} = \frac{-1.40}{0.30}$$

$$Z\text{(obtained)} = -4.67$$

Note that, even though sample size is large, we have used "$N - 1$" rather than "N" under the square root sign. We will do this routinely whenever σ is unknown, even when the effect on the value of Z(obtained) will be very small.

Step 5. Making a Decision and Interpreting the Test Results. With a Z(critical) of ±1.96 and an obtained Z score of −4.67, the null hypothesis is rejected. The difference between the prison sentences of felons convicted in the local court and felons convicted statewide is statistically significant. The difference is so large that we may conclude that it did not occur by random chance. The decision to reject the null hypothesis has a 0.05 probability of being wrong.

The Distribution of *t*: Using Appendix B

The *t* distribution is summarized in Appendix B. The *t* table differs from the *Z* table in three ways.

1. There is a column at the left of the table labeled *df* for "degrees of freedom."[2] Because the exact shape of the *t* distribution varies by sample size, the exact location of the critical region varies as well. Degrees of freedom, which are equal to $N - 1$ in the case of a single-sample mean, must be computed before the critical region for any alpha can be located.
2. Alpha levels are arrayed across the top of Appendix B in two rows, one row for one-tailed tests and one for two-tailed tests. To use the table, begin by locating the selected alpha level in the appropriate row.
3. The entries in the table are the actual scores, called ***t*(critical)**, that mark the beginnings of the critical regions and not areas under the sampling distribution.

To illustrate the use of this table with single-sample means, find the critical region for alpha equal to 0.05, two-tailed test, for $N = 30$. The degrees of freedom will be $N = 1$, or 29; reading down the proper column, you should find a value of 2.045. Thus, the critical region for this test will begin at $t(\text{critical}) = \pm 2.045$.

Notice that this *t*(critical) is larger in value than the comparable *Z*(critical), which for a two-tailed test at an alpha of 0.05 would be ±1.96. This is because the *t* distribution is flatter than the *Z* distribution (see Figure 7.7). On the *t* distribution, the critical regions will begin farther away from the mean of the sampling distribution; therefore, the null hypothesis will be harder to reject. Furthermore, the smaller the sample size (the lower the degrees of freedom), the larger the value of ***t*(obtained)** necessary to reject the H_0.

[2]Degrees of freedom refers to the number of values in a distribution that are free to vary. For a sample mean, a distribution has $N - 1$ degrees of freedom. This means that for a specific value of the mean and of N, $N - 1$ scores are free to vary. For example, if the mean is 3 and $N = 5$, the distribution of five scores would have $5 - 1 = 4$ degrees of freedom. When the values of four of the scores are known, the value of the fifth is fixed. If four scores are 1, 2, 3, and 4, the fifth must be 5 and no other value.

Also note that the values of t(critical) decrease as degrees of freedom increase. For 1 degree of freedom, the t(critical) for an alpha of 0.05, two-tailed test is ± 12.706, but this score becomes smaller for larger samples. For degrees of freedom greater than 120, the value of t(critical) is the same as the comparable value of Z(critical), or ± 1.96. As sample size increases, the t distribution resembles the Z distribution more and more until, with sample sizes greater than 120, the two distributions are essentially identical.[3]

Using the *t* Distribution in a Test of Hypothesis

To demonstrate the uses of the t distribution in more detail, we will work through an example problem. Note that, in terms of the five-step model, the changes occur mostly in steps 3 and 4. In step 3, the sampling distribution will be the t distribution, and degrees of freedom (df) must be computed before locating the critical region or the t(critical) score. In step 4, a slightly different formula for computing the test statistic, t(obtained), will be used. As compared with the formula for Z(obtained), s will replace σ and $N - 1$ will replace N. Specifically,

FORMULA 7.2

$$t(\text{obtained}) = \frac{\bar{X} - \mu}{s/\sqrt{N-1}}$$

A researcher wonders whether commuter students are different from the general student body in terms of academic achievement. She has gathered a random sample of 30 commuter students and has learned from the registrar that the mean grade-point average for all students is 2.50 ($\mu = 2.50$), but the standard deviation of the population (σ) has never been computed. Sample data are reported here. Is the sample from a population that has a mean of 2.50?

Student Body	Commuter Students
$\mu = 2.50$	$\bar{X} = 2.78$
$\sigma = ?$	$s = 1.23$
	$N = 30$

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Random sampling
Level of measurement is interval-ratio
Sampling distribution is normal

Step 2. Stating the Null Hypothesis. You can see from the research hypothesis that the researcher has not predicted a direction for the difference. This will be a two-tailed test.

$$H_0: \mu = 2.50$$
$$(H_1: \mu \neq 2.50)$$

[3]Appendix B abbreviates the t distribution by presenting a limited number of critical t scores for degrees of freedom between 31 and 120. If the degrees of freedom for a specific problem equal 77 and alpha equals 0.05, two-tailed, we have a choice between a t(critical) of 2.000 ($df = 60$) and a t(critical) of 1.980 ($df = 120$). In situations such as these, take the larger table value as t(critical). This will make rejection of H_0 less likely and is therefore the more conservative course of action.

ONE STEP AT A TIME **Completing Step 4 of the Five-Step Model: Computing *t*(Obtained)**

Follow these procedures when using the Student's t distribution.

Step	***Operation***
To compute the test statistic using Formula 7.2:	
1.	Find the square root of $N - 1$.
2.	Divide the quantity you found in step 1 into the sample standard deviation (s).
3.	Subtract the population mean (μ) from the sample mean ($\bar{X}$).
4.	Divide the quantity you found in step 3 by the quantity you found in step 2. This value is t(obtained).

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region. Since σ is unknown and the sample size is small, the t distribution will be used to find the critical region. Alpha will be set at 0.01.

$$\text{Sampling distribution} = t \text{ distribution}$$

$$\alpha = 0.01, \text{ two-tailed test}$$

$$df = (N - 1) = 29$$

$$t(\text{critical}) = \pm 2.756$$

Step 4. Computing the Test Statistic.

$$t(\text{obtained}) = \frac{\bar{X} - \mu}{s/\sqrt{N - 1}}$$

$$t(\text{obtained}) = \frac{2.78 - 2.50}{1.23/\sqrt{29}}$$

$$t(\text{obtained}) = \frac{0.28}{1.23/5.39}$$

$$t(\text{obtained}) = \frac{0.28}{0.23}$$

$$t(\text{obtained}) = 1.22$$

ONE STEP AT A TIME **Completing Step 5 of the Five-Step Model: Making a Decision and Interpreting Results**

Step	***Operation***
1.	Compare the t(obtained) to the t(critical). If t(obtained) is *in* the critical region, *reject* the null hypothesis. If t(obtained) is *not in* the critical region, *fail to reject* the null hypothesis.
2.	Interpret your decision in terms of the original question. For example, our conclusion for the example problem used in this section was "There is no significant difference between the average GPAs of commuter students and the general student body."

FIGURE 7.8 Sampling Distribution Showing *t*(Obtained) Versus *t*(Critical) (α = 0.01, Two-Tailed Test, *df* = 29)

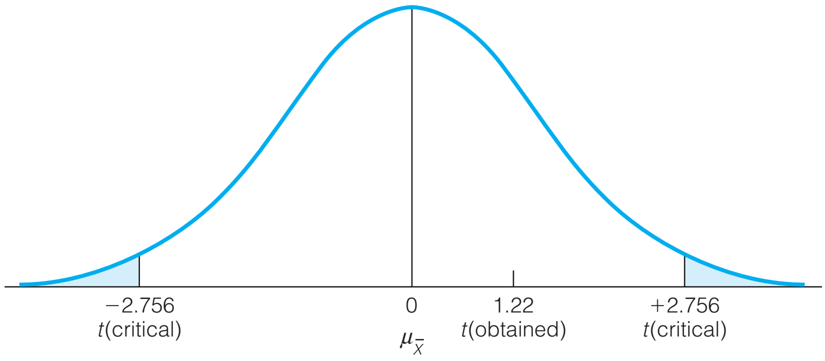


TABLE 7.6 Choosing a Sampling Distribution When Testing Single-Sample Means for Significance

If Population Standard Deviation (σ) Is	Sampling Distribution Is the
Known	*Z* distribution
Unknown and sample size (*N*) is large	*Z* distribution
Unknown and sample size (*N*) is small	*t* distribution

Step 5. Making a Decision and Interpreting Test Results. The test statistic does not fall in the critical region. Therefore, the researcher fails to reject H_0. The difference between the sample mean (2.78) and the population mean (2.50) is not statistically significant. The difference is no greater than what would be expected if only random chance were operating. The test statistic and critical regions are displayed in Figure 7.8.

To summarize, when testing single-sample means, we must make a choice regarding the theoretical distribution we will use to establish the critical region. The choice is straightforward. If the population standard deviation (σ) is known or sample size is large, the *Z* distribution (summarized in Appendix A) will be used. If σ is unknown and the sample is small, the *t* distribution (summarized in Appendix B) will be used. These decisions are summarized in Table 7.6. *(For practice in using the t distribution in a test of hypothesis, see problems 7.8 to 7.10 and 7.17.)*

Tests of Hypothesis for Single-Sample Proportions (Large Samples)

In many cases, we work with variables that are not interval-ratio in level of measurement. One alternative in this situation would be to use a sample proportion (P_s), rather than a sample mean, as the test statistic. As we shall see, the overall procedures for testing single-sample proportions are the same as those for testing means. The central question is still "Does the population from which the sample was drawn have a certain characteristic?" We still conduct the test based on the assumption that the null hypothesis is true, and we still evaluate the probability of the obtained sample outcome against a sampling distribution of all possible sample outcomes. Our decision at the end of the test is also the same. If the obtained test statistic falls in the critical region (and thus is unlikely, given the assumption that H_0 is true), we reject H_0.

Of course, there are also some important differences in significance tests for sample proportions. These differences are best related in terms of the five-step model:

1. In step 1, we assume that the variable is measured at the nominal level.
2. In step 2, the symbols used to state the null hypothesis are different even though it is still a statement of "no difference."
3. In step 3, we will use only the Z distribution to find areas under the sampling distribution and locate the critical region. This will be appropriate as long as sample size is large. We will not consider small-sample tests of hypothesis for proportions in this text.
4. In step 4, computing the test statistic, the form of the formula remains the same. That is, the test statistic, Z(obtained), equals the sample statistic minus the mean of the sampling distribution, divided by the standard deviation of the sampling distribution. However, the symbols change because we are basing the tests on sample proportions. The formula can be stated as

FORMULA 7.3

$$Z(\text{obtained}) = \frac{P_s - P_u}{\sqrt{P_u(1 - P_u)/N}}$$

5. Step 5, making a decision, is exactly the same as before. If the test statistic, Z(obtained), falls in the critical region, reject H_0.

A Test of Hypothesis Using Sample Proportions

An example should clarify these procedures. A random sample of 122 households in a low-income neighborhood revealed that 53 (or a proportion of 0.43) of the households were headed by females. In the city as a whole, the proportion of female-headed households is 0.39. Are households in the lower-income neighborhood significantly different from the city as a whole on this characteristic?

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Random sampling
Level of measurement is nominal
Sampling distribution is normal in shape

Step 2. Stating the Null Hypothesis. The research question asks whether the sample proportion is *different from* the population proportion. A two-tailed test will be used because a direction for the difference has not been predicted.

$$H_0: P_u = 0.39$$

$$(H_1: P_u \neq 0.39)$$

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region.

$$\text{Sampling distribution} = Z \text{ distribution}$$
$$\alpha = 0.10, \text{ two-tailed}$$
$$Z(\text{critical}) = \pm 1.65$$

Step 4. Computing the Test Statistic.

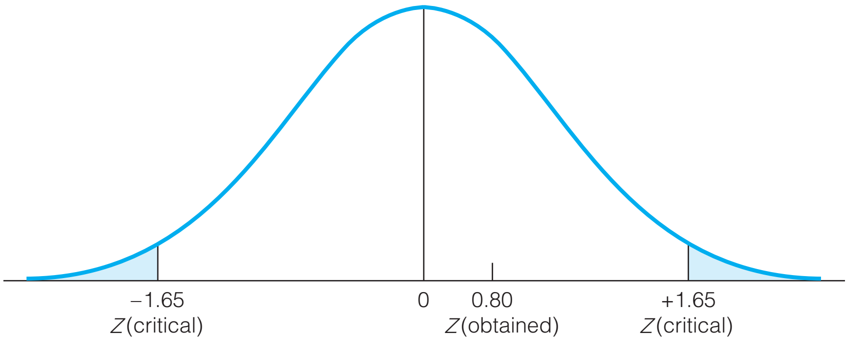
$$Z(\text{obtained}) = \frac{P_s - P_u}{\sqrt{P_u(1 - P_u)/N}}$$
$$Z(\text{obtained}) = \frac{0.43 - 0.39}{\sqrt{(0.39)(0.61)/122}}$$
$$Z(\text{obtained}) = \frac{0.04}{\sqrt{0.24/122}}$$
$$Z(\text{obtained}) = \frac{0.04}{\sqrt{0.002}}$$
$$Z(\text{obtained}) = \frac{0.04}{0.05}$$
$$Z(\text{obtained}) = 0.80$$

Step 5. Making a Decision and Interpreting Test Results. The test statistic, Z(obtained), does not fall in the critical region. Therefore, we fail to reject H_0. There is no statistically significant difference between the low-income community and the city as a whole in terms of the proportion of households headed by females. Figure 7.9 displays the sampling distribution, the critical region, and the Z(obtained). *(For practice in tests of significance using sample proportions, see problems 7.1c, 7.11 to 7.14, 7.15a to d, 7.16, and 7.19a and b.)*

ONE STEP AT A TIME | **Completing Step 4 of the Five-Step Model: Computing *Z*(obtained)**

Step	***Operation***
	To compute the test statistic using Formula 7.3:
1.	Start with the denominator of Formula 7.3 and substitute in the value for P_u. This value will be given in the statement of the problem.
2.	Subtract the value of P_u from 1.
3.	Multiply the value you found in step 2 by the value of P_u.
4.	Divide the quantity you found in step 3 by N.
5.	Take the square root of the value you found in step 4.
6.	Subtract P_u from P_s.
7.	Divide the quantity you found in step 6 by the quantity you found in step 5. This value is Z(obtained).

FIGURE 7.9 Sampling Distribution Showing *Z*(Obtained) Versus *Z*(Critical) (α = 0.10, Two-Tailed Test)



ONE STEP AT A TIME **Completing Step 5 of the Five-Step Model: Making a Decision and Interpreting Results**

Step	***Operation***
1.	Compare your *Z*(obtained) to your *Z*(critical). If *Z*(obtained) is *in* the critical region, *reject* the null hypothesis. If *Z*(obtained) is *not in* the critical region, *fail to reject* the null hypothesis.
2.	Interpret the decision in terms of the original question. For example, our conclusion for the example problem used in this section was "There is no significant difference between the low-income community and the city as a whole in the proportion of households that are headed by females."

Applying Statistics 7.2 Testing a Sample Proportion for Significance

A researcher draws a random sample ($N = 103$) from the most affluent neighborhood in a community and finds that 76% of the respondents voted Republican in the most recent presidential election. For the community as a whole, 66% of the electorate voted Republican. Was the affluent neighborhood significantly more likely to have voted Republican?

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Random sampling
Level of measurement is nominal
Sampling distribution is normal

This is a large sample, so we may assume a normal sampling distribution. Political party preference is a nominal-level variable.

Step 2. Stating the Null Hypothesis (H_0). The null hypothesis says that the affluent neighborhood is not different from the community as a whole.

$$H_0: P_u = 0.66$$

The original question ("Was the affluent neighborhood more likely to vote Republican?") suggests a one-tailed research hypothesis:

$$(H_1: P_u > 0.66)$$

(continued)

Applying Statistics 7.2 (*continued*)

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region.

$$\text{Sampling distribution} = Z \text{ distribution}$$
$$\alpha = 0.05$$
$$Z(\text{critical}) = +1.65$$

The research hypothesis says that we will be concerned only with outcomes in which the neighborhood is *more* likely to vote Republican or with sample outcomes in the upper tail of the sampling distribution.

Step 4. Computing the Test Statistic. The information necessary for a test of the null hypothesis, expressed in the form of proportions, is

Neighborhood	Community
$P_s = 0.76$	$P_u = 0.66$
$N = 103$	

The test statistic, Z(obtained), would be

$$Z(\text{obtained}) = \frac{P_s - P_u}{\sqrt{P_u(1 - P_u)/N}}$$

$$Z(\text{obtained}) = \frac{0.76 - 0.66}{\sqrt{(0.66)(0.34)/103}}$$

$$Z(\text{obtained}) = \frac{0.10}{\sqrt{0.22/103}}$$

$$Z(\text{obtained}) = \frac{0.10}{\sqrt{0.0021}}$$

$$Z(\text{obtained}) = \frac{0.10}{0.05}$$

$$Z(\text{obtained}) = 2.00$$

Step 5. Making a Decision and Interpreting Test Results. With alpha set at 0.05, one-tailed, the critical region begins at $Z(\text{critical}) = +1.65$. With an obtained Z score of 2.00, the null hypothesis is rejected. The difference between the affluent neighborhood and the community as a whole is statistically significant and in the predicted direction. Residents of the affluent neighborhood were significantly more likely to have voted Republican in the last presidential election.

SUMMARY

1. All the basic concepts and techniques for testing hypotheses were presented in this chapter. We saw how to test the null hypothesis of "no difference" for single-sample means and proportions. In both cases, the central question is whether the population represented by the sample has a certain characteristic.
2. All significance tests involve finding the probability of the observed sample outcome, given that the null hypothesis is true. If the outcome has a low probability, we reject the null hypothesis. In the usual research situation, we will wish to reject the null hypothesis and thereby support the research hypothesis.
3. The five-step model will be our framework for decision making throughout the hypothesis-testing chapters. However, what we do during each step will vary, depending on the specific test being conducted.
4. If we can predict a direction for the difference in stating the research hypothesis, a one-tailed test is called for. If no direction can be predicted, a two-tailed test is appropriate.
5. There are two kinds of errors in hypothesis testing. Type I, or alpha, error is rejecting a true null hypothesis; Type II, or beta, error is failing to reject a false null hypothesis. The probabilities of committing these two types of error are inversely related and cannot be simultaneously minimized in the same test. In selecting an alpha level, we try to balance the probability of these two errors.
6. When testing sample means, we use the *t* distribution to find the critical region when the population standard deviation is unknown and sample size is small.
7. Tests of significance based on sample proportions assume a nominal level of measurement, use different symbols to state the null hypothesis, and use a different formula (7.3) to compute Z(obtained).

8. If you are still confused about the uses of inferential statistics described in this chapter, don't be alarmed or discouraged. A sizeable volume of complex material has been presented, and only rarely will someone fully comprehend the unique logic of hypothesis testing on first exposure. After all, it is not every day that you learn how to test a statement you don't believe (the null hypothesis) against a distribution that doesn't exist (the sampling distribution)!

SUMMARY OF FORMULAS

FORMULA 7.1 Single-sample means, large samples:

$$Z(\text{obtained}) = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

FORMULA 7.2 Single-sample means when samples are small and population standard deviation is unknown:

$$t(\text{obtained}) = \frac{\bar{X} - \mu}{s/\sqrt{N-1}}$$

FORMULA 7.3 Single-sample proportions, large samples:

$$Z(\text{obtained}) = \frac{P_s - P_u}{\sqrt{P_u(1 - P_u)/N}}$$

GLOSSARY

Alpha level (α). The proportion of area under the sampling distribution that contains unlikely sample outcomes, given that the null hypothesis is true. Also, the probability of Type I error.

Critical region (region of rejection). The area under the sampling distribution that, in advance of the test itself, is defined as including unlikely sample outcomes, given that the null hypothesis is true.

Five-step model. A step-by-step guideline for conducting tests of hypotheses. A framework that organizes decisions and computations for all tests of significance.

Hypothesis testing. Statistical tests that estimate the probability of sample outcomes if assumptions about the population (the null hypothesis) are true.

Null hypothesis (H_0). A statement of "no difference." In the context of single-sample tests of significance, the null hypothesis states that the population from which the sample was drawn has a certain characteristic or value.

One-tailed test. A type of hypothesis test used when (1) the direction of the difference can be predicted or (2) concern focuses on outcomes in only one tail of the sampling distribution.

Research hypothesis (H_1). A statement that contradicts the null hypothesis. In the context of single-sample tests of significance, the research hypothesis says that the population from which the sample was drawn does not have a certain characteristic or value.

Significance testing. See Hypothesis testing.

Student's *t* distribution. A distribution used to find the critical region for tests of sample means when σ is unknown and sample size is small.

***t*(critical).** The *t* score that marks the beginning of the critical region of a *t* distribution.

***t*(obtained).** The test statistic computed in step 4 of the five-step model. The sample outcome expressed as a *t* score.

Test statistic. The value computed in step 4 of the five-step model that converts the sample outcome into either a *t* score or a *Z* score.

Two-tailed test. A type of hypothesis test used when (1) the direction of the difference cannot be predicted or (2) concern focuses on outcomes in both tails of the sampling distribution.

Type I error (alpha error). The probability of rejecting a null hypothesis that is, in fact, true.

Type II error (beta error). The probability of failing to reject a null hypothesis that is, in fact, false.

***Z*(critical).** The *Z* score that marks the beginnings of the critical region on a *Z* distribution.

***Z*(obtained).** The test statistic computed in step 4 of the five-step model. The sample outcomes expressed as a *Z* score.

PROBLEMS

7.1 **a.** For each situation, find Z(critical).

Alpha (α)	Form of the Test	Z(Critical)
0.05	One-tailed	
0.10	Two-tailed	
0.06	Two-tailed	
0.01	One-tailed	
0.02	Two-tailed	

b. For each situation, find the critical t score.

Alpha(α)	Form of the Test	N	t(Critical)
0.10	Two-tailed	31	
0.02	Two-tailed	24	
0.01	Two-tailed	121	
0.01	One-tailed	31	
0.05	One-tailed	61	

c. Compute the appropriate test statistic (Z or t) for each situation.

	Population	Sample	Z(obtained) or t(obtained)
1.	$\mu = 2.40$ $\sigma = 0.75$	$\bar{X} = 0.20$ $N = 200$	
2.	$\mu = 17.1$	$\bar{X} = 16.8$ $s = 0.9$ $N = 45$	
3.	$\mu = 10.2$	$\bar{X} = 9.4$ $s = 1.7$ $N = 150$	
4.	$P_u = 0.57$	$P_s = 0.60$ $N = 117$	
5.	$P_u = 0.32$	$P_s = 0.30$ $N = 322$	

7.2 **a.** SOC The students at Littlewood Regional High School cut an average of 3.3 classes per month. A random sample of 117 seniors averages 3.8 cuts per month, with a standard deviation of 0.53. Are seniors significantly different from the student body as a whole? *(HINT: The wording of the research question suggests a two-tailed test. This means that the alternative, or research, hypothesis in step 2 will be stated as H_1: $\mu \neq 3.3$ and that the critical region will be split between the upper and lower tails of the sampling distribution. See Table 7.3 for values of Z(critical) for various alpha levels.)*

b. What if the research question were changed to "Do seniors cut a significantly *greater* number of classes"? How would the test conducted in problem 7.2a change? (*HINT: This wording implies a one-tailed test of significance. How would the research hypothesis change? For the alpha you used in problem 7.2a, what would the value of Z(critical) be?*)

7.3 **a.** SW Statewide, social workers average 10.2 years of experience. In a random sample, 203 social workers in greater metropolitan Shinbone average only 8.7 years, with a standard deviation of 0.52. Are social workers in Shinbone significantly less experienced? *(NOTE: The wording of the research hypothesis may justify a one-tailed test of significance. For a one-tailed test, what form would the research hypothesis take, and where would the critical region begin?)*

b. The same sample of social workers reports an average annual salary of \$25,782, with a standard deviation of \$622. Is this figure significantly higher than the statewide average of \$24,509? *(NOTE: The wording of the research hypothesis suggests a one-tailed test. What form would the research hypothesis take, and where would the critical region begin?)*

7.4 SOC Statewide, the average score on the verbal portion of the college entrance exam is 453, with a standard deviation of 95. A random sample of 137 seniors at Littlewood Regional High School shows a mean score of 502. Is there a significant difference?

7.5 SOC A random sample of 423 Asian Americans drawn from the population of a particular state has finished an average of 12.7 years of formal education, with a standard deviation of 1.7. Is this significantly different from the statewide average of 12.2 years?

7.6 SOC A sample of 105 sanitation workers for the city of Euonymus, Texas, earns an average of \$24,375 per year. The average salary for all Euonymus city workers is \$24,230, with a standard deviation of \$523. Are the sanitation workers overpaid? Conduct both one- and two-tailed tests.

7.7 **a.** SOC Statewide, the population as a whole watches 6.2 hours of TV per day. A random sample of 1017 senior citizens in the state reports watching an average of 5.9 hours per day, with a standard deviation of 0.7. Is the difference significant?

b. The same sample of senior citizens reports that they belong to an average of 2.1 voluntary organizations and clubs, with a standard deviation of 0.5. Statewide, the average is 1.7. Is the difference significant?

7.8 *SOC* A school system has assigned several hundred "chronic and severe underachievers" to an alternative educational experience. To assess the program, a random sample of 35 has been selected for comparison with all students in the system.

a. In terms of GPA, did the program work?

Systemwide GPA	Program GPA
$\mu = 2.47$	$\bar{X} = 2.55$
	$s = 0.70$
	$N = 35$

b. In terms of absenteeism (number of days missed per year), what can be said about the success of the program?

Systemwide	Program
$\mu = 6.137$	$\bar{X} = 4.78$
	$s = 1.11$
	$N = 35$

c. In terms of standardized test scores in math and reading, was the program successful?

Math Test—Systemwide	Math Test—Program
$\mu = 103$	$\bar{X} = 106$
	$s = 2.0$
	$N = 35$

Reading Test—Systemwide	Reading Test—Program
$\mu = 110$	$\bar{X} = 113$
	$s = 2.0$
	$N = 35$

(HINT: Note the wording of the research questions. Is a one-tailed test justified? Is the program a success if the students in the program are no different from students systemwide? What if the program students were performing at lower levels? If a one-tailed test is used, what form should the research hypothesis take? Where will the critical region begin?)

7.9 *SOC* A random sample of 26 local sociology graduates scored an average of 458 on the GRE advanced sociology test, with a standard deviation of 20. Is this significantly different from the national average ($\mu = 440$)?

7.10 *SOC* Nationally, the per capita monthly fuel oil bill is $110. A random sample of 36 cities in the Southeast average $78, with a standard deviation of $4. Is the difference significant? Summarize your conclusions in a sentence or two.

7.11 *SOC/CJ* A survey shows that 10% of the population is victimized by property crime each year. A random sample of 527 older citizens (65 years or more of age) shows a victimization rate of 14%. Are older people more likely to be victimized? Conduct both one- and two-tailed tests of significance.

7.12 *SOC/CJ* A random sample of 113 felons convicted of nonviolent crimes in a state prison system completed a program designed to improve their employment possibilities before being released on parole. Fifty-eight eventually became repeat offenders. Is this recidivism rate significantly different from the rate for all offenders in that state (57%)? Summarize your conclusions in a sentence or two. *(HINT: You must use the information given in the problem to compute a sample proportion. Remember to convert the population percentage to a proportion.)*

7.13 *PS* In a recent statewide election, 55% of the voters rejected a proposal to institute a state lottery. In a random sample of 150 urban precincts, 49% of the voters rejected the proposal. Is the difference significant? Summarize your conclusions in a sentence or two.

7.14 *SOC/CJ* Statewide, the police clear by arrest 35% of the robberies and 42% of the aggravated assaults reported to them. A researcher takes a random sample of all the robberies ($N = 207$) and aggravated assaults ($N = 178$) reported to a metropolitan police department in one year and finds that 83 of the robberies and 80 of the assaults were cleared by arrest. Are the local arrest rates significantly different from the statewide rates? Write a sentence or two interpreting your decision.

7.15 *SOC/SW* A researcher has compiled a file of information on a random sample of 317 families that have chronic, long-term patterns of child abuse. Reported here are some of the characteristics of the sample, along with values for the city as a whole. For each trait, test the null hypothesis of "no difference" and summarize your findings.

	Variable	City	Sample
a.	Mother's educational level (proportion completing high school)	$P_u = 0.65$	$P_s = 0.61$
b.	Family size (proportion of families with four or more children)	$P_u = 0.21$	$P_s = 0.26$
c.	Mother's work status (proportion of mothers with jobs outside the home)	$P_u = 0.51$	$P_s = 0.27$
d.	Relations with kin (proportion of families that have contact with kin at least once a week)	$P_u = 0.82$	$P_s = 0.43$
e.	Father's educational achievement (average years of formal schooling)	$\mu = 12.3$	$\bar{X} = 12.5$ $s = 1.7$
f.	Father's occupational stability (average years in present job)	$\mu = 5.2$	$\bar{X} = 3.7$ $s = 0.5$

7.16 SW You are the head of an agency seeking funding for a program to reduce unemployment among teenage males. Nationally, the unemployment rate for this group is 18%. A random sample of 323 teenage males in your area reveals an unemployment rate of 21.7%. Is the difference significant? Can you demonstrate a need for the program? Should you use a one-tailed test in this situation? Why? Explain the result of your test of significance as you would to a funding agency.

7.17 PA The city manager has received a complaint from the local union of firefighters to the effect that they are underpaid. Not having much time, the city manager gathers the records of a random sample of 27 firefighters and finds that their average salary is \$38,073, with a standard deviation of \$575. If she knows that the average salary nationally is \$38,202, how can she respond to the complaint? Should she use a one-tailed test in this situation? Why? What would she say in a memo to the union that would respond to the complaint?

7.18 SW The following essay questions review the basic principles and concepts of inferential statistics. The order of the questions roughly follows the five-step model.

a. Hypothesis testing or significance testing can be conducted only with a random sample. Why?

b. Under what specific conditions can it be assumed that the sampling distribution is normal in shape?

c. Explain the role of the sampling distribution in a test of hypothesis.

d. The null hypothesis is an assumption about reality that makes it possible to test sample outcomes for their significance. Explain.

e. What is the critical region? How is the size of the critical region determined?

f. Describe a research situation in which a one-tailed test of hypothesis would be appropriate.

g. With regard to the shape of the sampling distribution, why does use of the *t* distribution (as opposed to the *Z* distribution) make it more difficult to reject the null hypothesis?

h. What exactly can be concluded in the one-sample case when the test statistic falls in the critical region?

7.19 SOC A researcher is studying changes in the student body at her university and has selected a random sample of 163 from the freshman class. The table below compares their characteristics to the student body as a whole. Which differences are significant?

	Variable	Freshmen	All Students
a.	Proportion Republican	$P_s = 0.44$	$P_u = 0.49$
b.	Proportion majoring in math and science	$P_s = 0.17$	$P_u = 0.09$
c.	Average family income	$\bar{X} = 63{,}542$ $s = 1568$	$\mu = 59{,}235$
d.	Average number of siblings	$\bar{X} = 0.9$ $s = 0.2$	$\mu = 1.2$
e.	Average number of telephone calls to parents per week	$\bar{X} = 1.5$ $s = 0.4$	$\mu = 0.4$
f.	Average number of close friends living on campus	$\bar{X} = 4.5$ $s = 0.3$	$\mu = 7.8$

8 Hypothesis Testing II
The Two-Sample Case

LEARNING OBJECTIVES

By the end of this chapter, you will be able to:

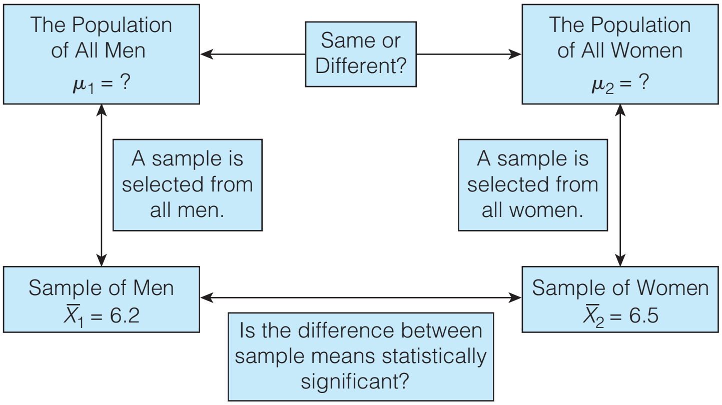
1. Identify and cite examples of situations in which the two-sample test of hypothesis is appropriate.
2. Explain the logic of hypothesis testing as applied to the two-sample case.
3. Explain what an independent random sample is.
4. Perform a test of hypothesis for two-sample means or two-sample proportions following the five-step model and correctly interpret the results.
5. List and explain each of the factors (especially sample size) that affect the probability of rejecting the null hypothesis. Explain the differences between statistical significance and importance.
6. Use SPSS to conduct a test of significance for the difference between sample means.

USING STATISTICS

This chapter presents statistical techniques used to compare two random samples. If the samples are significantly different, we conclude that the populations from which they were selected are different. This conclusion has a known probability of error, often set at 0.05. Examples of these situations include:

1. A researcher is examining the differences in support for gun control between men and women in a sample that is representative of the U.S. population. If the samples are significantly different, she will conclude that the populations of *all* U.S. men and *all* U.S. women differ on this issue.
2. Random samples of the freshman and senior classes at a large university are compared for differences in their political views and religious values. If the researcher finds significant differences between the samples, he will conclude that the populations (*all* freshmen and *all* seniors at the university) differ on these dimensions.
3. Is bullying more of a problem in suburban or city schools? Researchers compare random samples of each type of school and, if the samples are significantly different, they will conclude that all suburban schools are different from all city schools.

FIGURE 8.1 A Test of Hypothesis for Two Sample Means



In Chapter 7, we dealt with hypothesis testing in the one-sample case and were concerned with the significance of the difference between a sample statistic and a population parameter. In this chapter, we consider a new research situation. We will be concerned with the difference between two samples and our question will be: Is the difference between the samples large enough for us to conclude (with a known probability of error) that the populations represented by the samples are different?

For example, do men and women in the United States vary in their support for gun control? Obviously, we cannot ask every male and female for their opinions on this issue. Instead, we must draw random samples and use the information gathered from these samples to infer population patterns. If the difference between the samples is large enough, we will conclude (with a known probability of error) that the populations represented by the samples are different and that there is a real difference between men and women in the population.

Figure 8.1 illustrates the logic of the two-sample case, using gender differences in support for gun control as an example. It may be helpful to compare this diagram with Figure 7.1, which depicted the one-sample case.

In this chapter, we consider tests for the significance of the difference between sample means and sample proportions. In both tests, the five-step model will be the framework for our decision making. The hypothesis-testing process is very similar to that of the one-sample case, but there are some important differences as well.

The One-Sample Case Versus the Two-Sample Case

There are three important differences between the one-sample case considered in Chapter 7 and the two-sample case covered in this chapter:

1. The model in step 1 will include a new assumption: **independent random sampling**. To satisfy this assumption, we select cases for each sample randomly *and* separately from each other. In our example, this means that selecting a specific male would not affect the probability of selecting

any particular female: Each case must be selected independently of all other cases.

This requirement can be satisfied by drawing EPSEM samples from separate lists (for example, one for females and one for males) but it is usually more convenient to draw a single EPSEM sample from a single list of the population and then subdivide the cases into separate groups (males and females, for example). As long as the original sample is random, any subsamples created by the researcher will meet the assumption of independence.

2. The null hypothesis will be stated differently. In the one-sample case, the null hypothesis stated that the population from which the sample was drawn had a certain characteristic. Now, the null hypothesis will say that the two populations are the same (e.g., "There is no significant difference between men and women in their support of gun control"). If the test statistic falls in the critical region, the null hypothesis of no difference between the populations can be rejected, and the argument that the populations are different will be supported.

3. The sampling distribution—the distribution of all possible sample outcomes—will be new. In Chapter 7, the sample outcome was a single statistic, either a mean or a proportion. Now we are dealing with two samples (e.g., samples of men and women), and the sample outcome is the *difference between* the sample statistics. In our example, the sampling distribution would include all possible differences in sample means for support of gun control between men and women. If the null hypothesis is true and men and women do *not* have different views about gun control, the difference between the population means would be zero, the mean of the sampling distribution would be zero, and the huge majority of differences between sample means would be zero or very close to zero. The greater the differences between the sample means, the further the sample outcome (the *difference* between the two sample means) will be from the mean of the sampling distribution (zero), and the more likely that the difference reflects a real difference between the populations represented by the samples.

Hypothesis Testing with Sample Means (Large Samples)

To illustrate the procedure for testing sample means, assume that a researcher has access to a nationally representative random sample and that the individuals in the sample have answered a survey that measures attitudes toward gun control. The sample is divided by gender, and sample statistics are computed for males and females.

As long as sample size is large (that is, as long as the combined number of cases in the two samples exceeds 100), the sampling distribution of the differences in sample means will be normal, and the normal curve (Appendix A) can be used to establish the critical regions. The test statistic, *Z*(obtained), will be computed by the usual formula: sample outcome (the difference between the sample means) minus the mean of the sampling distribution, divided by the standard deviation of the sampling distribution.

The formula for *Z*(obtained) is presented as Formula 8.1. Note that numerical subscripts are used to identify the samples and populations. The "$\overline{X} - \overline{X}$"

subscript attached to σ indicates that we are dealing with the sampling distribution of the *differences* in sample means.

FORMULA 8.1

$$Z(\text{obtained}) = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\overline{X}-\overline{X}}}$$

Where $(\overline{X}_1 - \overline{X}_2)$ = the difference in the sample means
$(\mu_1 - \mu_2)$ = the difference in the population means
$\sigma_{\overline{X}-\overline{X}}$ = the standard deviation of the sampling distribution of the differences in sample means

Recall that tests of significance are always based on the assumption that the null hypothesis is true. If the means of the two populations are equal, then the term $(\mu_1 - \mu_2)$ will be 0 and can be dropped from the equation. In effect, then, the formula we will actually use to compute the test statistic in step 4 will be

FORMULA 8.2

$$Z(\text{obtained}) = \frac{(\overline{X}_1 - \overline{X}_2)}{\sigma_{\overline{X}-\overline{X}}}$$

For large samples, the standard deviation of the sampling distribution of the difference in sample means is defined as

FORMULA 8.3

$$\sigma_{\overline{X}-\overline{X}} = \sqrt{\frac{\sigma_1^2}{N_1} - \frac{\sigma_2^2}{N_2}}$$

Because we will rarely, if ever, know the values of the population standard deviations (σ_1 and σ_2), we must use the sample standard deviations, corrected for bias, to estimate them. Formula 8.4 displays the equation used to estimate the standard deviation of the sampling distribution in this situation. This is called a **pooled estimate** because it combines information from both samples.

FORMULA 8.4

$$\sigma_{\overline{X}-\overline{X}} = \sqrt{\frac{s_1^2}{N_1 - 1} + \frac{s_2^2}{N_2 - 1}}$$

The sample outcomes for support of gun control are

Sample 1 (Men)	Sample 2 (Women)
$\overline{X}_1 = 6.2$	$\overline{X}_2 = 6.5$
$s_1 = 1.3$	$s_2 = 1.4$
$N_1 = 324$	$N_2 = 317$

We see from the sample statistics that men have a lower average score and are less supportive of gun control. The test of hypothesis will tell us whether this difference is large enough to conclude that it did not occur by random chance and reflects an actual difference between all men and all women on this issue.

Step 1. Making Assumptions and Meeting Test Requirements. We now assume that the random samples are independent, but the rest of the model is the same as in the one-sample case.

Model: Independent random samples
Level of measurement is interval-ratio
Sampling distribution is normal

Step 2. Stating the Null Hypothesis. The null hypothesis states that the *populations* represented by the samples are not different on this variable. No direction for the difference has been predicted, so a two-tailed test is called for:

$$H_0: \mu_1 = \mu_2$$
$$(H_1: \mu_1 \neq \mu_2)$$

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region. For large samples, the Z distribution can be used to find areas under the sampling distribution and establish the critical region. Alpha will be set at 0.05.

$$\text{Sampling distribution} = Z \text{ distribution}$$
$$\text{Alpha} = 0.05$$
$$Z(\text{critical}) = \pm 1.96$$

Step 4. Computing the Test Statistic. The population standard deviations are unknown, so Formula 8.4 will be used to estimate the standard deviation of the sampling distribution. This value will then be substituted into Formula 8.2 and Z(obtained) will be computed.

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{s_1^2}{N_1 - 1} + \frac{s_2^2}{N_2 - 1}}$$
$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{(1.3)^2}{324 - 1} + \frac{(1.4)^2}{317 - 1}}$$
$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{1.69}{323} + \frac{1.96}{316}}$$
$$\sigma_{\bar{X}-\bar{X}} = \sqrt{0.0052 + 0.0062}$$
$$\sigma_{\bar{X}-\bar{X}} = \sqrt{0.0114}$$
$$\sigma_{\bar{X}-\bar{X}} = 0.107$$
$$Z(\text{obtained}) = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}-\bar{X}}}$$
$$Z(\text{obtained}) = \frac{6.2 - 6.5}{0.107}$$
$$Z(\text{obtained}) = \frac{-0.300}{0.107}$$
$$Z(\text{obtained}) = -2.80$$

Step 5. Making a Decision and Interpreting the Results of the Test. Comparing the test statistic with the critical region:

$$Z(\text{obtained}) = -2.80$$
$$Z(\text{critical}) = \pm 1.96$$

We see that the Z score falls in the critical region, which means that a difference as large as -0.3 $(6.2 - 6.5)$ between the sample means is unlikely if the null

ONE STEP AT A TIME **Completing Step 4 of the Five-Step Model: Computing Z(obtained)**

Use these procedures when samples are large. Solve Formula 8.4 first and then solve Formula 8.2.

Step	***Operation***
To solve Formula 8.4:	
1.	Subtract 1 from N_1.
2.	Square the value of the standard deviation for the first sample (s_1^2).
3.	Divide the quantity you found in step 2 by the quantity you found in step 1.
4.	Subtract 1 from N_2.
5.	Square the value of the standard deviation for the second sample (s_2^2).
6.	Divide the quantity you found in step 5 by the quantity you found in step 4.
7.	Add the quantity you found in step 6 to the quantity you found in step 3.
8.	Take the square root of the quantity you found in step 7.
To solve Formula 8.2:	
1.	Subtract $\bar{X}_2$ from $\bar{X}_1$.
2.	Divide the value you found in step 1 by the quantity you found in step 8 above. This is Z(obtained).

hypothesis is true. The null hypothesis of no difference is rejected and the alternative hypothesis ("men and women differ in their support of gun control") is supported. The decision to reject the null hypothesis has only a 0.05 probability (the alpha level) of being incorrect.

Note that the value for Z(obtained) is negative, indicating that men have significantly lower support for gun control. The sign of the test statistic reflects our arbitrary decision to label men sample 1 and women sample 2. If we had reversed the labels and called women sample 1 and men sample 2, the sign of the Z(obtained) would have been positive, but its value (2.80) would have been exactly the same, as would our decision in step 5. *(For practice in testing the significance of the difference between sample means for large samples, see problems 8.1 to 8.6, 8.9, 8.15d to f, 8.16d and e, and 8.17 to 8.18.)*

ONE STEP AT A TIME **Completing Step 5 of the Five-Step Model: Making a Decision and Interpreting Results**

Step	***Operation***
1.	Compare the Z(obtained) to your Z(critical). If Z(obtained) is *in* the critical region, *reject* the null hypothesis. If Z(obtained) is *not in* the critical region, *fail to reject* the null hypothesis.
2.	Interpret the decision to reject or fail to reject the null hypothesis in terms of the original question. For example, our conclusion for the example problem was "There is a significant difference between men and women in their support for gun control."

Applying Statistics 8.1 A Test of Significance for Sample Means

Do men and women have different ideas about the ideal number of children a family should have? Means and standard deviations have been computed for both groups, from the full 2012 General Social Survey, a nationally representative sample of U.S. adults:

Sample 1 (Men)	Sample 2 (Women)
$\bar{X}_1 = 2.96$	$\bar{X}_2 = 3.28$
$s_1 = 1.72$	$s_2 = 2.00$
$N_1 = 578$	$N_2 = 681$

We can see from the sample results that women have a higher score than men, on the average. Is this difference significant?

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Independent random samples
Level of measurement is interval-ratio
Sampling distribution is normal

Step 2. Stating the Null Hypothesis.

$$H_0: \mu_1 = \mu_2$$
$$H_0: \mu_1 \neq \mu_2$$

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region.

$$\text{Sampling distribution} = Z \text{ distribution}$$
$$\text{Alpha} = 0.05, \text{two-tailed}$$
$$Z(\text{critical}) = \pm 1.96$$

Step 4. Computing the Test Statistic.

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{s_1^2}{N_1 - 1} + \frac{s_2^2}{N_2 - 1}}$$
$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{(1.72)^2}{578 - 1} + \frac{(2.00)^2}{681 - 1}}$$
$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{2.96}{577} + \frac{4.00}{680}}$$
$$\sigma_{\bar{X}-\bar{X}} = \sqrt{0.0051 + 0.0059}$$
$$\sigma_{\bar{X}-\bar{X}} = \sqrt{0.011}$$
$$\sigma_{\bar{X}-\bar{X}} = 0.11$$
$$Z(\text{obtained}) = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{X}-\bar{X}}}$$
$$Z(\text{obtained}) = \frac{2.96 - 3.28}{0.11}$$
$$Z(\text{obtained}) = \frac{-0.32}{0.11}$$
$$Z(\text{obtained}) = -2.91$$

Step 5. Making a Decision and Interpreting the Results of the Test. Comparing the test statistic with the critical region,

$$Z(\text{obtained}) = -2.91$$
$$Z(\text{critical}) = \pm 1.96$$

We reject the null hypothesis. Men and women are significantly different in their view of the ideal number of children. Given the direction of the difference, we can also note that women's idea of the ideal family size is larger, on the average, than men's. What are some possible reasons for this difference? How could you test your proposed explanation?

STATISTICS IN EVERYDAY LIFE

Reading and Math

The reading scores of American fourth graders in 2012 were significantly higher than scores in the early 1970s, when the tests were first administered. However, the average score in 2012 (221) was essentially unchanged from 2008 (220). The same pattern holds for scores in mathematics: the average score in 2012 was significantly higher than the early 1970s but virtually the same as in 2008. Tests are administered every four years to very large, randomly-selected samples of about 180,000 4th-graders.

Source: National Center for Education Statistics. 2012. The Nations Report Card: Trends in Academic Progress, 2012. Reports are available at http://nces.ed.gov/nationsreportcard/pubs/main2012/2013456.aspx

Hypothesis Testing with Sample Means (Small Samples)

When the population standard deviation is unknown and sample size is small (combined N's of less than 100), the Z distribution cannot be used to find areas under the sampling distribution. Instead, we will use the t distribution to find the critical region and identify unlikely sample outcomes. To do this, we need to perform one additional calculation and make one additional assumption. The calculation is for degrees of freedom, which we need in order to use the t table (Appendix B). In the two-sample case, degrees of freedom are equal to $N_1 + N_2 - 2$.

The additional assumption is a more complex matter. When samples are small, we must assume that the population variances are equal in order to justify the assumption of a normal sampling distribution and to form a pooled estimate of the standard deviation of the sampling distribution. There are ways to test the assumption of equal variances in the population but, for our purposes, we will simply assume equal population variances without formal testing. This assumption is safe as long as sample sizes are approximately equal.

The Five-Step Model and the *t* Distribution

To illustrate this procedure, assume that a researcher believes that center-city families have significantly more children than suburban families. Random samples from both areas are gathered and the following sample statistics computed.

Sample 1 (Suburban)	Sample 2 (Center-City)
$\bar{X}_1 = 2.37$	$\bar{X}_2 = 2.78$
$s_1 = 0.63$	$s_2 = 0.95$
$N_1 = 42$	$N_2 = 37$

The sample data show a difference in the predicted direction. The significance of this observed difference can be tested with the five-step model.

Step 1. Making Assumptions and Meeting Test Requirements. Sample size is small, and the population standard deviation is unknown. Hence, we must assume equal population variances in the model.

Model: Independent random samples
Level of measurement is interval-ratio
Population variances are equal
Sampling distribution is normal

Step 2. Stating the Null Hypothesis. A direction has been predicted, so a one-tailed test will be used. The research hypothesis is stated accordingly.

$$H_0: \mu_1 = \mu_2$$
$$(H_1: \mu_1 < \mu_2)$$

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region. With small samples, the t distribution is used to establish the critical region. Alpha will be set at 0.05, and a one-tailed test will be used.

$$\text{Sampling distribution} = t \text{ distribution}$$
$$\text{Alpha} = 0.05\text{, one-tailed}$$
$$\text{Degrees of freedom} = N_1 + N_2 - 2 = 42 + 37 - 2 = 77$$
$$t(\text{critical}) = -1.671$$

Note that the critical region is placed in the lower tail of the sampling distribution in accordance with the direction specified in H_1.

Step 4. Computing the Test Statistic. With small samples, a different formula (Formula 8.5) is used for the pooled estimate of the standard deviation of the sampling distribution. This value is then substituted directly into the denominator of the formula for t(obtained) in Formula 8.6.

FORMULA 8.5

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}}\sqrt{\frac{N_1 + N_2}{N_1 N_2}}$$

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{(42)(0.63)^2 + (37)(0.95)^2}{42 + 37 - 2}}\sqrt{\frac{42 + 37}{(42)(37)}}$$

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{(42)(0.40) + (37)(0.90)}{77}}\sqrt{\frac{79}{1554}}$$

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{50.10}{77}}\sqrt{\frac{79}{1554}}$$

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{0.65}\sqrt{0.05}$$

$$\sigma_{\bar{X}-\bar{X}} = (.81)(.22)$$

$$\sigma_{\bar{X}-\bar{X}} = 0.18$$

FORMULA 8.6

$$t(\text{obtained}) = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{X}-\bar{X}}}$$

$$t(\text{obtained}) = \frac{2.37 - 2.78}{0.18}$$

$$t(\text{obtained}) = \frac{-0.41}{0.18}$$

$$t(\text{obtained}) = -2.28$$

Step 5. Making a Decision and Interpreting the Results of the Test. The test statistic falls in the critical region:

$$t(\text{obtained}) = -2.28$$
$$t(\text{critical}) = -1.671$$

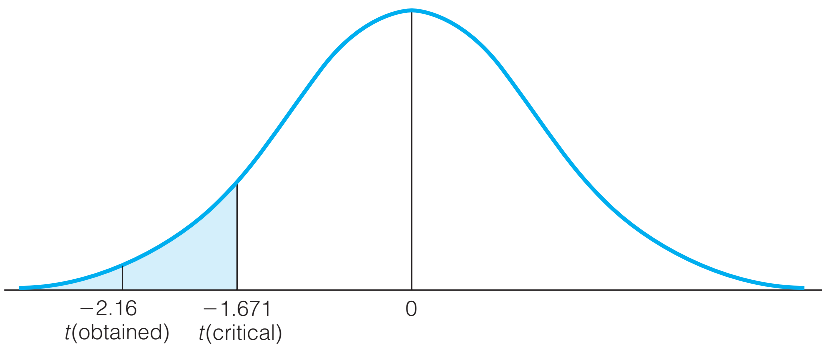
ONE STEP AT A TIME — Completing Step 4 of the Five-Step Model: Computing *t*(obtained)

Solve Formula 8.5 first.

Step	**Operation**
To solve Formula 8.5:	
1.	Add N_1 and N_2 and then subtract 2 from this total.
2.	Square the standard deviation for the first sample (s_1^2) and multiply the result by N_1.
3.	Square the standard deviation for the second sample (s_2^2) and multiply the result by N_2.
4.	Add the quantities you found in steps 2 and 3.
5.	Divide the quantity you found in step 4 by the quantity you found in step 1 and take the square root of the result.
6.	Multiply N_1 by N_2.
7.	Add N_1 and N_2.
8.	Divide the quantity you found in step 7 by the quantity you found in step 6 and take the square root of the result.
9.	Multiply the quantity you found in step 8 by the quantity you found in step 5.
To solve Formula 8.6:	
1.	Subtract $\bar{X}_2$ from $\bar{X}_1$.
2.	Divide the difference between the sample means by the quantity you found in step 9 above. This is *t*(obtained).

If the null hypothesis (H_0: $\mu_1 = \mu_2$) were true, this would be a very unlikely outcome, so the null hypothesis can be rejected. There is a statistically significant difference (a difference so large that it is unlikely to be due to random chance) in the sizes of center-city and suburban families. Furthermore, center-city families are significantly larger in size. The test statistic and sampling distribution are depicted in Figure 8.2. *(For practice in testing the significance of the difference between sample means for small samples, see problems 8.7 and 8.8.)*

FIGURE 8.2 The Sampling Distribution with Critical Region and Test Statistic Displayed



ONE STEP AT A TIME	Completing Step 5 of the Five-Step Model: Making a Decision and Interpreting Results
Step	***Operation***
1.	Compare the *t*(obtained) to your *t*(critical). If *t*(obtained) is *in* the critical region, *reject* the null hypothesis. If *t*(obtained) is *not in* the critical region, *fail to reject* the null hypothesis.
2.	Interpret the decision to reject or fail to reject the null hypothesis in terms of the original question. For example, our conclusion for the example problem was "There is a significant difference between the size of center city and suburban families."

Using SPSS to Test Sample Means for Significant Differences

SPSS provides several tests for the significance of the difference between means. We'll use the Independent-Samples T Test, which is similar to the test presented in this chapter for large samples.

To provide an illustration, we will compare the social class standings of high school graduates and college graduates. Level of education is the independent variable and the dependent variable will be *rank*, the respondent's self-ranking of his or her social position on a scale of 1 to 10. We expect that college graduates will have higher social class standing and, because *rank* is scored so that 1 is the highest score, a significantly lower average score on this variable. If the difference between the sample means is large enough, we can conclude that similar differences exist in the population. You will have the opportunity to use this procedure on your own later in this chapter.

To begin,

1. Click the SPSS icon on your desktop.
2. Load the *GSS2012* database.
   a. Find the **File** command on the far left of the menu bar and then click **File → Open → Data**.
   b. Find the *GSS2012* database and click the file name.
3. From the main menu bar, click **Analyze**, then **Compare Means**, and then **Independent-Samples T Test**. The "Independent-Samples T Test" dialog box will open with the usual list of variables on the left.
4. Find *rank*, the dependent variable, and click the top arrow in the middle of the window to move the variable name to the "Test Variable(s)" box.
5. To select the independent variable, find *degree* in the list of variables and click the bottom arrow in the middle of the window to move *degree* to the "Grouping Variable" box.
6. Two question marks will appear in the "Grouping Variable" box, and the **Define Groups** button will become active. SPSS needs to know which cases

go in which groups. There are five different levels of education on *degree* and we could compare any two of them with this procedure. We need to tell SPSS that we want to compare high school graduates (indicated by a score of 1 on *degree*) and college graduates (a score of 3 on *degree*).

7. Click the **Define Groups** button, and the "Define Groups" window will appear. The cursor will be blinking in the box beside Group 1. SPSS is asking for the score that will determine which cases go in this group. Type 1 (for high school graduates) in this box and then click the box next to Group 2 and type 3 (for college graduates).
8. Click **Continue** to return to the "Independent-Samples T Test" window and click **OK**.

The output is presented in two boxes, the first of which looks like this:

GROUP STATISTICS

	RS HIGHEST DEGREE	N	Mean	Std. Deviation	Std. Error Mean
RS SELF RANKING OF SOCIAL POSITION	HIGH SCHOOL	653	4.90	1.794	.070
	BACHELOR	264	4.34	1.591	.098

This "Group Statistics" box presents descriptive statistics. There were 653 high school graduates, and their average score on *rank* was 4.90; the 264 college graduates averaged 4.34. We can see from this output that the sample means are different and that college graduates average a lower score (as we predicted). Is the difference between the sample means significant?

The results of the test for significance are reported in the next block of output. SPSS does a separate test for each of two assumptions about the population variance, but we will look only at the "Equal Variances Assumed" reported in the top row. This is the model used in this chapter.

		Levene's Test for Equality of Variances		*t*-Test of Equality of Means		
		F	Sig.	*t*	df	Sig. (2-tailed)
R's Self Ranking of Social Position	Equal Variances Assumed	.118	.731	4.384	915	.000
	Equal Variances Not Assumed			4.613	544.860	.000

Note: This box has been edited for clarity and some output has been omitted.

Skip over the first columns of the output block (which report the results of a test for equality of the population variances). In the top row, SPSS reports a t value (4.384), the degrees of freedom ($df = 915$), and a "Sig (2-tailed)" of .000. This last piece of information is an alpha level but it is the *exact* probability, rounded to three decimal places, of getting the observed difference in sample means if only chance is operating. Thus, there is no need to look up the test statistic ($t = 4.384$) in Appendix A or B. The value for alpha is less than 0.000, much less than our usual indicator of significance of 0.05. We conclude that the difference in *rank* is statistically significant. College graduates rank themselves significantly higher in social position than high school graduates.

Hypothesis Testing with Sample Proportions (Large Samples)

Testing for the significance of the difference between two sample proportions follows the pattern for testing sample means. The null hypothesis states that there is no difference between the populations from which the samples are drawn. We compute a test statistic in step 4, which is then compared with the critical region. When sample sizes are large (combined N's of more than 100), the Z distribution is used to find the critical region. In this text, we will not consider tests of significance for proportions based on small samples.

In order to find the value of the test statistic, several preliminary equations must be solved. Formula 8.7 uses the values of the two sample proportions (P_s) to give us an estimate of the population proportion (P_u), which is the proportion of cases in the population that have the trait under consideration assuming the null hypothesis is true.

FORMULA 8.7

$$P_u = \frac{N_1 P_{s1} + N_2 p_{s2}}{N_1 + N_2}$$

The value of P_u is then used to compute the standard deviation of the sampling distribution of the difference in sample proportions in Formula 8.8.

FORMULA 8.8

$$\sigma_{p-p} = \sqrt{P_u(1 - P_u)}\sqrt{\frac{N_1 + N_2}{N_1 N_2}}$$

This value is then substituted into the formula for computing the test statistic, presented as Formula 8.9.

FORMULA 8.9

$$Z(\text{obtained}) = \frac{(P_{s1} - P_{s2}) - (P_{u1} - P_{u2})}{\sigma_{p-p}}$$

where $(P_{s1} - P_{s2})$ = the difference between the sample proportions
$(P_{u1} - P_{u2})$ = the difference between the population proportions
σ_{p-p} = the standard deviation of the sampling distribution of the difference between sample proportions

As was the case with sample means, the second term in the numerator is assumed to be zero by the null hypothesis. Therefore, the formula reduces to

FORMULA 8.10

$$Z(\text{obtained}) = \frac{(P_{s1} - P_{s2})}{\sigma_{p-p}}$$

Remember to solve these equations in order, starting with Formula 8.7 (and skipping Formula 8.9).

Conducting a Test of Hypothesis with Sample Proportions (Large Samples)

An example will make these procedures clearer. Suppose we are researching social networks among senior citizens and wonder whether black people and white people differ in the number of memberships in clubs and other organizations they hold. Random samples of black and white senior citizens have been selected and classified as high or low in terms of their number of memberships. Is there a statistically significant difference in the participation patterns of black and white elderly? The proportions of each group classified as "high" in participation and sample size for both groups are

Black Senior Citizens (Sample 1)	White Senior Citizens (Sample 2)
$P_{s1} = 0.34$	$P_{s2} = 0.25$
$N_1 = 83$	$N_2 = 103$

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Independent random samples
Level of measurement is nominal
Sampling distribution is normal

Step 2. Stating the Null Hypothesis. Because no direction has been predicted, this will be a two-tailed test.

$$H_0: P_{u1} = P_{u2}$$
$$(H_1: P_{u1} \neq P_{u2})$$

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region. Because sample size is large, the Z distribution will be used to establish the critical region. Setting alpha at 0.05, we have

Sampling distribution = Z distribution
Alpha = 0.05, two-tailed
$Z(\text{critical}) = \pm 1.96$

Step 4. Computing the Test Statistic. Solve Formula 8.7 first, substitute the resultant value into Formula 8.8, and then solve for Z(obtained) with Formula 8.10.

$$P_u = \frac{N_1 P_{s1} + N_2 P_{s2}}{N_1 + N_2}$$

$$P_u = \frac{(83)(0.34) + (103)(0.25)}{83 + 103}$$

$$P_u = \frac{28.22 + 25.27}{186}$$

$$P_u = \frac{53.49}{186}$$

$$P_u = 0.29$$

$$\sigma_{p-p} = \sqrt{P_u(1 - P_u)}\sqrt{\frac{N_1 + N_2}{N_1N_2}}$$

$$\sigma_{p-p} = \sqrt{(0.29)(0.71)}\sqrt{\frac{83 + 103}{(83)(103)}}$$

$$\sigma_{p-p} = \sqrt{0.2059}\sqrt{\frac{186}{8549}}$$

$$\sigma_{p-p} = \sqrt{0.2059}\sqrt{0.0218}$$

$$\sigma_{p-p} = (0.45)(0.15)$$

$$\sigma_{p-p} = 0.07$$

$$Z(\text{obtained}) = \frac{(P_{s1} - P_{s2})}{\sigma_{p-p}}$$

$$Z(\text{obtained}) = \frac{0.34 - 0.25}{0.07}$$

$$Z(\text{obtained}) = \frac{0.09}{0.07}$$

$$Z(\text{obtained}) = 1.29$$

Step 5. Making a Decision and Interpreting the Results of the Test. The test statistic does not fall in the critical region:

$$Z(\text{critical}) = \pm 1.96$$
$$Z(\text{obtained}) = 1.29$$

We fail to reject the null hypothesis. The difference between the sample proportions is no greater than what would be expected if the null hypothesis were true and only random chance were operating. Black and white senior citizens are not significantly different in their participation in voluntary associations. *(For practice in testing the significance of the difference between sample proportions, see problems 8.10 to 8.14, 8.15a to c, and 8.16a to c.)*

ONE STEP AT A TIME **Completing Step 4 of the Five-Step Model: Computing *Z*(obtained)**

Solve Formulas 8.7, 8.8, and 8.10 to find the test statistic.

Step	***Operation***
To solve Formula 8.7:	
1.	Add N_1 and N_2.
2.	Multiply P_{s1} by N_1.
3.	Multiply P_{s2} by N_2.
4.	Add the quantity you found in step 3 to the quantity you found in step 2.
5.	Divide the quantity you found in step 4 by the quantity you found in step 1. This is P_u.

(*continued*)

ONE STEP AT A TIME *(continued)*

Step	*Operation*
To solve Formula 8.8:	
1.	Multiply P_u by $(1 - P_u)$.
2.	Take the square root of the quantity you found in step 1.
3.	Multiply N_1 by N_2.
4.	Add N_1 and N_2. (*Note: You already found this value when solving Formula 8.7. See step 1*)
5.	Divide the quantity you found in step 4 by the quantity you found in step 3.
6.	Take the square root of the quantity you found in step 5.
7.	Multiply the quantity you found in step 6 by the quantity you found in step 2.
To solve Formula 8.10:	
1.	Subtract P_{s2} from P_{s1}.
2.	Divide the quantity you found in step 1 by the quantity you found in step 7 above. This value is Z(obtained).

ONE STEP AT A TIME **Completing Step 5 of the Five-Step Model: Making a Decision and Interpreting Results**

Step	*Operation*
1.	Compare Z(obtained) to Z(critical). If Z(obtained) is *in* the critical region, *reject* the null hypothesis. If Z(obtained) is *not in* the critical region, *fail to reject* the null hypothesis.
2.	Interpret the decision to reject or fail to reject the null hypothesis in terms of the original question. For example, our conclusion for the example problem was "There is no significant difference between black and white senior citizens in their level of participation in voluntary associations."

Applying Statistics 8.2 Testing the Significance of the Difference Between Sample Proportions

Do attitudes toward premarital sex vary by gender? Based on a representative sample of the adult U.S. population, the proportion of each sex that feels that premarital sex is "always wrong" is

Males (Sample 1)	Females (Sample 2)
$P_{s1} = 0.18$	$P_{s2} = 0.24$
$N_1 = 597$	$N_1 = 807$

Females are more likely to say that premarital sex is always wrong, but is the difference significant? We will conduct a test of the null hypothesis following the familiar five-step model with alpha set at 0.05, two-tailed test.

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Independent random samples
Level of measurement is nominal
Sampling distribution is normal

(*continued*)

Step 2. Stating the Null Hypothesis.

$$H_0: P_{u1} = P_{u2}$$
$$H_1: P_{u1} \neq P_{u2}$$

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region.

Sampling distribution = Z distribution

Alpha = 0.05, two-tailed

Z(critical) = ±1.96

Step 4. Computing the Test Statistic. Start with Formula 8.7, substitute the value for P_u into Formula 8.8, and then substitute that value into Formula 8.10 to solve for Z(obtained).

$$P_u = \frac{N_1 P_{s1} + N_2 P_{s2}}{N_1 + N_2}$$
$$P_u = \frac{(597)(0.18) + (807)(0.24)}{597 + 807}$$
$$P_u = \frac{107.46 + 193.68}{1404}$$
$$P_u = \frac{301.14}{1404}$$
$$P_u = 0.21$$

$$\sigma_{p-p} = \sqrt{P_u(1 - P_u)}\sqrt{\frac{N_1 + N_2}{N_1 N_2}}$$
$$\sigma_{p-p} = \sqrt{(0.21)(0.79)}\sqrt{\frac{597 + 807}{(597)(807)}}$$
$$\sigma_{p-p} = \sqrt{0.1659}\sqrt{\frac{1404}{481{,}779}}$$
$$\sigma_{p-p} = (0.41)\sqrt{0.0029}$$
$$\sigma_{p-p} = (0.41)\,(0.05)$$
$$\sigma_{p-p} = 0.02$$

$$Z(\text{obtained}) = \frac{(P_{s1} - P_{s2})}{\sigma_{p-p}}$$
$$Z(\text{obtained}) = \frac{(0.18 - 0.24)}{0.02}$$
$$Z(\text{obtained}) = \frac{-0.06}{0.02}$$
$$Z(\text{obtained}) = -3.00$$

Step 5. Making a Decision and Interpreting the Results of the Test. With an obtained Z score of −3.00, we reject the null hypothesis. There is a statistically significant difference between the proportion of females and males who feel that premarital sex is always wrong.

STATISTICS IN EVERYDAY LIFE

Abortion and Polarization in American Politics

According to Gallup polls, Republicans and Democrats have become increasingly polarized over the abortion issue. In 1975, one year after the landmark *Roe v. Wade* Supreme Court decision, 18% of Republicans and 19% of Democrats agreed that abortion should be legal “under any circumstances.” In 2013, the percentage of Republicans approving abortion under any circumstances had fallen to 13% while the percentage of Democrats approving had risen to 39%. The 1 percentage point difference in 1975 might have been *statistically* significant but it was clearly not important. The 26% point difference in 2013 is statistically significant, important, and extremely consequential for the possibility of civil discourse in American politics. These results are based on random samples of about 1000 adult Americans.

How might we explain these trends? One possibility is that the Republican Party has lost many moderate, pro-choice members since the 1970s, leaving the party smaller and more ideologically homogenous. What information would you need to investigate this possibility?

Source: Saad, Lydia. 2013. “American’s Abortion Views Steady Amid Gosnell Trial.” Available at http://www.gallup.com/poll/162374/americans-abortion-views-steady-amid-gosnell-trial.aspx

The Limitations of Hypothesis Testing: Significance Versus Importance

We are usually interested in rejecting the null hypothesis and should take a moment to consider systematically the factors that affect our decision in step 5. Generally speaking, the probability of rejecting the null hypothesis is a function of four independent factors, only the first of which is not under the direct control of the researcher:

1. The size of the difference between the sample statistics. The larger the difference between the means or proportions, the greater the likelihood that the null hypothesis will be rejected.
2. The alpha level. As the alpha level increases, the critical region grows larger, and the probability of rejecting the null hypothesis increases. Thus, it is easier to reject H_0 at the 0.05 level than at the 0.01 level, and easier still at the 0.10 level. The danger here, of course, is that higher alpha levels increase the chance of Type I errors, and we might declare small differences to be statistically significant.
3. One-tailed vs. two-tailed test. A one-tailed test will increase the probability of rejecting the null hypothesis (assuming that the proper direction has been predicted).
4. The size of the sample. With all other factors constant, the probability of rejecting H_0 increases with sample size: The larger the sample, the more likely we are to reject the null hypothesis. With very large samples, we may declare small, unimportant differences to be statistically significant.

The relationship between sample size and the probability of rejecting the null hypothesis may be surprising and we will consider it in more detail. The probability of rejecting H_0 increases with sample size because sample size (N) is in the "denominator of the denominator" in all of the formulas for computing the test statistic (step 4). Algebraically, this is equivalent to being in the numerator of the formula and means that the value of the test statistic is directly proportional to N and that the two will increase together.

To illustrate, consider Table 8.1, which shows the value of the test statistic for single-sample means from samples of various sizes. The value of the test statistic, Z(obtained), increases as N increases, even though none of the other terms in the formula change.

With a small sample ($N = 50$), the test statistic of 1.41 is not significant at the 0.05 level. However, doubling the sample size to 100 produces a significant result, even though none of the other values in the equation change. The test statistic—Z(obtained)—continues to increase as sample size increases. This pattern holds for all tests of significance: The larger the sample, the greater the value of the test statistic and the higher the probability of rejecting the null hypothesis.

TABLE 8.1 Test Statistics for Single-Sample Means Computed from Samples of Various Sizes ($\bar{X} = 80$, $\mu = 79$, $s = 5$ Throughout)

Sample Size (N)	Computing the Test Statistic	Test Statistic, Z(obtained)
50	$Z(\text{obtained}) = \frac{\bar{X} - \mu}{s/\sqrt{N-1}} = \frac{80 - 79}{5/\sqrt{49}} = \frac{1}{0.71} =$	1.41
100	$Z(\text{obtained}) = \frac{\bar{X} - \mu}{s/\sqrt{N-1}} = \frac{80 - 79}{5/\sqrt{99}} = \frac{1}{0.50} =$	2.00
500	$Z(\text{obtained}) = \frac{\bar{X} - \mu}{s/\sqrt{N-1}} = \frac{80 - 79}{5/\sqrt{499}} = \frac{1}{0.22} =$	4.55
1000	$Z(\text{obtained}) = \frac{\bar{X} - \mu}{s/\sqrt{N-1}} = \frac{80 - 79}{5/\sqrt{999}} = \frac{1}{0.16} =$	6.25
10,000	$Z(\text{obtained}) = \frac{\bar{X} - \mu}{s/\sqrt{N-1}} = \frac{80 - 79}{5/\sqrt{9{,}999}} = \frac{1}{0.05} =$	20.00

On one hand, the relationship between sample size and the probability of rejecting the null hypothesis should not alarm us unduly. Larger samples are, after all, better approximations of the populations they represent. Thus, decisions based on larger samples can be trusted more than decisions based on smaller samples.

On the other hand, this relationship clearly underlines what is perhaps the most important limitation of hypothesis testing: *Simply because a difference is statistically significant does not guarantee that it is important in any other sense*. Particularly with very large samples, relatively small differences may be statistically significant. Even with small samples, of course, differences that are otherwise trivial or uninteresting may be statistically significant.

The crucial point is that statistical significance and theoretical or practical importance can be two very different things. Statistical significance is a necessary but not sufficient condition for theoretical or practical importance. A difference that is not statistically significant is almost certainly unimportant. However, significance by itself does not guarantee importance. Even when it is clear that the research results were not produced by random chance, the researcher must still assess their importance. Do they firmly support a theory or hypothesis? Are they clearly consistent with a prediction or analysis? Do they strongly indicate a line of action in solving some problem? These are the kinds of questions a researcher must ask when assessing the importance of the results of a statistical test.

Also, we should note that researchers have access to some very powerful ways of analyzing the importance (vs. the statistical significance) of research results. These statistics, which include bivariate measures of association and multivariate statistical techniques, are introduced in Parts III and IV of this text.

STATISTICS
IN EVERYDAY LIFE

Is the Gender Gap in Income Decreasing?

It is well known that there is a gap between the average incomes of men and women. Is this gender gap closing? Comparing *only* full-time workers, the median income for women in1955 was 64% of the median income for men. The gender gap actually grew larger until the early 1970s but has generally (but not always) shrunk since that time and, as of 2012, it had closed to 79%. In other words, comparing only full-time, year-round workers, women earned 79 cents for every dollar earned by men.

Note that it may not be possible to close this gap much further as long as there is a "glass ceiling" on women's chances for promotion to the highest (most lucrative) positions and, therefore, relatively few women in the highest income brackets. For example, in 2012, 17% of men earned more than $100,000 versus only 8% of women.

What does the glass ceiling imply about the overall shape of the distribution of men's and women's income? Which distribution will be more positively skewed? What effect will that have on the mean? (See Chapter 3 for a review of the effect of skew on measures of central tendency.)

Source: U.S. Bureau of the Census. Table P-36 available at http://www.census.gov/search-results.html?q=table+p-36&search.x=0&search.y=0&search=submit&page=1&stateGeo=none&utf8=%26%2310003%3B&affiliate=census

BECOMING A CRITICAL CONSUMER: When is a difference a difference?

How big does a difference have to be to be in order to be considered a difference? The question may sound silly, but this is a serious issue because it relates to our ability to identify the truth when we see it. Using the income gap between the genders as an example, how big a difference must there be in average income before we conclude that there is gender inequality? Would we be concerned if U.S. men averaged $50,000 and women averaged $49,500? Across millions of cases, a difference of $500—about 1% of the average incomes—seems small and unimportant. How about if the difference was $1,000? $5,000? $10,000? At what point do we declare the difference to be important?

Very large or very small differences are easy to deal with. But what about differences between these two extremes? How big is big?

There are, of course, no absolute rules that would always enable us to identify important differences. However, we can discuss some guidelines that will help us know when a difference is consequential. We'll do this in general terms first and then relate this discussion to significance testing.

Differences in General

We can use three guidelines to help identify important differences:

1. As I suggested above, it is helpful to think about the difference in terms of the scale of the variable. A drop of a cent in the cost of a gallon of gas when the average is $4.00 probably won't make a difference for most people. In very general (and arbitrary) terms, a change of 5% or 10% or more (a rise or fall of 20 or 40 cents if gas costs $4.00 a gallon) would signal an important difference for many social indicators—population growth, crime rates, and birthrates, for example.
2. It will often be helpful to look at the raw frequencies to judge the importance of a change. Most people would be alarmed by a headline that reported that the number of teen pregnancies had doubled. However, consider two towns, both with a population of 250,000. In the first, the number of teen pregnancies doubled from 10 to 20 in a year and, in the second, the number doubled from 2500 to 5000. The latter case is clearly more alarming than the first, even though, in both cases, the numbers doubled. The raw frequencies can add a valuable context to the perception of a change (which is one reason they are always reported in the professional research literature).
3. Another way to add some context to a change is to look at a broader time period. A report that voter turnout declined by 20% in a state between 2012 and 2014 might cause some alarm. However, turnout often declines between years featuring presidential elections (2012) and those that do not (2014). It would be more meaningful to compare 2014 with 2010 and even more revealing to get the data for earlier years.

(*continued*)

Differences in Social Science Research

In social research based on random samples, the problem of identifying important differences is complicated by the vagaries of chance. That is, the size of a difference between sample statistics may be the result of random chance rather than (or in addition to) actual differences in the population.

One of the great strengths of hypothesis testing is that it provides a system for identifying important differences. When we say that a difference is statistically significant, we reject the argument that random chance alone is responsible (with a known probability of error—the alpha level) and support the idea that the difference in the sample statistics reflects a difference in the population. Small differences (e.g., a difference of only a few hundred dollars in average income between the genders) are unlikely to be significant at the 0.05 level. The larger the difference between the sample statistics, the more likely it is to be significant. The larger the difference and the lower the alpha level, the more confidence we can have that the difference reflects actual patterns in the population.

Obviously, this form of decision making is not infallible. Remember that there is always a chance of making an incorrect decision: we could declare a trivial difference to be important (a Type I error) and, vice versa, we could conclude that an important difference is trivial (a Type II error).

Reading Social Research

Professional researchers use a vocabulary that is much terser than ours. This is partly because of space limitations in scientific journals and partly because they can assume a certain level of statistical literacy in their audiences. Thus, they omit many of the elements—such as the null hypothesis or the critical region—that we have been so careful to state.

Instead, researchers report only the sample statistics (for example, means or proportions), the value of the test statistic (for example, a Z or t score), the alpha level, the degrees of freedom (if applicable), and the sample size. The results of one of our example problems might be reported as "the difference between the sample means of 2.37 (suburban families) and 2.78 (central-city families) was tested and found to be significant ($t = -2.16$, $df = 77$, $p < 0.05$)."

Note that the alpha level is reported as "$p < 0.05$." This is shorthand for "the probability of a difference of this size occurring by chance alone, if the null hypothesis of no difference is true, is less than 0.05" and is a good illustration of how researchers can convey a great deal of information in just a few symbols. In a similar fashion, our somewhat longwinded phrase "the test statistic falls in the critical region and, therefore, the null hypothesis is rejected" is rendered simply as: "the difference . . . was . . . found to be significant."

When researchers need to report the results of many tests of significance, they will often use a summary table. If you read the researcher's description and analysis of such tables, you should have little difficulty interpreting and understanding them. These comments about how significance tests are reported in the literature apply to all of the tests of hypotheses covered in Part II of this text.

SUMMARY

1. A common research situation is to test for the significance of the difference between two populations. Sample statistics are calculated for random samples of each population, and then we test for the significance of the difference between the samples as a way of inferring differences between the populations.

2. When sample information is summarized in the form of sample means and N is large, the Z distribution is used to find the critical region. When N is small, the t distribution is used to establish the critical region.

3. Differences in sample proportions may also be tested for significance. For large samples, the Z distribution is used to find the critical region.

4. In all tests of hypothesis, a number of factors, including sample size, affect the probability of rejecting the null hypothesis. Statistical significance is not the same thing as theoretical or practical importance. Even after a difference is found to be statistically significant, the researcher must still demonstrate the relevance or importance of his or her findings. The statistics presented in Parts III and IV of this text will give us the tools we need to deal directly with issues beyond statistical significance.

SUMMARY OF FORMULAS

FORMULA 8.1	Test statistic for two sample means, large samples:	$Z(\text{obtained}) = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}-\bar{X}}}$
FORMULA 8.2	Test statistic for two sample means, large samples (simplified):	$Z(\text{obtained}) = \dfrac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{X}-\bar{X}}}$
FORMULA 8.3	Standard deviation of the sampling distribution of the difference in sample means, large samples:	$\sigma_{\bar{X}-\bar{X}} = \sqrt{\dfrac{\sigma_1^2}{N_1} - \dfrac{\sigma_2^2}{N_2}}$
FORMULA 8.4	Pooled estimate of the standard deviation of the sampling distribution of the difference in sample means, large samples:	$\sigma_{\bar{X}-\bar{X}} = \sqrt{\dfrac{s_1^2}{N_1 - 1} + \dfrac{s_2^2}{N_2 - 1}}$
FORMULA 8.5	Pooled estimate of the standard deviation of the sampling distribution of the difference in sample means, small samples:	$\sigma_{\bar{X}-\bar{X}} = \sqrt{\dfrac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}} \sqrt{\dfrac{N_1 + N_2}{N_1 N_2}}$
FORMULA 8.6	Test statistic for two sample means, small samples:	$t(\text{obtained}) = \dfrac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{X}-\bar{X}}}$
FORMULA 8.7	Pooled estimate of population proportion, large samples:	$P_u = \dfrac{N_1 P_{s1} + N_2 P_{s2}}{N_1 + N_2}$
FORMULA 8.8	Standard deviation of the sampling distribution of the difference in sample proportions, large samples:	$\sigma_{p-p} = \sqrt{P_u(1 - P_u)} \sqrt{\dfrac{N_1 + N_2}{N_1 N_2}}$
FORMULA 8.9	Test statistic for two sample proportions, large samples:	$Z(\text{obtained}) = \dfrac{(P_{s1} - P_{s2}) - (P_{u1} - P_{u2})}{\sigma_{p-p}}$
FORMULA 8.10	Test statistic for two sample proportions, large samples (simplified):	$Z(\text{obtained}) = \dfrac{(P_{s1} - P_{s2})}{\sigma_{p-p}}$

GLOSSARY

Independent random samples. Random samples gathered in such a way that the selection of a particular case for one sample has no effect on the probability that any other particular case will be selected for the other samples.

Pooled estimate. An estimate of the standard deviation of the sampling distribution of the difference in sample means based on the standard deviations of both samples.

σ_{p-p}. Symbol for the standard deviation of the sampling distribution of the differences in sample proportions.

$\sigma_{\bar{X}-\bar{X}}$. Symbol for the standard deviation of the sampling distribution of the differences in sample means.

PROBLEMS

8.1 For each of the following, test for the significance of the difference in sample statistics using the five-step model. *(HINT: Remember to solve Formula 8.4 before attempting to solve Formula 8.2. Also, in Formula 8.4, perform the mathematical operations in the proper sequence. First square each sample standard deviation, then divide by the proper N, add the resultant values, and then find the square root of the sum.)*

a.

Sample 1	Sample 2
$\bar{X}_1 = 72.5$	$\bar{X}_2 = 76.0$
$s_1 = 14.3$	$s_2 = 10.2$
$N_1 = 136$	$N_2 = 257$

b.

Sample 1	Sample 2
$\bar{X}_1 = 107$	$\bar{X}_2 = 103$
$s_1 = 14$	$s_2 = 17$
$N_1 = 175$	$N_2 = 200$

8.2 SOC Two sociologists administered questionnaires to samples of undergraduates that measured attitudes toward interpersonal violence (higher scores indicate greater approval of interpersonal violence). Test the results as reported here for gender, racial, and social-class differences.

a.

Sample 1 (Males)	Sample 2 (Female)
$\bar{X}_1 = 2.99$	$\bar{X}_2 = 2.29$
$s_1 = 0.88$	$s_2 = 0.91$
$N_1 = 122$	$N_2 = 251$

b.

Sample 1 (Blacks)	Sample 2 (Whites)
$\bar{X}_1 = 2.76$	$\bar{X}_2 = 2.49$
$s_1 = 0.68$	$s_2 = 0.91$
$N_1 = 43$	$N_2 = 304$

c.

Sample 1 (White Collar)	Sample 2 (Blue Collar)
$\bar{X}_1 = 2.46$	$\bar{X}_2 = 2.67$
$s_1 = 0.91$	$s_2 = 0.87$
$N_1 = 249$	$N_2 = 97$

Summarize your results in terms of the significance and the direction of the differences. Which of these three factors seems to make the biggest difference in attitudes toward interpersonal violence?

8.3 SOC Do athletes in different sports vary in terms of their readiness for college? Reported here are college entrance exam scores of fictitious random samples of college basketball and football players and male and female athletes. Is there a significant difference between groups? Write a sentence or two explaining the results.

a.

Sample 1 (Basketball Players)	Sample 2 (Basketball Players)
$\bar{X}_1 = 460$	$\bar{X}_2 = 442$
$s_1 = 92$	$s_2 = 57$
$N_1 = 102$	$N_2 = 117$

b.

Sample 1 (Males)	Sample 2 (Females)
$\bar{X}_1 = 452$	$\bar{X}_2 = 480$
$s_1 = 88$	$s_2 = 75$
$N_1 = 107$	$N_2 = 105$

8.4 PA A number of years ago, the road and highway maintenance department of Sin City, Nevada, began recruiting minority group members through an affirmative action program. In terms of efficiency ratings, as compiled by their superiors, how do the affirmative action employees rate? The ratings of random samples of both groups were collected, and the results are reported here (higher ratings indicate greater efficiency). Write a few sentences reporting and interpreting the results of the test.

a.

Sample 1 (Affirmative Action)	Sample 2 (Regular)
$\bar{X}_1 = 15.2$	$\bar{X}_2 = 15.5$
$s_1 = 3.9$	$s_2 = 2.0$
$N_1 = 97$	$N_2 = 100$

8.5 SOC Are middle-class families more likely than working-class families to maintain contact with kin? Write a paragraph summarizing the results of these tests.

a. A sample of middle-class families reported an average of 7.3 visits per year with close kin, while a sample of working-class families averaged 8.2 visits. Is the difference significant?

Sample 1 (Middle Class)	Sample 2 (Working Class)
$\bar{X}_1 = 7.3$	$\bar{X}_2 = 8.2$
$s_1 = 0.3$	$s_2 = 0.5$
$N_1 = 89$	$N_2 = 55$

b. The middle-class families averaged 2.3 phone calls and 8.7 e-mail messages per month with close kin. The working-class families averaged 2.7 calls and 5.7 e-mail messages per month. Are these differences significant?

Phone Calls per Month	
Sample 1 (Middle Class)	Sample 2 (Working Class)
$\bar{X}_1 = 2.3$	$\bar{X}_2 = 2.7$
$s_1 = 0.5$	$s_2 = 0.8$
$N_1 = 89$	$N_2 = 55$

E-Mail Messages per Month	
Sample 1 (Middle Class)	Sample 2 (Working Class)
$\bar{X}_1 = 8.7$	$\bar{X}_2 = 5.7$
$s_1 = 0.3$	$s_2 = 1.1$
$N_1 = 89$	$N_2 = 55$

8.6 **SOC** Are college students who live in dormitories significantly more involved in campus life than students who commute to campus? The following data report the average number of hours per week students devote to extracurricular activities. Is the difference between these randomly selected samples of commuter and residential students significant?

Sample 1 (Residential)	Sample 2 (Commuter)
$\bar{X}_1 = 12.4$	$\bar{X}_2 = 10.2$
$s_1 = 2.0$	$s_2 = 1.9$
$N_1 = 158$	$N_2 = 173$

8.7 **GER** Are senior citizens who live in retirement communities more socially active than those who live in age-integrated communities? Write a sentence or two explaining the results of these tests. *(HINT: Remember to use the proper formulas for small sample sizes.)*

a. A random sample of senior citizens living in a retirement village reported that they had an average of 1.42 face-to-face interactions per day with their neighbors. A random sample of those living in age-integrated communities reported 1.58 interactions. Is the difference significant?

Sample 1 (Retirement Community)	Sample 2 (Age-Integrated Neighborhood)
$\bar{X}_1 = 1.42$	$\bar{X}_2 = 1.58$
$s_1 = 0.10$	$s_2 = 0.78$
$N_1 = 43$	$N_2 = 37$

b. Senior citizens living in the retirement village reported that they had an average of 7.43 telephone calls with friends and relatives each week, while those in the age-integrated communities reported an average of 5.50 calls. Is the difference significant?

Sample 1 (Retirement Community)	Sample 2 (Age-Integrated Neighborhood)
$\bar{X}_1 = 7.43$	$\bar{X}_2 = 5.50$
$s_1 = 0.75$	$s_2 = 0.25$
$N_1 = 43$	$N_2 = 37$

8.8 **SW** As the director of the local Boys Club, you have claimed for years that membership in your club reduces juvenile delinquency. Now a cynical member of your funding agency has demanded proof of your claim. Random samples of members and nonmembers are gathered and interviewed with respect to their involvement in delinquent activities. Each respondent is asked to enumerate the number of delinquent acts he has engaged in over the past year. The average numbers of admitted acts of delinquency are reported below. What can you tell the funding agency?

Sample 1 (Members)	Sample 2 (Nonmembers)
$\bar{X}_1 = 10.3$	$\bar{X}_2 = 12.3$
$s_1 = 2.7$	$s_2 = 4.2$
$N_1 = 40$	$N_2 = 55$

8.9 SOC A survey has been administered to random samples of respondents in each of five nations. Respondents were asked: "How satisfied are you with your life as a whole?" Responses varied from 1 (very dissatisfied) to 10 (very satisfied). Conduct a test for the significance of the difference in mean scores for each nation between men and women.

Brazil	
Males	Females
$\bar{X}_1 = 7.7$	$\bar{X}_2 = 7.6$
$s_1 = 2.0$	$s_2 = 2.2$
$N_1 = 712$	$N_2 = 783$

Ukraine	
Males	Females
$\bar{X}_1 = 5.8$	$\bar{X}_2 = 5.5$
$s_1 = 2.3$	$s_2 = 2.4$
$N_1 = 446$	$N_2 = 549$

Vietnam	
Males	Females
$\bar{X}_1 = 7.2$	$\bar{X}_2 = 7.0$
$s_1 = 1.9$	$s_2 = 1.8$
$N_1 = 762$	$N_2 = 720$

South Africa	
Males	Females
$\bar{X}_1 = 6.8$	$\bar{X}_2 = 7.2$
$s_1 = 2.5$	$s_2 = 2.3$
$N_1 = 1492$	$N_2 = 1482$

Egypt	
Males	Females
$\bar{X}_1 = 5.6$	$\bar{X}_2 = 5.9$
$s_1 = 2.7$	$s_2 = 2.7$
$N_1 = 1557$	$N_2 = 1493$

8.10 For each problem, test the sample proportions for the significance of the difference.

a.

Sample 1	Sample 2
$P_{s1} = 0.20$	$P_{s2} = 0.17$
$N_1 = 114$	$N_2 = 101$

b.

Sample 1	Sample 2
$P_{s1} = 0.60$	$P_{s2} = 0.62$
$N_1 = 478$	$N_2 = 532$

8.11 CJ About half of the police officers in Shinbone, Kansas, have completed a special course in investigative procedures. Has the course increased their efficiency in clearing crimes by arrest? The proportions of cases cleared by arrest for samples of trained and untrained officers are reported here.

Sample 1 (Trained)	Sample 2 (Untrained)
$P_{s1} = 0.47$	$P_{s2} = 0.43$
$N_1 = 157$	$N_2 = 113$

8.12 SW A large counseling center needs to evaluate several experimental programs. Write a paragraph summarizing the results of these tests. Did the new programs work?

a. One program is designed for divorce counseling; the key feature of the program is its counselors, who are married couples working in teams. About half of all clients have been randomly assigned to this special program and half to the regular program, and the proportion of cases that eventually ended in divorce was recorded for both. The results for random samples of couples from both programs are reported here. In terms of preventing divorce, did the new program work?

Sample 1 (Special Program)	Sample 2 (Regular Program)
$P_{s1} = 0.53$	$P_{s2} = 0.59$
$N_1 = 78$	$N_2 = 82$

b. The agency is also experimenting with a program that includes peer counseling for depressed teenagers. About half of all clients were randomly assigned to the new program. After a year, a random sample of teens from the new program was compared with a random sample that received standard counseling. In terms of the percentage of children who were judged to be "much improved," did the new program work?

Sample 1 (Peer Counseling)	Sample 2 (Standard Program)
$P_{s1} = 0.10$	$P_{s2} = 0.15$
$N_1 = 52$	$N_2 = 56$

8.13 SOC At St. Algebra College, the sociology and psychology departments have been feuding for years about the respective quality of their programs. In an attempt to resolve the dispute, you have gathered data about the graduate school experience of random samples of both groups of majors. As measured by these data, is there a significant difference in program quality?

a. Proportion of majors who applied to graduate school:

Sample 1 (Seciology)	Sample 2 (Psychology)
$P_{s1} = 0.53$	$P_{s2} = 0.40$
$N_1 = 150$	$N_2 = 175$

b. Proportion accepted by program of first choice:

Sample 1 (Seciology)	Sample 2 (Psychology)
$P_{s1} = 0.75$	$P_{s2} = 0.86$
$N_1 = 80$	$N_2 = 70$

c. Proportion completing the program:

Sample 1 (Seciology)	Sample 2 (Psychology)
$P_{s1} = 0.75$	$P_{s2} = 0.70$
$N_1 = 60$	$N_2 = 60$

8.14 CJ The mayor of a large city started a "crimeline" program some years ago and wonders whether it is working. The program publicizes unsolved violent crimes in the local media and offers cash rewards for information leading to arrests. Are "featured" crimes more likely to be cleared by arrest than other violent crimes? Results from random samples of both types of crimes are reported as follows:

Sample 1 (Crimeline Crimes Cleared by Arrest)	Sample 2 (Non-Crimeline Crimes Cleared by Arrest)
$P_{s1} = 0.35$	$P_{s2} = 0.25$
$N_1 = 178$	$N_2 = 212$

8.15 SOC Some results from a survey administered to a nationally representative sample are reported here in terms of differences by gender. Which of these differences, if any, are significant? Write a sentence or two of interpretation for each test.

a. Proportion favoring the legalization of marijuana:

Males	Females
$P_{s1} = 0.37$	$P_{s2} = 0.31$
$N_1 = 202$	$N_2 = 246$

b. Proportion strongly agreeing that "kids are life's greatest joy":

Males	Females
$P_{s1} = 0.47$	$P_{s2} = 0.58$
$N_1 = 251$	$N_2 = 351$

c. Proportion voting for President Obama in 2012:

Males	Females
$P_{s1} = 0.45$	$P_{s2} = 0.53$
$N_1 = 399$	$N_2 = 509$

d. Average hours spent with e-mail each week:

Males	Females
$\bar{X}_1 = 4.18$	$\bar{X}_2 = 3.38$
$s_1 = 7.21$	$s_2 = 5.92$
$N_1 = 431$	$N_2 = 535$

e. Average rate of church attendance (number of times per year):

Males	Females
$\bar{X}_1 = 3.19$	$\bar{X}_2 = 3.99$
$s_1 = 2.60$	$s_2 = 2.72$
$N_1 = 641$	$N_2 = 808$

f. Number of children:

Males	Females
$\bar{X}_1 = 1.49$	$\bar{X}_2 = 1.93$
$s_1 = 1.50$	$s_2 = 1.50$
$N_1 = 635$	$N_2 = 803$

8.16 SOC A researcher is studying the effects of the college experience on attitudes, values, and behaviors and is comparing random samples of freshmen and seniors at the same university. Which of the following differences are significant?

a. Proportion calling parents at least once a day:

Freshmen	Seniors
$P_{s1} = 0.24$	$P_{s2} = 0.08$
$N_1 = 153$	$N_2 = 117$

b. Proportion with a "close friend" from another racial or ethnic group:

Freshmen	Seniors
$P_{s1} = 0.15$	$P_{s2} = 0.26$
$N_1 = 155$	$N_2 = 114$

c. Proportion saying that they are "very interested" in national politics:

Freshmen	Seniors
$P_{s1} = 0.46$	$P_{s2} = 0.48$
$N_1 = 147$	$N_2 = 111$

d. Average score on a scale that measures political ideology. On this scale, 10 means "very conservative" and 1 means "very liberal."

Freshmen	Seniors
$\bar{X}_1 = 5.23$	$\bar{X}_2 = 5.12$
$s_1 = 1.78$	$s_2 = 1.07$
$N_1 = 145$	$N_2 = 105$

e. Average score on a scale that measures support for traditional gender roles. This scale ranges from 7 ("Very traditional") to 1 ("Very non-traditional")

Freshmen	Seniors
$\bar{X}_1 = 2.07$	$\bar{X}_2 = 2.17$
$s_1 = 1.23$	$s_2 = 0.78$
$N_1 = 143$	$N_2 = 111$

Statistical Analysis Using SPSS

8.17 *SOC* In this exercise, you will use SPSS and the 2012 General Social Survey (GSS) to compare Protestants and Catholics for significant differences in their average social class standing, using level of education (*educ*) and self-rated social position (*rank*) as measures of social class. You will use the **Independent-Samples T Test** to conduct the tests.

- Click the SPSS icon on your desktop.
- Load the *GSS2012* database.
- Click **Analyze** → **Compare Means** → **Independent-Samples T Test**.
- Find *educ* and *rank* in the list of variables and move them into the "Test Variable(s)" box.
- Find *relig* in the list of variables and move it into the "Grouping Variable" box.
- Click **Define Groups** and type **1** (for Protestants) in the box for Group 1 and **2** (for Catholics) in the box for Group 2. Click **Continue**.
- Click **OK** and the results of the test will be printed to the SPSS output window.
- Find the group means, the *t* score, degrees of freedom, and the "Sig. (2-tailed)" (which is the exact probability of getting this difference if H_0 is true), and report the results of the test in a few sentences.

8.18 *SOC* Repeat the test you conducted in Problem 8.17 but, this time, compare Protestants with people with no religious affiliation ("Nones"). Repeat all the steps in the previous problem but, at the **Define Groups** command, change the second group from "2" (Catholics) to "4" (Nones). Finish all other steps as in the previous problem and report the results of the test in a few sentences.

YOU ARE THE RESEARCHER

Gender Gaps and Support for Traditional Gender Roles

Two projects follow. The first tests for significant differences in sample means between men and women on four variables of your own choosing. The second uses the **Compute** command to explore attitudes toward abortion or traditional gender roles. You are urged to complete both projects. Follow the instructions in the "Using SPSS" demonstration presented earlier in this chapter to produce the output.

Project 1: Exploring the Gender Gap

In this enlightened age, with its heavy stress on gender equality, how many important differences persist between the sexes? In this project, you will select four dependent variables and test whether the genders are significantly different on the variables you select.

Step 1: Choosing Dependent Variables

Select four dependent variables from the *GSS2012* dataset. Choose *only* interval-ratio variables or ordinal variables with three or more scores or categories. As you select variables, you might keep in mind the issues at the forefront of the debate over gender equality: income, education, and other measures of equality. Or you might choose variables that relate to lifestyle choices and patterns of everyday life: religiosity, TV viewing habits, or political ideas.

List your four dependent variables in the table below.

Variable	SPSS Name	What Exactly Does This Variable Measure?
1		
2		
3		
4		

Step 2: Stating Hypotheses

For each dependent variable, state a hypothesis about the difference you expect to find. For example, you might hypothesize that men will be more liberal or women will be more educated. Of course, you can hypothesize that there will be no significant difference between the genders. You can base your hypotheses on your own experiences or on the information about gender differences that you have acquired in your courses or from other sources.

Hypotheses:

1. ______________________________
2. ______________________________
3. ______________________________
4. ______________________________

Step 3: Getting the Output

Load the *GSS2012* data set and use the **Independent-Samples T Test** to produce output. See the "Using SPSS" demonstration in this chapter for detailed instructions. Find the four dependent variables you selected in the list and click the top arrow in the middle of the window to move the variable names to the "Test Variable(s)" box.

Next, highlight *sex* and click the bottom arrow in the middle of the window to move *sex* to the "Grouping Variable" box. Click the **Define Groups** button and type **1** in the first box (for males) and **2** in the second box (for females). Click **Continue** to return to the "Independent-Samples T Test" window and click **OK**.

Step 4: Reading the Output

Remember that the first block of output ("Group Statistics") presents descriptive statistics, and the second block reports the test for significance. In the top row of the second block of output, find the *t* value, the degrees of freedom, and the "Sig.

(2-tailed)," which, as you recall, is the *exact* probability of getting the observed difference in sample means if only chance is operating.

Step 5: Recording Your Results

Record your results in the table below. Write the SPSS variable name in the first column and then record the descriptive statistics. Next, record the results of the test of significance, using the top row ("Equal Variance Assumed") of the second output block. Record the *t* score, the degrees of freedom (*df*), and whether the difference is significant at the 0.05 level. If the value of "Sig. (2-tailed)" is less than 0.05, write YES in this column. If the value of "Sig. (2-tailed)" is more than 0.05, write NO in the column.

Dependent Variables		$\bar{X}$	*s*	*N*	*t* score	*df*	Significant?
	Men						
	Women						
	Men						
	Women						
	Men						
	Women						
	Men						
	Women						

Step 6: Interpreting Your Results

Summarize your findings. For each dependent variable, write

1. At least one sentence summarizing the test, in which you identify the variable being tested, the sample means for each group, *N*, the *t* score, and the significance level. In the professional research literature, you might find the results reported as "For a sample of 1417 respondents, there was no significant difference between the average age of men (48.21) and the average age of women (48.12) ($t = 0.95$, $df = 1416$, $p > 0.05$)."
2. A sentence relating to your hypotheses. Were they supported? How?

Project 2: Using the Compute Command to Explore Gender Differences

In this project, you will use the **Compute** command, which was introduced in Chapter 4, to construct a summary scale for either support for legal abortion or support for traditional gender roles. Do these attitudes vary significantly by gender? You will also choose a second independent variable, other than gender, to test for significant differences.

Step 1: Creating Summary Scales

To refresh your memory, we used the **Compute** command in Chapter 4 to create a summary scale (*abscale*) for attitudes toward abortion by adding the scores on two items (*abany* and *abpoor*). Remember that, once created, a computed variable is added to the active file and can be used like any of the variables actually recorded in the file. If you did not save the data file with *abscale* included, you can quickly recreate the variable by following the instructions in Chapter 4.

As an alternative, you can also create a scale to measure support for traditional gender roles. The GSS data set supplied with this text includes two variables that measure gender attitudes. One of these (*fefam*) states, "It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family." There are four possible responses to this item, ranging from "Strongly agree" (1) to "Strongly disagree" (4). Note that the lowest score (1) is most supportive of traditional gender roles.

The other item (*fechld*) states, "A working mother doesn't hurt the children." This item also has four possible responses and, again, the lowest score—strongly disagree (1)—is the most consistent with support of traditional gender roles.

The scores of the two items vary in the same direction (for both, the lowest score indicates the strongest support for traditional gender roles), so we can create a summary scale by simply adding the two variables together. Follow the commands in Chapter 4 to add the scores of *fefam* and *fechld* and create the scale, which we will call *fescale*. The computed variable (*fescale*) will have a total of seven possible scores, with lower scores indicating more support for traditional gender roles and higher scores indicating less support.

Step 2: Stating Hypotheses

Choose either *abscale* or *fescale*, and state a hypothesis about what difference you expect to find between men and women. Which gender will be more supportive (have a lower average score on *abscale*) of legal abortion? Why? Will men or women be more supportive (have a lower average score on *fescale*) of traditional gender roles? Why?

Step 3: Getting and Interpreting the Output

Run the **Independent Samples T Test** as before with your chosen scale as the **Test Variable** and *sex* as the **Grouping Variable**. See the instructions for Project 1 above.

Step 4: Interpreting Your Results

Summarize your results as in Project 1, step 6. Was your hypothesis confirmed? How?

Step 5: Extending the Test by Selecting an Additional Independent Variable

What other independent variable besides gender might be related to attitudes toward abortion or traditional gender roles? Select another independent variable besides *sex*, and conduct an additional *t* test with either *abscale* or *fescale* as the dependent variable. Remember that the *t* test requires that the independent variable have *only* two categories. For variables with more than two categories (*relig* or *racecen1*, for example), you can meet this requirement by using the **Define Groups** button in the "Grouping Variables" box to select specific categories of a variable. You could, for example, compare Protestants and Catholics on *relig* by choosing scores of 1 (Protestants) and 2 (Catholics).

Step 6: Stating Hypotheses

State a hypothesis about what differences you expect to find between the categories of your independent variable. Which category will be more supportive of legal abortion or more supportive of traditional gender roles? Why?

Step 7: Getting and Interpreting the Output

Run the **Independent Samples T Test** as before with the scale you selected as the **Test Variable** and your independent variable as the **Grouping Variable**.

Step 8: Interpreting Your Results

Summarize your results as in Project 1, step 6. Was your hypothesis confirmed? How?

9 Hypothesis Testing III
The Analysis of Variance

LEARNING OBJECTIVES

By the end of this chapter, you will be able to:

1. Identify and cite examples of situations in which analysis of variance (ANOVA) is appropriate.
2. Explain the logic of hypothesis testing as applied to ANOVA.
3. Perform the ANOVA test, using the five-step model as a guide, and correctly interpret the results.
4. Define and explain the concepts of population variance, total sum of squares, sum of squares between, sum of squares within, and mean square estimates.
5. Explain the difference between the statistical significance and the importance of relationships between variables.
6. Use SPSS to conduct the analysis of variance test.

USING STATISTICS

This chapter presents statistical techniques used to compare differences between more than two random samples. If the differences are significant, we conclude that the populations from which the samples have been selected are different. This conclusion will have a known probability of error, which is often set at the 0.05 level. Examples of situations in which the analysis of variance is appropriate include:

1. A researcher is examining the differences in support for the death penalty between people of various religious affiliations. If there are significant differences between random samples of Protestants, Catholics, and other religions, she will conclude that the respective populations have different views on this issue.
2. Random samples of freshmen, sophomores, juniors, and seniors at a large university are compared for differences in their political views and religious values. If the researcher finds significant differences between the samples, he will conclude that the populations (the classes) differ on these dimensions.
3. Is bullying more of a problem in suburban, city, or rural schools? Researchers compare random samples of each type of school to see whether there are statistically significant differences.

In this chapter, we examine a very flexible and widely used test of significance called the **analysis of variance** (often abbreviated as **ANOVA**). This test is designed to be used with interval-ratio-level dependent variables and is a powerful tool for analyzing the most sophisticated and precise measurements you are likely to encounter.

It is perhaps easiest to think of ANOVA as an extension of the *t* test for the significance of the difference between two sample means, which was presented in Chapter 8. The *t* test is used in situations in which our independent variable has exactly two categories (e.g., Protestants and Catholics). The analysis of variance, on the other hand, is appropriate for independent variables with more than two categories (e.g., Protestants, Catholics, Jews, people with no religious affiliation, Muslims, Hindus, and so forth).

To illustrate, suppose we were interested in analyzing the relationship between support for capital punishment (the dependent variable) and age group (the independent variable). Do attitudes change as people age? Are older people more supportive of the death penalty than younger people?

Suppose that we administered a scale that measures support for capital punishment at the interval-ratio level to a randomly selected sample and have divided the respondents into four age groups: 18–29, 30–45, 46–64, and 65 and older. We have four categories of subjects, and we want to see whether opinion varies significantly by the category (age group) into which a person is classified.

The Logic of the Analysis of Variance

For ANOVA, the null hypothesis is that the populations from which the samples are drawn have the same mean score on the dependent variable. As applied to our problem, the null hypothesis could be phrased as "People from different age groups do not vary in their support for the death penalty" or, symbolically, as $\mu_1 = \mu_2 = \mu_3 = \mu_4$. (Note that this is an extended version of the null hypothesis for the two-sample *t* test.) The researcher will normally want to reject the null hypothesis and, in this case, show that support is related to age group.

If the null hypothesis of "no difference" in the populations is true, then any means calculated from randomly selected samples should be roughly equal in value. The average score for the 18- to 29-year-olds should be about the same as the average score for each of the other age groups. Note that the averages are unlikely to be exactly the same value even if the null hypothesis really is true, because we will always encounter some error or chance fluctuations in the measurement process. We are *not* asking: "Are there differences between the samples (or, in our example, the age groups)?" Rather, we are asking: "Are the differences between the samples large enough to reject the null hypothesis and justify the conclusion that the populations are different?"

Consider what kinds of outcomes we might encounter if we actually administered a "Support of Capital Punishment Scale" and organized the scores by age group. Of the infinite variety of possibilities, let's focus on the two extreme outcomes presented in Tables 9.1 and 9.2.

TABLE 9.1 Support for Capital Punishment by Age Group (Fictitious Data)

	18–29	30–45	46–64	65+
Mean =	10.3	11.0	10.1	9.9
Standard Deviation =	2.4	1.9	2.2	1.7

TABLE 9.2 Support for Capital Punishment by Age Group (Fictitious Data)

	18–29	30–45	46–64	65+
Mean =	10.0	13.0	16.0	22.0
Standard Deviation =	2.4	1.9	2.2	1.7

In the first set of hypothetical results (Table 9.1), we see that the means and standard deviations of the groups are quite similar. The average scores are about the same, and all four groups exhibit about the same dispersion. These results would be consistent with the null hypothesis of no difference between the populations. Neither the average score nor the dispersion of the scores varies in any important way by age group.

Now consider another set of fictitious results, as displayed in Table 9.2. Here we see substantial differences in average score, with the youngest age group showing the least support and the oldest group showing the most. Also, the standard deviations are low and similar from category to category, indicating that there is not much variation within the age groups. Table 9.2 shows marked differences *between* age groups combined with homogeneity (or low standard deviations) *within* age groups. In other words, the age groups are different from each other and the variation within each age group is low. These results would contradict the null hypothesis and support the notion that support for the death penalty does vary by age group.

ANOVA proceeds by making the kinds of comparisons outlined above. The test compares the amount of variation *between* categories (that is, between age groups) with the amount of variation *within* categories (within each age group). The greater the difference *between* categories (as measured by the means) relative to the differences *within* categories (as measured by the standard deviations), the more likely that the null hypothesis of "no difference" can be rejected. If support for capital punishment actually does vary by age group, then the sample mean for each age group should be quite different from the others and dispersion within the categories should be relatively low.

The Computation of ANOVA

Even though we have been thinking of ANOVA as a test for the significance of the difference between sample means, the computational routine actually involves developing two separate estimates of the population variance, σ^2 (hence the name *analysis of variance*). Recall from Chapter 4 that the variance is the standard deviation squared. One estimate of the population variance is based on the amount of variation *within* each of the categories of the independent variable, and the other is based on the amount of variation *between* the categories.

Before constructing these estimates, we need to introduce some new concepts and statistics. The first new concept is the total variation of the scores, which is measured by a quantity called the **total sum of squares**, or ***SST***:

FORMULA 9.1

$$SST = \Sigma X_i^2 - N\bar{X}^2$$

To solve this formula, first find the sum of the squared scores (in other words, square each score and then add up the squared scores). Next, square the mean of all scores, multiply that value by the total number of cases in the sample (N), and subtract that quantity from the sum of the squared scores.

Formula 9.1 may seem vaguely familiar. A similar expression, $\Sigma(X_i - \bar{X})^2$ appears in the formula for the standard deviation and variance (see Chapter 4). All three statistics incorporate information about the variation of the scores (or, in the case of *SST*, the squared scores) around the mean (or, in the case of *SST*, the square of the mean multiplied by N). In other words, all three statistics are measures of the variation, or dispersion, of the scores.

To construct the two separate estimates of the population variance, we divide the total variation (*SST*) into two components. One component reflects the pattern of variation *within* each of the categories and is called the **sum of squares within (*SSW*)**. In our example problem, *SSW* would measure the amount of variation in support for the death penalty within each of the age groups.

The other component is based on the variation *between* categories and is called the **sum of squares between (*SSB*)**. Again using our example to illustrate, *SSB* measures how different the people in each age group are from each other in their support for capital punishment. *SSW* and *SSB* are components of *SST*, as reflected in Formula 9.2:

FORMULA 9.2

$$SST = SSB + SSW$$

Let's start with the computation of *SSB*, our measure of the variation between categories. We use the category means as summary statistics to determine the size of the difference from category to category. In other words, we compare the average support for the death penalty for each age group with the average support for all age groups to determine *SSB*. The formula for the sum of squares between (*SSB*) is

FORMULA 9.3

$$SSB = \Sigma N_k(\bar{X}_k - \bar{X})^2$$

where SSB = the sum of squares between the categories
N_k = the number of cases in a category
$\bar{X}_k$ = the mean of a category

To find *SSB*, subtract the overall mean of all scores ($\bar{X}$) from each category mean ($\bar{X}_k$), square the difference, multiply by the number of cases in the category (N_k), and add the results across all the categories.

The second estimate of the population variance (*SSW*) is based on the amount of variation within the categories. Formula 9.2 shows that the total sum of squares (*SST*) is equal to the sum of *SSW* and *SSB*. This relationship means that we can find *SSW* by simple subtraction. Formula 9.4 rearranges the symbols in Formula 9.2:

FORMULA 9.4

$$SSW = SST - SSB$$

Let's pause for a second to remember what we are after here. If the null hypothesis is *true*, then there should not be much variation from category to category (see Table 9.1), and *SSW* and *SSB* should be roughly equal. If the null hypothesis is *not true*, there will be large differences between categories (see Table 9.2) relative to the differences within categories, and *SSB* should be much larger than *SSW*. *SSB* will increase as the difference *between* category means increases, especially when there is not much variation *within* the categories (*SSW*). The larger *SSB* is compared to *SSW*, the more likely we are to reject the null hypothesis.

The next step in the computational routine is to construct the estimates of the population variance. To do this, we divide each sum of squares by its respective degrees of freedom. To find the degrees of freedom associated with *SSW*, subtract the number of categories (k) from the number of cases (N). The degrees of freedom associated with *SSB* are the number of categories minus 1. In summary,

FORMULA 9.5

$$dfw = N - k$$

where dfw = degrees of freedom associated with *SSW*
N = total number of cases
k = number of categories

FORMULA 9.6

$$dfb = k - 1$$

where dfb = degrees of freedom associated with *SSB*
k = number of categories

The actual estimates of the population variance, called the **mean square estimates**, are calculated by dividing each sum of squares by its respective degrees of freedom:

FORMULA 9.7

$$\text{Mean Square Within} = \frac{SSW}{dfw}$$

FORMULA 9.8

$$\text{Mean Square Between} = \frac{SSB}{dfb}$$

The test statistic calculated in step 4 of the five-step model is called the ***F* ratio**, and its value is determined by the following formula:

FORMULA 9.9

$$F(\text{obtained}) = \frac{\text{Mean Square Between}}{\text{Mean Square Within}}$$

As you can see, the value of the *F* ratio is a function of the amount of variation between categories (based on *SSB*) compared to the amount of variation within the categories (based on *SSW*). The greater the variation between the categories relative to the variation within, the higher the value of the *F* ratio and the more likely we will reject the null hypothesis.

An Example of Computing the Analysis of Variance

Assume that we have administered our Support for Capital Punishment Scale to a sample of 16 individuals who are equally divided among the four age categories. (Obviously, this sample is much too small for any serious research and is intended solely for purposes of illustration.) All scores are reported in Table 9.3 along with the other quantities needed to complete the computations. The scores (X_i) are listed for each of the four categories, and a column has been added for the squared scores (X_i^2). The sums of both X_i and X_i^2 are reported at the bottom of each column. The category means ($\overline{X}_k$) show that the four 18–29-year-olds averaged 10.0 on the Support for Capital Punishment Scale, the four 30–45-year-olds averaged 13.0, and so forth. Finally, the overall mean (sometimes called the *grand mean*) is reported in the bottom row of the table. This shows that the average score for all 16 respondents was 15.25.

To organize our computations, we'll follow the routine summarized in the "One Step at a Time" box. We begin by finding *SST* (Formula 9.1):

$$SST = \sum X_i^2 - N\overline{X}^2$$
$$SST = (438 + 702 + 1058 + 1994) - (16)(15.25)^2$$
$$SST = 4192 - (16)(232.56)$$
$$SST = 4192 - 3720.96$$
$$SST = 471.04$$

TABLE 9.3 Support for Capital Punishment by Age Group (Fictitious Data)

18–29		30–45		46–64		65+	
X_i	X_i^2	X_i	X_i^2	X_i	X_i^2	X_i	X_i^2
7	49	10	100	12	144	17	289
8	64	12	144	15	225	20	400
10	100	13	169	17	289	24	576
15	225	17	289	20	400	27	729
40	438	52	702	64	1058	88	1994
$\overline{X}_k = 10.0$		$\overline{X}_k = 13.0$		$\overline{X}_k = 16.0$		$\overline{X}_k = 22.0$	
			$\overline{X} = 15.25$				

The sum of squares between (*SSB*) is found by means of Formula 9.3:

$$SSB = \sum N_k(\bar{X}_k - \bar{X})^2$$
$$SSB = 4(10 - 15.25)^2 + 4(13 - 15.25)^2 + 4(16 - 15.25)^2 + 4(22 - 15.25)^2$$
$$SSB = 4(-5.25)^2 + 4(-2.25)^2 + 4(0.75)^2 + 4(6.75)^2$$
$$SSB = 4(27.56) + 4(5.06) + 4(0.56) + 4(45.56)$$
$$SSB = 110.24 + 20.24 + 2.24 + 182.24$$
$$SSB = 314.96$$

Now *SSW* can be found by subtraction (Formula 9.4):

$$SSW = SST - SSB$$
$$SSW = 471.04 - 314.96$$
$$SSW = 156.08$$

To find the degrees of freedom for the two sums of squares, we use Formulas 9.5 and 9.6:

$$dfw = N - k = 16 - 4 = 12$$
$$dfb = k - 1 = 4 - 1 = 3$$

Finally, we are ready to construct the mean square estimates of the population variance. For the estimate based on *SSW*, we use Formula 9.7:

$$\text{Mean Square Within} = \frac{SSW}{dfw} = \frac{156.08}{12} = 13.00$$

For the between estimate, we use Formula 9.8:

$$\text{Mean Square Between} = \frac{SSB}{dfb} = \frac{314.96}{3} = 104.99$$

The test statistic, or *F* ratio, is found by using Formula 9.9:

$$F = \frac{\text{Mean Square Between}}{\text{Mean Square Within}} = \frac{104.99}{13.00} = 8.08$$

This statistic must still be evaluated for its significance. *(Solve any of the end-of-chapter problems to practice computing these quantities and solving these formulas.)*

A Test of Significance for ANOVA

Now we will see how to test an *F* ratio for significance. We will also take a look at some of the assumptions underlying the ANOVA test. As usual, we will follow the five-step model as a convenient way of organizing the decision-making process.

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Independent random samples
Level of measurement is interval-ratio
Populations are normally distributed
Population variances are equal

The model assumptions are quite strict and underscore the fact that ANOVA should be used only with dependent variables that have been carefully and precisely measured. However, as long as the categories are roughly equal in size, ANOVA can tolerate some violation of the model assumptions. In situations where you are uncertain or have samples of very different sizes, it is probably advisable to use an alternative test. (Chi square, described in Chapter 10, is one option.)

Step 2. Stating the Null Hypothesis. For ANOVA, the null hypothesis always states that the means of the populations from which the samples were drawn are equal. For our example problem, we are concerned with four different populations, or categories, so our null hypothesis would be

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

where μ_1 represents the mean for the 18–29 age group, μ_2 is the mean for people 30–45, and so forth.

The alternative hypothesis states simply that at least one of the population means is different. The wording here is important. If we reject the null hypothesis, ANOVA does not identify which mean or means are significantly different.

(H_1: At least one of the population means is different.)

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region. The sampling distribution for ANOVA is the *F* distribution, which is summarized in Appendix D. Note that there are separate tables for alphas of 0.05 and 0.01. As with the *t* table, the value of the critical *F* score will vary by degrees of freedom. For ANOVA, there are two separate degrees of freedom, one for each estimate of the population variance. The numbers across the top of the table are the degrees of freedom associated with the between estimate (*dfb*), and the numbers down the side of the table are those associated with the within estimate (*dfw*). In our example, *dfb* is $(k - 1)$, or 3, and *dfw* is $(N - k)$, or 12 (see Formulas 9.5 and 9.6). So, if we set alpha at 0.05, our critical *F* score will be 3.49.

Summarizing these considerations,

$$\begin{aligned} \text{Sampling distribution} &= F \text{ distribution} \\ \text{Alpha } (\alpha) &= 0.05 \\ \text{Degrees of freedom within } (dfw) &= (N - k) = 12 \\ \text{Degrees of freedom between } (dfb) &= (k - 1) = 3 \\ F(\text{critical}) &= 3.49 \end{aligned}$$

Take a moment to inspect the two *F* tables and you will notice that all the values are greater than 1.00. This is because ANOVA is a one-tailed test and we are concerned only with outcomes in which there is more variance between categories than within categories. *F* values of less than 1.00 would indicate that the between estimate was lower in value than the within estimate, and, because we would always fail to reject the null hypothesis in such cases, we simply ignore this class of outcomes.

Step 4. Computing the Test Statistic. This was done in the previous section, where we found an obtained *F* ratio of 8.08.

Step 5. Making a Decision and Interpreting the Results of the Test. Compare the test statistic with the critical value:

$$F(\text{critical}) = 3.49$$
$$F(\text{obtained}) = 8.08$$

The test statistic falls in the critical region, so our decision would be to reject the null hypothesis. Support for capital punishment differs significantly by age group.

ONE STEP AT A TIME Completing Step 4 of the Five-Step Model: Computing *F*(obtained)

To compute the obtained *F* ratio, use Formulas 9.1, 9.3, and 9.4 to find *SST*, *SSB*, and *SSW*. Then, calculate the degrees of freedom, mean square estimates of the population variance, and the obtained *F* ratio. It is strongly recommended that you use a computing table like Table 9.3 to organize the computations.

Step	***Operation***
To solve Formula 9.1 (SST):	
1.	To find ΣX_i^2,
	a. Square each score.
	b. Find the sum of the squared scores.
2.	To find $N\bar{X}^2$,
	a. Square the value of the mean of all scores.
	b. Multiply $\bar{X}^2$ by *N*.
3.	Subtract the quantity you found in step 2 from the quantity you found in step 1 to find *SST*.
To solve Formula 9.3 (SSB):	
1.	Subtract the mean of all scores ($\bar{X}$) from the mean of each category ($\bar{X}_k$) and then square each difference.
2.	Multiply each of the squared differences you found in step 1 by the number of cases in the category (N_k).
3.	Add all of the quantities you found in step 2 to find *SSB*.
To solve Formula 9.4 (SSW):	
1.	Subtract the value of *SSB* from the value of *SST.*
To solve Formulas 9.5 and 9.6 (dfw *and* dfb):	
1.	For *dfw*, subtract the number of categories (*k*) from the number of cases (*N*).
2.	For *dfb*, subtract 1 from the number of categories (*k*).
To solve Formulas 9.7 and 9.8 (Mean Square Estimates):	
1.	To find the Mean Square Within estimate, divide *SSW* by *dfw*.
2.	To find the Mean Square Between estimate, divide *SSB* by *dfb*.
To solve Formula 9.9 (Obtained F *ratio):*	
1.	Divide the Mean Square Between estimate by the Mean Square Within estimate.

ONE STEP AT A TIME — Completing Step 5 of the Five-Step Model: Making a Decision and Interpreting Results

Step	***Operation***
1.	Compare *F*(obtained) to *F*(critical). If *F*(obtained) is greater than *F*(critical), it is *in* the critical region. *Reject* the null hypothesis. If *F*(obtained) is less than *F*(critical), it is *not* in the critical region. *Fail to reject* the null hypothesis.
2.	Interpret the decision to reject or fail to reject the null hypothesis in terms of the original question. For example, our conclusion for the example problem was "Support for capital punishment varies significantly by age group."

STATISTICS IN EVERYDAY LIFE

Battling the Bulge

Want to stay slim and trim? A study that followed over 34,000 women for 15 years is not particularly encouraging: On the average, the women in the sample put on pounds during the study. However, exercise did help to limit the weight gain. The women were divided into three groups, and those who exercised at the highest levels gained significantly fewer pounds than the women in the moderate- and low-exercise groups. This study used sophisticated statistics not covered in this text to analyze the data, but the statistical significance of the differences in mean weight gain for the three groups was established by an ANOVA technique.

Source: Lee, I., Djousse, L, Sesso, H., Wang, L., and Buring J. 2010. "Physical Activity and Weight Gain Prevention." *JAMA*, March 24–31, 1173–1179.

Applying Statistics 9.1 The Analysis of Variance

A random sample of 20 nations from four income levels has been selected. We used these economic categories in Chapter 4 when discussing birth rates in connection with boxplots. To remind you, "Low income" nations are largely agricultural and have the lowest quality of life. "High Income" nations are industrial and the most affluent and modern. "Upper- and Lower-Middle income" nations are between these extremes. Are these income levels reflected in differences in life expectancy (the number of years the average citizen can expect to live at birth)?

Life Expectancy (L.E.) by Income Level

INCOME LEVELS							
Low		Lower-Middle		Upper-Middle		High	
Nation	L.E.	Nation	L.E.	Nation	L.E.	Nation	L.E.
Cambodia	62	Bolivia	67	China	75	Australia	82
Malawi	54	Guatemala	71	Colombia	74	Canada	81
Nepal	68	India	66	Jordan	73	Japan	83
Niger	57	Philippines	69	Libya	75	Spain	82
Zimbabwe	56	Vietnam	73	Russia	69	U.S.	79

Source: Population Reference Bureau. 2012 World Population Data Sheet. Available at http://www.prb.org/Publications/Datasheets/2012/world-population-data-sheet.aspx

(*continued*)

Applying Statistics 9.1 *(continued)*

To find *F*(obtained) and conduct the ANOVA test, computations will be organized into table format:

	Low		Lower-Middle		Upper-Middle		High	
	X_i	X_i^2	X_i	X_i^2	X_i	X_i^2	X_i	X_i^2
	62	3844	67	4489	75	5625	82	6724
	54	2916	71	5041	74	5476	81	6561
	68	4624	66	4356	73	5329	83	6889
	57	3249	69	4761	75	5625	82	6724
	56	3136	73	5329	69	4761	79	6241
$\Sigma X_i =$	297		346		366		407	
$\Sigma X_i^2 =$		17,769		23,976		26,816		33,139
$\bar{X}_k =$	59.4		69.2		73.2		81.4	
				$\bar{X} = 70.8$				

The ANOVA test will tell us if these differences are large enough to justify the conclusion that they did not occur by chance alone. Following the usual computational routine,

$$SST = \Sigma X_i^2 - N\bar{X}^2$$
$$SST = (17{,}769 + 23{,}976 + 26{,}816 + 33{,}139) - (20)(70.8)^2$$
$$SST = (101{,}700) - (20)(5012.64)$$
$$SST = 101{,}700 - 100{,}252.80$$
$$SST = 1447.20$$

$$SSB = \Sigma N_k(\bar{X}_k - \bar{X})^2$$
$$SSB = 5(59.4 - 70.8)^2 + 5(69.2 - 70.8)^2 + 5(73.2 - 70.8)^2 + 5(81.4 - 70.8)^2$$
$$SSB = 5(-11.40)^2 + 5(-1.60)^2 + 5(2.40)^2 + 5(10.60)^2$$
$$SSB = 5(129.96) + 5(2.56) + 5(5.76) + 5(112.36)$$
$$SSB = 649.80 + 12.80 + 28.80 + 561.80$$
$$SSB = 1253.20$$

$$SSW = SST - SSB$$
$$SSW = 1447.20 - 1253.20$$
$$SSW = 194$$
$$dfw = N - k = 20 - 4 = 16$$
$$dfb = k - 1 = 4 - 1 = 3$$

$$\text{Mean Square Within} = \frac{SSW}{dfw} = \frac{194}{16} = 12.13$$

$$\text{Mean Square Between} = \frac{SSB}{dfb} = \frac{1253.20}{3} = 417.73$$

$$F(\text{obtained}) = \frac{\text{Mean Square Between}}{\text{Mean Square Within}} = \frac{417.73}{12.13} = 34.44$$

We can now conduct the test of significance.

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Independent random samples
Level of measurement is interval-ratio
Populations are normally distributed
Population variances are equal

Step 2. Stating the Null Hypothesis.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

(H_1: At least one of the population means is different.)

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region.

Sampling distribution = *F* distribution
Alpha = 0.05
dfw = 16
dfb = 3
F(critical) = 3.24

Step 4. Computing the Test Statistic. We found an obtained *F* ratio of 34.44.

(continued)

Step 5. Making a Decision and Interpreting the Results of the Test. Compare the test statistic with the critical value

$$F(\text{critical}) = 3.24$$
$$F(\text{obtained}) = 34.44$$

The null hypothesis ("The population means are equal") can be rejected. The differences in life expectancy between nations at different income levels are statistically significant and reflect differences in the populations from which these samples were selected.

An Additional Example of Computing and Testing the Analysis of Variance

In this section, we work through an additional example of the computation and interpretation of the ANOVA test. We first review matters of computation, find the obtained F ratio, and then test the statistic for its significance.

A researcher has been asked to evaluate the efficiency of three social service agencies. One area of concern is how fast the agencies are determining the eligibility of clients for a particular program. The researcher has information on the number of days it took to process random samples of ten cases in each agency. Is there a significant difference between the agencies? The data are reported in Table 9.4, which also includes the information we will need to complete our calculations.

TABLE 9.4 Number of Days Required to Process Cases for Three Agencies (Fictitious Data)

	Agency A		Agency B		Agency C	
Client	X_i	X_i^2	X_i	X_i^2	X_i	X_i^2
1	5	25	12	144	9	81
2	7	49	10	100	8	64
3	8	64	19	361	12	144
4	10	100	20	400	15	225
5	4	16	12	144	20	400
6	9	81	11	121	21	441
7	6	36	13	169	20	400
8	9	81	14	196	19	361
9	6	36	10	100	15	225
10	6	36	9	81	11	121
$\Sigma X_i =$	70		130		150	
$\Sigma X_i^2 =$		524		1816		2462
$\bar{X}_k =$	7.0		13.0		15.0	

$$\bar{X} = \frac{350}{30} = 11.67$$

To find *SST* by means of Formula 9.1,

$$SST = \sum X_i^2 - N\bar{X}^2$$
$$SST = (524 + 1816 + 2462) - (30)(11.67)^2$$
$$SST = 4802 - (30)(136.19)$$
$$SST = 4802 - 4085.70$$
$$SST = 716.30$$

To find *SSB* by means of Formula 9.3,

$$SSB = \sum N_k(\bar{X}_k - \bar{X})^2$$
$$SSB = (10)(7.0 - 11.67)^2 + (10)(13.0 - 11.67)^2 + (10)(15.0 - 11.67)^2$$
$$SSB = (10)(-4.67)^2 + (10)(1.33)^2 + (10)(3.33)^2$$
$$SSB = (10)(21.81) + (10)(1.77) + (10)(11.09)^2$$
$$SSB = 218.10 + 17.70 + 110.90$$
$$SSB = 346.70$$

Now we can find *SSW* using Formula 9.4:

$$SSW = SST - SSB$$
$$SSW = 716.30 - 346.70$$
$$SSW = 369.60$$

The degrees of freedom are found through Formulas 9.5 and 9.6:

$$dfw = N - k = 30 - 3 = 27$$
$$dfb = k - 1 = 3 - 1 = 2$$

The estimates of the population variances are found by means of Formulas 9.7 and 9.8:

$$\text{Mean Square Within} = \frac{SSW}{dfw} = \frac{369.6}{27} = 13.69$$

$$\text{Mean Square Between} = \frac{SSB}{dfb} = \frac{346.70}{2} = 173.35$$

The *F* ratio (Formula 9.9) is

$$F(\text{obtained}) = \frac{\text{Mean Square Between}}{\text{Mean Square Within}} = \frac{173.35}{13.69} = 12.66$$

And we can now test this value for its significance.

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Independent random samples
Level of measurement is interval-ratio
Populations are normally distributed
Population variances are equal

The researcher will always be in a position to judge the adequacy of the first two assumptions in the model. The second two assumptions are more problematic, but remember that ANOVA will tolerate some deviation from its assumptions as long as sample sizes are roughly equal.

Step 2. Stating the Null Hypothesis.

$$H_0: \mu_1 = \mu_2 = \mu_3$$
(H_1: At least one of the population means is different.)

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region.

Sampling distribution = F distribution
Alpha = 0.05
$dfw = 27$
$dfb = 2$
$F(\text{critical}) = 3.35$

Step 4. Computing the Test Statistic. We found an obtained F ratio of 12.66.

Step 5. Making a Decision and Interpreting the Results of the Test. Compare the test statistic with the critical value:

$F(\text{critical}) = 3.35$
$F(\text{obtained}) = 12.66$

The test statistic is in the critical region, and we would reject the null hypothesis of no difference. The differences between the three agencies are very unlikely to have occurred by chance alone. The agencies are significantly different in the speed with which they process paperwork and determine eligibility. *(For practice in conducting the ANOVA test, see problems 9.2 to 9.10. Begin with the lower-numbered problems since they have smaller datasets, fewer categories, and, therefore, the simplest calculations.)*

Using SPSS: The ANOVA Test

SPSS provides several different ways of conducting the analysis of variance test. The procedure summarized here is called **One-Way-ANOVA** and is the most accessible of these, but it still incorporates options and capabilities that we have not covered in this chapter. If you wish to explore these possibilities, please use the online **Help** facility.

To provide an illustration, we will follow up on the example used in Chapter 9, in which we tested for significant differences in *rank* between high school and college graduates. ANOVA will allow us to expand the number of categories in the independent variable (*degree*) beyond the two we used in Chapter 9. As you recall, the dependent variable (*rank*) is the respondent's self-ranking of their social position on a scale of 1 to 10, with 1 representing the highest position.

To begin,

1. Click the SPSS icon on your desktop.
2. Load the *GSS2012* database.
   a. Find the **File** command on the far left of the menu bar and then click **File → Open → Data**.
   b. Find the *GSS2012* database and click the file name.
3. From the main menu bar, click **Analyze**, then **Compare Means**, and then "One-Way ANOVA". The "One-Way ANOVA" window appears. Find *rank* in the variable list on the left and click the arrow to move the variable name into the **Dependent List** box. Note that you can request more than one dependent variable at a time. Next, find the *degree* and click the arrow to move the variable name into the "Factor" box.
4. Click **Options** and then click the box next to **Descriptive** in the "Statistics" box to request means and standard deviations along with the analysis of variance. Click **Continue** and then click **OK**, and the following output will be produced:

Descriptives

RS Self Ranking of Social Position

	N	Mean	Std. Devia2tion	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
LT HIGH SCHOOL	213	5.51	2.136	.146	5.22	5.80	1	10
HIGH SCHOOL	653	4.90	1.794	.070	4.76	5.04	1	10
JUNIOR COLLEGE	106	4.63	1.785	.173	4.29	4.98	1	10
BACHELOR	264	4.34	1.591	.098	4.15	4.54	1	9
GRADUATE	143	4.21	1.674	.140	3.93	4.49	1	10
Total	1379	4.80	1.844	.050	4.70	4.89	1	10

The output box labeled "Descriptives" presents the category means and shows, not surprisingly, that the respondents with the least education—"LT (or less than) High School"—had the highest average score on *rank*. Remember that higher scores on this variable indicate lower social status. The scores on rank decrease as education increases and respondents with graduate degrees had the lowest score (or highest social position). Are the differences significant?

ANOVA

RS Self Ranking of Social Position

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	222.000	4	55.500	17.084	.000
Within Groups	4463.740	1374	3.249		
Total	4685.740	1378			

The output box labeled "ANOVA" reports the degrees of freedom, all of the sums of squares, the Mean Square estimates, the *F* ratio (17.084), and, at the far right, the exact probability ("Sig.") of getting these results if the null hypothesis is true. This is reported as .000, much lower than our usual alpha level of 0.05. The differences in *rank* for the various levels of education are statistically significant.

STATISTICS
IN EVERYDAY LIFE

Generation Gaps?

Have you noticed that opinions about controversial issues such as gay marriage tend to vary by age group? Younger people tend to be more tolerant of homosexuality and more supportive of gay rights, as they are on many issues. For example, the 2012 General Social Survey shows that almost two-thirds (65.4%) of the oldest respondent (ages 65 and older) said that homosexual relations were "always wrong." In contrast, the majority (57%) of younger respondents (ages 18 to 34) said homosexual relations were "not wrong at all."

An analysis of variance test shows that the relationship between age group and opinion on homosexuality is significant (see the table below). Tolerance is highest in the youngest age group and decreases as age group increases.

Opinion on Homosexuality by Age Group

ITEM: What About Sexual Relations Between Two Adults of the Same Sex: Do You Believe it is Always Wrong (1), Almost Always Wrong (2), Wrong Only Sometimes (3), or Not Wrong at All (4)?

Age Group	*N*	Mean	*F* Ratio	Significance
18–34	329	2.91	23.03	0.000
35–50	377	2.58		
51–64	296	2.40		
65+	237	1.94		
	1239			

Why would opinions vary by age group? Do people grow more conservative as they grow older? Or does this pattern reflect the differing social worlds in which each generation is raised? How could you investigate this issue further?

The Limitations of the Test

ANOVA is appropriate when you want to test differences between the means of an interval-ratio–level dependent variable between three or more categories of an independent variable. This application is called a **one-way analysis of variance**, because it involves the effect of a single variable (for example, religion) on another (for example, support for capital punishment). This is the simplest application of ANOVA, and you should be aware that the technique has numerous more advanced and complex forms. For example, you may encounter research projects in which the effects of two separate variables (for example, religion and gender) on some third variable were observed.

One important limitation of ANOVA is that it requires an interval-ratio dependent variable and roughly equal numbers of cases in each category of the independent variable. The former condition may be difficult to meet with complete confidence for many variables of interest in the social sciences. The latter condition may create problems when the research hypothesis calls for comparisons between groups that are, by their nature, unequal in numbers (for example, white versus black Americans). Neither of these limitations should be particularly crippling, because ANOVA tolerates some deviation from its model assumptions, but you should be aware of these limitations in planning your own research as well as in judging the adequacy of research conducted by others.

Another limitation of ANOVA, introduced in Chapter 8, actually applies to all forms of significance testing. Tests of significance are designed to detect nonrandom differences or differences so large that they are unlikely to be produced by random chance alone. The problem is that differences that are statistically significant are not necessarily important in any other sense. Parts III and IV of this text provide some statistical techniques that can directly assess the importance of results.

A final limitation of ANOVA relates to the research hypothesis, which simply asserts that at least one of the population means is different from the others. Obviously, when we reject the null hypothesis, we would like to know *which* differences between the sample means are significant. We can sometimes make this determination by simple inspection. In our problem involving social service agencies, for example, it is pretty clear from Table 9.4 that Agency A is the source of most of the differences. This informal "eyeball" method can be misleading, however, and you should exercise caution in drawing conclusions.

BECOMING A CRITICAL CONSUMER: Reading the Professional Literature

It is unlikely that you would encounter a report using ANOVA in everyday life or in the popular media, and I will confine this section to the professional research literature. As I have pointed out previously, reports about tests of significance in social science research journals may be short on detail, but you will still be able to locate all of the essential information needed to understand the results of the test.

We can use a recent research article to illustrate how to read ANOVA results in the professional literature. Several researchers were concerned with competing theories of the relationship between religion and sexual behavior. One theory, the secularization hypothesis, argues that the modern reliance on reason and science has softened the relationship between religious belief and sexual experience. The researchers used a large sample of university students and some of their results (for females only) are presented below. They developed a variety of categories to indicate religious affiliation, including "Spiritual" (which included a variety of modern "New Age" belief systems) and "Monotheist Christian" (which included all mainstream Protestants and Catholics).

The entries in the table are mean scores. Note that that they do not vary much across the religious affiliations. This impression is verified by the low *F* ratios in the right-hand column, none of which are statistically significant. These results are broadly consistent with the secularization hypothesis: The participant's professed religious views seem to have little effect on their sexual behavior.

Want to learn more? The citation is given below.

Female Sexual Behavior (means) by Religious Affiliation

	Spiritual	Agnostic	Atheist	None	Monotheist Christian	Fundamentalist	*F* Score
Age at First Sexual Intercourse	17.33	17.11	17.67	16.38	16.56	16.69	1.66
Number of Sex Partners in Past Year	1.50	2.00	2.13	1.89	1.93	1.52	1.48
Times Cheated on Partner	1.50	1.62	1.44	1.71	1.40	1.31	1.29

Based on Table 2, p. 859

Source: Farmer, Melissa, Trapnell, Paul, & Meston, Cindy. 2010. "The Relation Between Sexual Behavior and Religiosity Subtypes: A Test of the Secularization Hypothesis." *Archives of Sexual Behavior*: 38: 852–865.

SUMMARY

1. One-way analysis of variance is a powerful test of significance that is commonly used when comparisons across more than two categories or samples are of interest. It is perhaps easiest to conceptualize ANOVA as an extension of the test for the difference in sample means.
2. ANOVA compares the amount of variation within the categories to the amount of variation between categories. If the null hypothesis of no difference is false, there should be relatively great variation between categories and relatively little variation within categories. The greater the differences from category to category relative to the differences within the categories, the more likely we will be able to reject the null hypothesis.
3. The computational routine for even simple applications of ANOVA can quickly become quite complex (a fact that suggests the value of computerized statistical packages such as SPSS). The basic process is to construct separate estimates of the population variance based on the variation within the categories and the variation between the categories. The test statistic is the *F* ratio, which is based on a comparison of these two estimates.
4. The ANOVA test can be organized into the familiar five-step model for testing the significance of sample outcomes. Although the model assumptions (step 1) require high-quality data, the test can tolerate some deviation as long as sample sizes are roughly equal. The null hypothesis takes the familiar form of stating that there is no difference of any importance among the population values, while the alternative hypothesis asserts that at least one population mean is different. The sampling distribution is the *F* distribution, and the test is always one-tailed. The decision to reject or to fail to reject the null hypothesis is based on a comparison of the obtained *F* ratio with the critical *F* ratio as determined for a given alpha level and degrees of freedom. The decision to reject the null hypothesis indicates only that one or more of the population means is different from the others. We can often determine which of the sample means accounts for the difference by inspecting the sample data, but this informal method should be used with caution.

SUMMARY OF FORMULAS

FORMULA 9.1	Total sum of squares:	$SST = \sum X_i^2 - N\overline{X}^2$
FORMULA 9.2	The two components of the total sum of squares:	$SST = SSB + SSW$
FORMULA 9.3	Sum of squares between:	$SSB = \sum N_k(\overline{X}_k - \overline{X})^2$
FORMULA 9.4	Sum of squares within:	$SSW = SST - SSB$
FORMULA 9.5	Degrees of freedom for *SSW*:	$dfw = N - k$
FORMULA 9.6	Degrees of freedom for *SSB*:	$dfb = k - 1$
FORMULA 9.7	Mean square within:	$\text{Mean Square Within} = \dfrac{SSW}{dfw}$
FORMULA 9.8	Mean square between:	$\text{Mean Square Between} = \dfrac{SSB}{dfb}$
FORMULA 9.9	*F* ratio:	$F(\text{obtained}) = \dfrac{\text{Mean Square Between}}{\text{Mean Square Within}}$

GLOSSARY

Analysis of variance. A test of significance appropriate for situations in which we are concerned with the differences between more than two sample means.

ANOVA. See **Analysis of variance**.

***F* ratio.** The test statistic computed in step 4 of the ANOVA test.

Mean square estimate. An estimate of the variance calculated by dividing the sum of squares within (*SSW*) or the sum of squares between (*SSB*) by the proper degrees of freedom.

One-way analysis of variance. Applications of ANOVA in which the effect of a single independent variable on a dependent variable is observed.

Sum of squares between (*SSB*). The sum of the squared deviations of the sample means from the overall mean, weighted by sample size.

Sum of squares within (*SSW*). The sum of the squared deviations of scores from the category means. Equal to SST − SSB.

Total sum of squares (*SST*). The sum of the squared deviations of the scores from the overall mean.

PROBLEMS

Note: The number of cases in these problems is very low—a fraction of the sample size necessary for any serious research—in order to simplify computations.

9.1 Conduct an ANOVA test for each of the following sets of scores. (*HINT: Follow the computational routine outlined in the "One Step at a Time" box, and keep track of all sums and means by constructing computational tables like Table 9.3 or Table 9.4.*)

a.

Category		
A	B	C
5	10	12
7	12	16
8	14	18
9	15	20

b.

Category		
A	B	C
1	2	3
10	12	10
9	2	7
20	3	14
8	1	1

c.

Category			
A	B	C	D
13	45	23	10
15	40	78	20
10	47	80	25
11	50	34	27
10	45	30	20

9.2 *SOC* A local church has been experiencing steady growth in membership and wants to launch a membership drive. Before they begin, they would like to find out what types of people are most likely to attend services each week. A random sample of 15 existing members has been asked how many times they attended services over the past five weeks and for certain other information. Which independent variables have significant relationships with attendance?

a. Attendance by education

Less Than High School	High School	College
0	1	0
1	3	3
2	3	4
3	4	4
4	5	4

b. Attendance by length of residence in present community

Less Than 2 Years	2–5 Years	More Than 5 Years
0	0	1
1	2	3
3	3	3
4	4	4
4	5	4

c. Attendance by extent of television watching

Little or none	Moderate	High
0	3	4
0	3	4
1	3	4
1	3	4
2	4	5

d. Attendance by number of children

None	One Child	More Than One Child
0	2	0
1	3	3
1	4	4
3	4	4
3	4	5

9.3 *SOC* In a local community, a random sample of 18 couples has been assessed on a scale that measures the extent to which power and decision making are shared (lower scores) or monopolized by one party (higher scores) and on marital happiness (higher scores indicate higher levels of happiness). The couples were also classified by type of relationship: traditional (only the husband works outside the home), dual-career (both parties work), and cohabitational (parties living together but not legally married, regardless of work patterns). Does decision making or happiness vary significantly by type of relationship?

a. Decision making

Traditional	Dual-Career	Cohabitational
7	8	2
8	5	1
2	4	3
5	4	4
7	5	1
6	5	2

b. Happiness

Traditional	Dual-Career	Cohabitational
10	12	12
14	12	14
20	12	15
22	14	17
23	15	18
24	20	22

9.4 *CJ* Two separate after-school programs for at-risk teenagers have been implemented in the city of Redland. One is a big-brother/big-sister mentoring program, in which adult volunteers are paired with troubled teenagers. The second involves extracurricular team sports competitions. In terms of the percentage change in crimes reported to the police over a one-year period, were the programs successful? The results are for random samples of 18 neighborhoods drawn from the entire city.

Mentoring Program	Competition Program	No Program
−10	−21	+30
−20	−15	−10
+10	−80	+14
+20	−10	+80
+70	−50	+50
+10	−10	−20

9.5 *SOC* Are sexually active teenagers any better informed about AIDS and other potential health problems related to sex than teenagers who are sexually inactive? A 15-item test of general knowledge about sex and health was administered to random samples of teens who are sexually inactive, teens who are sexually active but with only a single partner, and teens who are sexually active with more than one partner. Is there any significant difference in the test scores?

Inactive	Active—One Partner	Active—More Than One Partner
10	11	12
12	11	12
8	6	10
10	5	4
8	15	3
5	10	15

9.6 *PS* Does the rate of voter turnout vary significantly by the type of election? A random sample of voting precincts displays the following pattern of voter turnout by election type. Assess the results for significance.

Local Only	State	National
33	35	42
78	56	40
32	35	52
28	40	66
10	45	78
12	42	62
61	65	57
28	62	75
29	25	72
45	47	51
44	52	69
41	55	59

9.7 *GER* Do older citizens lose interest in politics and current affairs? A brief quiz on recent headline stories was administered to random samples of respondents from each of four age groups. Is there a significant difference? The following data represent numbers of correct responses.

High School (15–18)	Young Adult (21–30)	Middle-Aged (30–55)	Retired (65+)
0	0	2	5
1	0	3	6
1	2	3	6
2	2	4	6
2	4	4	7
2	4	5	7
3	4	6	8
5	6	7	10
5	7	7	10
7	7	8	10
7	7	8	10
9	10	10	10

(*continued next column*)

(*continued*)

9.8 *SOC* A small random sample of respondents has been selected from the General Social Survey database. Each respondent has been classified as either an urban dweller, a suburbanite, or a rural dweller. Are there statistically significant differences by place of residence for any of the variables listed here? (See Appendix G for definitions and scores of the dependent variables.)

a. Hours per week on the Internet (*wwwhr*)

Urban	Suburban	Rural
5	5	5
5	6	6
7	10	7
8	11	10
8	12	10
10	14	11
10	15	12
11	16	12
12	18	12
13	20	13

b. Number of children (*childs*)

Urban	Suburban	Rural
1	0	1
1	1	4
0	0	2
2	0	3
1	2	3
0	2	2
2	3	5
2	2	0
1	2	4
0	1	6

c. Family income (*income06*)

Urban	Suburban	Rural
5	6	5
7	8	5
8	11	11
11	12	10
8	12	9
9	11	6
8	11	10
3	9	7
9	10	9
10	12	8

d. Church attendance (*attend*)

Urban	Suburban	Rural
0	0	1
7	0	5
0	2	4
4	5	4
5	8	0
8	5	4
7	8	8
5	7	8
7	2	8
4	6	5

e. Hours of TV watching per day (*tvhours*)

Urban	Suburban	Rural
5	5	3
3	7	7
12	10	5
2	2	0
0	3	1
2	0	8
3	1	5
4	3	10
5	4	3
9	1	1

9.9 *SOC* Does support for suicide ("death with dignity") vary by social class? Is this relationship different in different nations? Small samples in three nations were asked whether it is ever justified for a person with an incurable disease to take his or her own life. Respondents answered in terms of a 10-point scale, on which 10 was "always justified" (the strongest support for "death with dignity") and 1 was "never justified" (the lowest level of support). Results are reported here.

Mexico			
Lower Class	Working Class	Middle Class	Upper Class
5	2	1	2
2	2	1	4
4	1	3	5
5	1	4	7
4	6	1	8
2	5	2	10
3	7	1	10
1	2	5	9
1	3	1	8
3	1	1	8

Canada			
Lower Class	Working Class	Middle Class	Upper Class
7	5	1	5
7	6	3	7
6	7	4	8
4	8	5	9
7	8	7	10
8	9	8	10
9	5	8	8
9	6	9	5
6	7	9	8
5	8	5	9

United States			
Lower Class	Working Class	Middle Class	Upper Class
4	4	4	1
5	5	6	5
6	1	7	8
1	4	5	9
3	3	8	9
3	3	9	9
3	4	9	8
5	2	8	6
3	1	7	9
6	1	2	9

9.10 *SOC* A researcher is studying the effects of the college experience on attitudes, values, and behaviors and is comparing small random samples from each class at the same university. Which of the following differences are significant?

a. Number of "close friends" from another racial or ethnic group

Class			
Freshman	Sophomore	Junior	Senior
7	0	10	0
2	3	9	2
5	4	8	3
12	2	5	1
9	7	1	1
1	2	8	4
0	0	7	5
3	3	12	7
4	4	10	10
6	2	11	9

b. Score on a scale that measures political ideology. On this scale, 10 means "very conservative" and 1 means "very liberal."

Class			
Freshman	Sophomore	Junior	Senior
7	7	9	1
7	3	8	1
5	7	5	5
10	8	0	6
9	5	1	5
7	4	1	3
6	1	2	5
5	2	3	6
4	4	5	1
3	2	6	5

c. Number of minutes spent in the library each day studying

Class			
Freshman	Sophomore	Junior	Senior
0	30	45	60
30	90	45	0
45	120	90	0
60	100	180	120
90	75	145	45
45	60	100	45
30	30	90	30
15	45	110	60
0	20	120	60
30	45	110	180

(*continued next column*)

(*continued*)

Statistical Analysis Using SPSS

9.11 *SOC* In this exercise, you will follow up on Problems 8.17 and 8.18, in which you compared several pairs of religious affiliations (Protestants and Catholics, Protestants and "Nones") for significant differences in level of education (*educ*) and self-rated social position (*rank*). We will now use **One-Way ANOVA** to compare all five categories of the *relig* variable in a single test.

- Click the SPSS icon on your desktop.
- Load the *GSS2012* database.
- Click **Analyze → Compare Means → One-Way ANOVA**.
- Find *educ* and *rank* in the list of variables and move them into the "Dependent List" box.
- Find *relig* in the list of variables and move it into the "Factor" box.
- Click **Options** and click the box next to **Descriptive** in the "Statistics" box to get statistical information on each category of *relig*. Click **Continue**.
- Click **OK** and the results of the test will be printed to the SPSS output window.
- Find the group means, the *F* ratio, degrees of freedom, and the "Sig." (which is the exact probability of getting this difference if is true) and report the results of the test in a few sentences.

9.12 *SOC* Repeat the test in Problem 9.11 with some new dependent variables: *spanking* and *tvhours*. Are there significant differences between the religions in their support of spanking as a way of disciplining children or in their television-watching habits? Repeat all the steps listed in Problem 9.11 but substitute *spanking* and *tvhours* for *educ* and *rank*. Report the results of the test in a few sentences, including all of the essential statistical information. NOTE: *spanking* is an ordinal variable, coded so that *lower* scores indicate *greater* support of spanking. See Appendix G or click **Utilities → Variables** in SPSS to see the coding scheme.

YOU ARE THE RESEARCHER

Why Are Some People Liberal (or Conservative)?

This project investigates political ideology using *polviews*—a seven-point scale that measures how liberal or conservative a person is—as the dependent variable. This variable is ordinal in level of measurement but will be treated as interval-ratio for this test. What social forces affect political views?

One possible independent variable is *age*: People may grow more conservative as they grow older. It might be very appropriate to investigate the possible causal impact of aging, but note that the variable is in the wrong form. Independent variables in ANOVA must have just a few categories, and respondents to the 2012 GSS ranged in age from 18 to over 89. How can we change age to a form that fits the test? In the next section, I introduce the **Recode** command, which enables us to collapse scores on a variable.

Recoding Variables

We will recode *age* into a variable with just a few scores, which we can use as an independent variable in ANOVA. We will create a new version of *age* that has three categories. When we are finished, we will have two versions of the same variable in the data set: the original interval-ratio version with age measured in years, and a new ordinal-level version with collapsed categories. If we wish, we can add the new version of age to the permanent data file and use it in the future.

The decision to use three categories for *age* is arbitrary; we could easily have decided on four, five, or even six categories for the new, recoded independent variable. If we find that we are unhappy with the three-category version of the variable, we can always recode the original version of *age* into a different set of categories.

We will collapse the values of *age* into three broad categories with roughly equal numbers of cases in each category. How can we define these new categories? Begin by running the **Frequencies** command for *age* to inspect the distribution of the scores. I used the cumulative percent column of the frequency distribution to find the ages that divide the variable into thirds. I found that 32.3% of the 2012 GSS sample was younger than 37 and that 67.1% were younger than 56. I decided to use these ages as the cut points for the new categories, as summarized below:

Ages	Percent of Sample
18–37	32.3%
38–56	34.8%
57–89	32.9%
	100.0%

To recode *age* into these categories, follow these steps:

1. In the “SPSS Data Editor” window, click **Transform** from the menu bar and then click **Recode**. A window will open that gives us two choices: **into same variable** or **into different variable**. If we choose **into same variable**, the new version of the variable will *replace* the old and the variable *age* (measured in years) would disappear. We definitely do *not* want this to happen, so we will choose (click on) **into different variable**. This option will allow us to keep both the old and new versions of the variable.

2. The “Recode into Different Variable” window will open. The usual box listing the variables will be on the left. Use the cursor to highlight *age*, and then click on the arrow button to move the variable to the “Input Variable → Output Variable” box. The input variable is the old version of age, and the output variable is the new, recoded version we will soon create.
3. In the “Output Variable” box on the right, click in the “Name” box and type a name for the new (output) variable. I suggest *ager* (“**age r**ecoded”) for the new variable, but you can assign any name as long as it does not duplicate the name of some other variable in the data set. Click the **Change** button and the expression *age* → *ager* will appear in the “Input Variable → Output Variable” box.
4. Click on the **Old and New Values** button in the middle of the screen, and a new dialog box will open. Read down the left-hand column until you find the **Range** button. Click on the button, and the cursor will move to the small box immediately below. Type **18** (the youngest age in the sample) into the left-hand “Range” dialog box and then click on the right-hand box and type **37**.
5. In the “New Value” box in the upper-right corner of the screen, click the **Value** button. Type **1** in the “Value” dialog box and then click the **Add** button directly below. The expression “**18–37** → **1**” will appear in the “Old → New” dialog box.
6. Return to the “Range” dialog boxes. Type **38** in the left-hand box and **56** in the right-hand box, and then click the **Value** button in the “New Values” box. Type **2** in the “Value” dialog box and then click the **Add** button. The expression “**38–56**” appears in the “Old → New” dialog box.
7. Finish by returning to the “Range” dialog box and entering the values **57** and **89** in the correct boxes. Click the **Value** button in the “New Values” box. Type **3** in the “Value” dialog box and then click the **Add** button. The expression “**57–89** → **3**” appears in the “Old → New” dialog box.
8. Click the **Continue** button at the bottom of the screen, and you will return to the “Recode into Different Variable” dialog box. Click **OK**, and SPSS will execute the transformation.

You now have a data set with an additional variable named *ager* (or whatever name you gave the recoded variable). You can find the new variable in the last column to the right in the data window. You can make the new variable a permanent part of the data set by saving the data file at the end of the session. If you do not wish to save the new, expanded data file, click **No** when you are asked whether you want to save the data file. If you are using the student version of SPSS remember that you are limited to a maximum of 50 variables and you may not be able to save the new variable.

Investigating Political Ideology (*polviews*)

Why are some people more liberal and others more conservative? In this project, you will analyze the relationships between *polviews* (the dependent variable), recoded *age,* and two more independent variables of your own choosing. As you analyze and interpret your results, remember that *polviews* is coded so that higher scores indicate greater conservatism (see Appendix G or click **Utilities** → **Variables** on the SPSS menu bar).

Step 1: Choosing Independent Variables

Select recoded ***age*** and two more possible independent variables from the 2012 GSS. What factors might help to explain a person’s political views? Choose *only* independent variables with three to six scores or categories. Among other possibilities, you might consider education (use *degree*), religious denomination (*relig*), or social class (*class*). Use the **Recode** command to collapse independent variables with more than six categories into three or four categories. In general, variables that measure personal characteristics or traits (like gender or race) will work better than those that measure attitude or opinion (like *cappun* or *gunlaw*), which are more likely to be manifestations of political ideology, not causes.

List your independent variables in the table below:

Variable	SPSS Name	What Exactly Does This Variable Measure?
1.	*ager* (or whatever name you assigned to recoded *age*)	
2.		
3.		

Step 2: Stating Hypotheses

For each independent variable, state a hypothesis about its relationship with *polviews*. For example, you might hypothesize that people with greater education will be more liberal or that the more religious will be more conservative. You can base your hypotheses on your own experiences or on information you have acquired in your courses.

1.____
2.____
3.____

Step 3: Getting and Reading the Output

To use the **One-Way ANOVA** procedure, follow the instructions in the "Using SPSS: The ANOVA Test" in this chapter.

1. Place *polviews* in the "Dependent List" box and your first independent variable in the **Factor** box.
2. Don't forget to click **Options** and click the box next to **Descriptive** to get summary statistics. Click **OK** and the first test will be conducted.
3. Return to the **One-Way ANOVA** window and click the arrow next to the "Factor" box to move your first independent variable back to the variable list.
4. Move your second independent variable into the "Factor" box and click **OK**.
5. Repeat steps 3 and 4 to test your third independent variable.
6. Print or save your results.

Step 4: Recording Results

Record your results in the following table, using as many rows for each independent variable as necessary.

1. Write the SPSS variable name in the first column.
2. Write the names of the categories of that independent variable in the next column.
3. Record the descriptive statistics in the designated columns.
4. Write the value of the *F* ratio and, in the far right-hand column, indicate whether the results are significant at the 0.05 level. If the value in the "Sig." column of the ANOVA output is less than 0.05, write YES in this column. If the value in the "Sig." column of the ANOVA output is more than 0.05, write NO in this column.
5. For each independent variable, record the grand or overall means, standard deviations, and sample sizes in the row labeled "Totals = ."

Independent Variables	Categories	Mean	Standard Deviation	N	F ratio	Sig. at 0.05 level?
1. *ager*	1.				______	______
	2.					
	3.					
	Totals =					
2. ______________	1.				______	______
	2.					
	3.					
	4.					
	5.					
	6.					
	Totals =					
3. ______________	1.				______	______
	2.					
	3.					
	4.					
	5.					
	6.					
	Totals =					

Step 6: Interpreting Your Results

Summarize your findings. For each test, write:

1. At least one sentence summarizing the test in which you identify the variables being tested, the sample means for each group, *N*, the *F* ratio, and the significance level. In the professional research literature, you might find the results reported as "For a sample of 1417 respondents, there was no significant difference between the average age of Southerners (46.50), Northerners (44.20), Midwesterners (47.80), and Westerners (47.20) ($F = 1.77$, $p > 0.05$)."
2. A sentence relating to your hypotheses. Were they supported? How?

10 Hypothesis Testing IV
Chi Square

LEARNING OBJECTIVES

By the end of this chapter, you will be able to:

1. Identify and explain examples of situations in which the chi square test is appropriate.
2. Explain the structure of a bivariate table and the concept of independence as applied to expected and observed frequencies in a bivariate table.
3. Explain the logic of hypothesis testing in terms of chi square.
4. Perform the chi square test using the five-step model and correctly interpret the results.
5. Explain the limitations of the chi square test and, especially, the difference between statistical significance and importance.
6. Use SPSS to conduct the chi square test of significance.

USING STATISTICS

This chapter presents statistical techniques used to test the statistical significance of the relationships between variables that have been arrayed in a table. If there is a significant difference between the frequencies in the table and what would be expected by random chance, we conclude that the variables are related in the population. This conclusion has a known probability of error, which is often set at the 0.05 level. Examples of situations in which the chi square test is appropriate include the following:

1. A researcher is studying membership in voluntary associations and hypothesizes that unmarried people will be more involved because they have fewer family obligations and more free time. She gathers a random sample from her community and finds a significant relationship between rates of membership and marital status in the sample. She concludes that the variables are related in the population.
2. Medical researchers wonder whether a particular weight loss plan will work equally well with people of all ages. A sample of volunteers is recruited to try the plan. The sample is divided into age groups and, after 2 months, the age groups are compared in terms of their weight loss. Results will be generalized to the community.
3. A researcher wonders about the sources of support for restrictive immigration policies and uses a nationally representative U.S. sample to

test the relationships between a variety of variables (including gender, occupation, and education) and support for the idea that immigration should be reduced. If the researcher finds significant relationships in the sample, he will conclude that those variables are related in the population (all adult Americans).

The **chi square (χ^2) test** has been one of the most frequently used tests of hypothesis in the social sciences—a popularity that is due largely to the fact that the assumptions and requirements in step 1 of the five-step model are easy to satisfy. The test can be conducted with variables measured at the nominal level (the lowest level of measurement) and is **nonparametric**, which means that it requires no assumption at all about the shape of the population or sampling distribution.

Why is it an advantage to have easy-to-satisfy assumptions and requirements? The decision to reject the null hypothesis (step 5) is not specific: It means only that one statement in the model (step 1) *or* the null hypothesis (step 2) is wrong. Usually, of course, we single out the null hypothesis for rejection. The more certain we are of the model stated in step 1, the greater our confidence that the null hypothesis is the faulty assumption. A "weak" or easily satisfied model means that our decision to reject the null hypothesis can be made with greater certainty.

Chi square has also been popular for its flexibility: It can be used with variables that have many categories or scores. For example, in Chapter 8, we tested the significance of the difference in the proportions of black and white citizens who were "highly participatory" in voluntary associations. What if the researcher wished to expand the test to include Hispanic and Asian Americans? The two-sample test would no longer be applicable, but chi square handles the more complex variable easily. Also, unlike the ANOVA test presented in Chapter 9, the chi square test can be conducted with variables at any level of measurement.

Bivariate Tables

Chi square is computed from **bivariate tables**, so called because they display the scores of cases on two different variables at the same time. Bivariate tables are used to test for significant relationships and for other purposes that we will investigate in later chapters. In fact, these tables are very commonly used in research, and a detailed examination of them is in order.

First of all, bivariate tables have (of course) two dimensions. We refer to the horizontal (across) dimension as **rows** and the vertical dimension (up and down) as **columns**. Each column or row represents a score on a variable, and the intersections of the rows and columns (**cells**) represent the combined scores on both variables.

TABLE 10.1 Rates of Participation in Voluntary Associations by Marital Status for 100 Senior Citizens

	Marital Status		
Participation Rates	Married	Unmarried	Totals
High			50
Low			50
Totals	50	50	100

Let's use an example to clarify. Suppose that a researcher is interested in the lives of senior citizens and, specifically, wonders whether their participation in voluntary groups is affected by their marital status. To simplify the analysis, the researcher has confined the sample to people who are presently married or not married (including people who are single and divorced) and has measured involvement in voluntary associations as a simple dichotomy: People have been classified as either high or low in participation.

By convention, we array the independent variable (the variable that is taken to be the cause) in the columns and the dependent variable in the rows. In the example at hand, marital status is the causal variable (the question was "Is participation *affected by* marital status?"), and each column will represent a score on this variable. Each row, on the other hand, will represent a score on level of involvement (high or low). Table 10.1 displays the outline of the bivariate table for a sample of 100 senior citizens.

Note some further details of the table. First, subtotals are included for each column and row. These are called the row or column **marginals**, and in this case they tell us that 50 members of the sample are married and 50 are not married (the column marginals), and that 50 are high in participation and 50 are low (the row marginals).

Second, the total number of cases in the sample ($N = 100$) is reported at the intersection of the row and column marginals. Finally, take careful note of the labeling of the table. Each row and column is identified, and the table has a descriptive title that includes the names of the variables, with the dependent variable listed first. Clear, complete labels and concise titles must be included in *all* tables, graphs, and charts.

As you have noticed, Table 10.1 lacks some crucial information: the numbers in the body of the table. To finish the table, we need to classify the marital status and level of participation of each member of the sample, keep count of how often each combination of scores occurs, and record these numbers in the appropriate cells of the table. Because each variable has two possible scores, four combinations of scores are possible, each corresponding to a cell in the table. For example, married people with high levels of participation would be counted in the upper-left cell, unmarried people with low levels of participation would be counted in the lower-right cell, and so forth. When we are finished counting, each cell will display the number of times each *combination* of scores occurred.

Finally, note that we could expand the table to accommodate variables with more than two scores. If we wished to include people with other marital statuses (widowed, separated, and so forth), we would simply add columns. More elaborate dependent variables could also be easily accommodated. If we had measured participation rates with three categories (e.g., high, moderate, and low), we would simply add a row to the table.

The Logic of Chi Square

The chi square test has several different uses, but we will cover only the *chi square test for independence*. We have encountered the term *independence* in connection with the requirements for the two-sample hypothesis-testing case (Chapter 8) and for the ANOVA test (Chapter 9). In the context of chi square, **independence** refers to the relationship between the variables, not between the samples. Two variables are independent if the classification of a case into a particular category of one variable has no effect on the probability that the case will fall into any particular category of the second variable. For example, the variables in Table 10.1 would be independent of each other if the classification of a case as married or not married has no effect on the classification of the case as high or low on participation. In other words, the variables are independent if level of participation and marital status are completely unrelated to each other.

Consider Table 10.1 again. If the variables were independent, the cell frequencies would be determined solely by random chance and we would find that about half of the married respondents would rank high on participation and half would rank low, just as an honest coin would show heads about 50% of the time. The same pattern would hold for the 50 unmarried respondents and, therefore, each of the four cells would have about 25 cases in it, as illustrated in Table 10.2. This pattern of cell frequencies indicates that marital status has no effect on a person's level of participation. The probability of being classified as high or low would be 0.50 for both marital statuses, and the variables would therefore be independent.

The null hypothesis for the chi square test for independence is that the variables are independent. Under the assumption that the null hypothesis is true, we compute the cell frequencies that we would expect to find if only random chance were operating. These frequencies are called **expected frequencies** (symbolized f_e) and we compare them, cell by cell, with the frequencies actually observed in the table (**observed frequencies**, symbolized f_o). If the null hypothesis is true and the variables are independent, then there should be little

TABLE 10.2 Rates of Participation in Voluntary Associations by Marital Status for 100 Senior Citizens

	Marital Status		
Participation Rates	Married	Unmarried	Totals
High	25	25	50
Low	25	25	50
Totals	50	50	100

difference between the expected and observed frequencies. If the null hypothesis is false, however, there should be large differences between the two. The greater the differences between expected (f_e) and observed (f_o) frequencies, the less likely that the variables are independent and the more likely that we will be able to reject the null hypothesis.

The Computation of Chi Square

To conduct the chi square test—as with all tests of hypothesis—we compute a test statistic, $\boldsymbol{\chi}^2$**(obtained)**, from the sample data and then place that value on the sampling distribution of all possible sample outcomes. Specifically, the χ^2(obtained) will be compared with the value of $\boldsymbol{\chi}^2$**(critical)**, which will be determined by consulting a chi square table (Appendix C) for a particular alpha level and degrees of freedom.

The procedure for calculating chi square is given in Formula 10.1:

FORMULA 10.1

$$\chi^2(\text{obtained}) = \sum \frac{(f_o - f_e)^2}{f_e}$$

where f_o = the cell frequencies observed in the bivariate table
f_e = the cell frequencies that would be expected if the variables were independent

For each cell, subtract the expected frequency from the observed frequency, square the result, and then divide by the expected frequency for that cell. Sum the resultant values for all cells.

This formula requires an expected frequency for each cell in the table. In Table 10.2, the marginals are the same value for all rows and columns, and the expected frequencies are obvious by intuition: $f_e = 25$ for all four cells. In the more usual case, the expected frequencies will not be obvious, the marginals will be unequal, and we must use Formula 10.2 to find the expected frequency for each cell:

FORMULA 10.2

$$f_e = \frac{\text{Row marginal} \times \text{Column marginal}}{N}$$

That is, the expected frequency for any cell is equal to the total number of cases in the row (the row marginal) times the total number of cases in the column (the column marginal) divided by the total number of cases in the table (N).

A Computational Example

An example using Table 10.3 should clarify these procedures. A random sample of 100 social work majors have been classified in terms of whether the Council on Social Work Education has accredited their undergraduate programs (the column, or independent, variable) and whether they were hired in social work positions within three months of graduation (the row, or dependent, variable).

TABLE 10.3 Employment of 100 Social Work Majors by Accreditation Status of Undergraduate Program

	Accreditation Status		
Employment Status	Accredited	Not Accredited	Totals
Working as a social worker	30	10	40
Not working as a social worker	25	35	60
Totals	55	45	100

TABLE 10.4 Expected Frequencies for Table 10.3

	Accreditation Status		
Employment Status	Accredited	Not Accredited	Totals
Working as a social worker	22	18	40
Not working as a social worker	33	27	60
Totals	55	45	100

TABLE 10.5 Computational Table for Table 10.3

(1)	(2)	(3)	(4)	(5)
f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
30	22	8	64	2.91
10	18	−8	64	3.56
25	33	−8	64	1.94
35	27	8	64	2.37
100	100	0		10.78

Beginning with the upper-left cell (graduates of accredited programs who are working as social workers), the expected frequency for this cell, using Formula 10.2, is (40 × 55)/100, or 22. For the other cell in this row (graduates of non-accredited programs who are working as social workers), the expected frequency is (40 × 45)/100, or 18. For the two cells in the bottom row, the expected frequencies are (60 × 55)/100, or 33, and (60 × 45)/100, or 27, respectively. The expected frequencies for all four cells are displayed in Table 10.4.

Note that the row and column marginals as well as the total number of cases in Table 10.4 are exactly the same as those in Table 10.3. The row and column marginals for the expected frequencies must *always* equal those of the observed frequencies, a relationship that provides a convenient way of checking your arithmetic to this point.

The value for chi square for these data can now be found by solving Formula 10.1. It will be helpful to use a computing table, such as Table 10.5, to organize the several steps required to compute chi square. The table lists the observed frequencies (f_o) in column 1 in order from the upper-left cell to the lower-right cell, moving from left to right across the table and from top to bottom. Column 2 lists the expected frequencies (f_e) *in exactly the same order*. Double-check to make sure you have listed the cell frequencies in the same order for both of these columns.

ONE STEP AT A TIME — **Computing Expected Frequencies (f_e)**

Step	***Operation***
Use Formula 10.2:	
1.	Start with the upper-left-hand cell of the bivariate table. Multiply the row marginal for this cell by the column marginal.
2.	Divide the quantity you found in step 1 by N. The result is the expected frequency for that cell.
3.	Record this f_e in a separate table (see Table 10.4) or in the second column of a computational table (see Table 10.5).
4.	Repeat steps 1–3 for each cell.
5.	*Double check to make sure that you are using the correct row and column marginal when computing the expected frequencies.*

The next step is to subtract the expected frequency from the observed frequency for each cell and list these values in column 3. To complete column 4, square the value in column 3 and then, in column 5, divide the column 4 value by the expected frequency for that cell. Finally, add up column 5. The sum of this column is obtained chi square. For Table 10.3, χ^2(obtained) = 10.78.

Note that the totals for columns 1 (f_o) and 2 (f_e) are exactly the same. This will always be the case, and if the totals do not match, you have made a computational error, probably in the calculation of the expected frequencies. Also note that the sum of column 3 will always be 0, another convenient way to check your math to this point.

This sample value for chi square must still be tested for its significance. *(For practice in computing chi square, see any problem at the end of this chapter, starting with the smallest tables.)*

ONE STEP AT A TIME — **Computing Chi Square**

Step	***Operation***
Use Formula 10.1 and a computing table like Table 10.5:	
1.	List the observed frequencies (f_o) in column 1. The total for this column is the number of cases (N).
2.	List the expected frequencies (f_e) in column 2. The total for this column *must* equal the total for column 1 (N).
3.	For each cell, subtract the expected frequency (f_e) from the observed frequency (f_o), and list these values in the third column of the computational table, labeled ($f_o - f_e$). Find the total for this column. If this total is not zero, you have made a mistake and need to check your computations.
4.	Square each of the values in the third column of the table, and record the result in the fourth column, labeled $(f_o - f_e)^2$.
5.	Divide each value in column 4 by the expected frequency for that cell, and record the result in the fifth column, labeled $(f_o - f_e)^2/f_e$.
6.	Find the total for the fifth column. This value is χ^2(obtained).

The Chi Square Test for Independence

We are now ready to conduct the chi square test for independence. Recall that if the variables are independent of each other, the score of a case on one variable will have no relationship with its score on the other variable. As always, the five-step model for significance testing will provide the framework for organizing our decision making. The data presented in Table 10.3 will serve as our example.

Step 1. Making Assumptions and Meeting Test Requirements. Note that, because the test is nonparametric, we make no assumptions at all about the shape of the sampling distribution.

Model: Independent random samples
Level of measurement is nominal

Step 2. Stating the Null Hypothesis. The null hypothesis states that the two variables are independent. If the null hypothesis is true, the differences between the observed and expected frequencies will be small. As usual, the research hypothesis directly contradicts the null hypothesis. Thus, if we reject H_0, the research hypothesis will be supported.

H_0: The two variables are independent.
(H_1: The two variables are dependent.)

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region. The sampling distribution of sample chi squares, unlike the Z and t distributions, is positively skewed, with higher values of sample chi squares in the upper tail of the distribution (to the right). Thus, with the chi square test, the critical region is established in the upper tail of the sampling distribution.

Values for χ^2(critical) are given in Appendix C. This table is similar to the t table, with alpha levels arrayed across the top and degrees of freedom down the side. However, with chi square, the degrees of freedom (df) are found by Formula 10.3:

FORMULA 10.3

$$df = (r - 1)(c - 1)$$

A table with two rows and two columns (a 2×2 table) has one degree of freedom regardless of the number of cases in the sample.[1] A table with two rows and

[1]Degrees of freedom are the number of values in a distribution that are free to vary for any particular statistic. A 2×2 table has one degree of freedom because, for a given set of marginals, once one cell frequency is determined, all other cell frequencies are fixed and no longer free to vary. In Table 10.3, for example, if any cell frequency is known, all others are determined. If the upper-left cell is known to be 30, the remaining cell in that row must be 10, because there are 40 cases total in the row and $40 - 30 = 10$. Once the frequencies of the cells in the top row are established, cell frequencies for the bottom row are determined by subtraction from the column marginals. *NOTE:* This relationship can be used to quickly compute expected frequencies. For example, in a 2×2 table, only one expected frequency needs to be computed. The f_e's for all other cells can then be found by subtraction.

three columns would have (2 − 1)(3 − 1), or two degrees of freedom. Our sample problem involves a 2 × 2 table with one degree of freedom, so if we set alpha at 0.05, the critical chi square score would be 3.841. Summarizing these decisions, we have

$$\text{Sampling distribution} = \chi^2 \text{ distribution}$$
$$\text{Alpha} = 0.05$$
$$\text{Degrees of freedom} = 1$$
$$\chi^2(\text{critical}) = 3.841$$

Step 4. Computing the Test Statistic. The computation of chi square was introduced in the previous section. As you recall, we had

$$\chi^2(\text{obtained}) = 10.78$$

Step 5. Making a Decision and Interpreting the Results of the Test. Comparing the test statistic with the critical region,

$$\chi^2(\text{obtained}) = 10.78$$
$$\chi^2(\text{critical}) = 3.841$$

we see that the test statistic falls in the critical region and, therefore, we reject the null hypothesis of independence. The pattern of cell frequencies observed in Table 10.3 is unlikely to have occurred by chance alone. The variables are dependent. Specifically, based on these sample data, the probability of securing employment in the field of social work is dependent on the accreditation status of the program. *(For practice in conducting and interpreting the chi square test for independence, see any problem at the end of the chapter.)*

We should be clear about exactly what the chi square test does and does not tell us. A significant chi square means that the variables are (probably) dependent on each other in the population. In terms of our example, this means that there is a relationship between accreditation and whether a person is working as a social worker. But what exactly is the relationship between the variables? Which type of graduate is more likely to find work in the profession? To make this determination, we must perform an additional calculation.

Column Percentages

We can figure out how the independent variable (accreditation status in our example) is affecting the dependent variable (employment as a social worker) by computing **column percentages** or by calculating percentages within each column of the bivariate table. This procedure is analogous to calculating percentages for frequency distributions (see Chapter 2).

TABLE 10.6 Column Percentages for Table 10.6

	Accreditation Status		
Employment Status	Accredited	Not Accredited	Totals
Working as a social worker	54.55%	22.22%	40.00%
Not working as a social worker	45.45%	77.78%	60.00%
Totals	100.00% (55)	100.00% (45)	100.00%

To calculate column percentages, divide each cell frequency by the total number of cases in the column (the column marginal) and multiply the result by 100. For Table 10.3, starting in the upper-left cell, we see that there are 30 cases in this cell and 55 cases in the column. So 30 of the 55 graduates of accredited programs are working as social workers. The column percentage for this cell is therefore $(30/55) \times 100 = 54.55\%$. For the lower-left cell, the column percent is $(25/55) \times 100 = 45.45\%$. For the two cells in the right-hand column (graduates of nonaccredited programs), the column percentages are $(10/45) \times 100 = 22.22\%$ and $(35/45) \times 100 = 77.78\%$. Table 10.6 displays all column percentages for Table 10.3.

Column percentages make the relationship between the variables more obvious, and we can see easily from Table 10.6 that it's the students from accredited programs who are more likely to be working as social workers. Nearly 55% of these students are working as social workers versus about 22% of the students from nonaccredited programs. We already know that this relationship is significant (unlikely to be caused by random chance), and now, with the aid of column percentages, we know how the two variables are related. According to these results, graduates from accredited programs have a decided advantage in securing social work jobs.

ONE STEP AT A TIME Computing Column Percentages

Step	*Operation*
1.	Start with the upper-left-hand cell. Divide the cell frequency (the number of cases in the cell) by the total number of cases in that column. Multiply the result by 100 to convert to a percentage.
2.	Move down one cell and repeat step 1. Continue moving down the column, cell by cell, until you have converted all cell frequencies into percentages.
3.	Move to the next column. Start with the cell in the top row and repeat step 1 (making sure that you use the correct column total in the denominator of the fraction).
4.	Continue moving down the second column until you have converted all cell frequencies to percentages.
5.	Continue these operations, moving from column to column one at a time, until you have converted all cell frequencies to percentages.

Applying Statistics 10.1 The Chi Square Test

Do men and women vary in their opinions about cohabitation? A random sample of 47 males and females has been rated as high or low in their support for "living together." The results are as follows:

Support for Cohabitation	Gender: Male	Gender: Female	Totals
High	15	5	20
Low	10	17	27
Totals	25	22	47

The expected frequencies are calculated cell-by-cell using Formula 10.2. Double-check to make sure that you have the correct row and column marginal when calculating expected frequencies.

Support for Cohabitation	Gender: Male	Gender: Female	Totals
High	10.64	9.36	20
Low	14.36	12.64	27
Totals	25	22	47

Organize the calculations of chi square with a computational table.

(1) f_o	(2) f_e	(3) $f_o - f_e$	(4) $(f_o - f_e)^2$	(5) $(f_o - f_e)^2/f_e$
15	10.64	4.36	19.01	1.79
5	9.36	−4.36	19.01	2.03
10	14.36	−4.36	19.01	1.32
17	12.64	4.36	19.01	1.50
47	47.00	0.00	χ^2(obtained) = 6.64	

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Independent random samples
Level of measurement is nominal

Step 2. Stating the Null Hypothesis.

H_0: The two variables are independent.
(H_1: The two variables are dependent.)

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region.

Sampling distribution = χ^2 distribution
Alpha = 0.05
Degrees of freedom = 1
χ^2(critical) = 3.841

Step 4. Computing the Test Statistic.

$$\chi^2(\text{obtained}) = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2(\text{obtained}) = 6.64$$

Step 5. Making a Decision and Interpreting the Results of the Test. The obtained chi square is in the critical region, so we reject the null hypothesis of independence. For this sample there is a statistically significant relationship between gender and support for cohabitation.

Which gender is more supportive of cohabitation? We can answer this question with column percentages:

Support for Cohabitation	Gender: Male	Gender: Female	Totals
High	60.00%	22.73%	42.55%
Low	40.00%	77.27%	57.45%
Totals	100.00%	100.00%	100.00%

The column percentages show that 60% of males in this sample are highly supportive versus only 23% of females. We have already concluded that the relationship is significant, and now we know the pattern of the relationship: Males are more supportive.

Let's highlight two points in summary:

1. Chi square is a test of statistical significance. It tests the null hypothesis that the variables are independent in the population. If we reject the null hypothesis, we are concluding, with a known probability of error (determined by the alpha level), that the variables are dependent on each other in the population. By itself, however, chi square does not tell us the nature of the relationship.

STATISTICS
IN EVERYDAY LIFE

There seems to be a gender gap in national elections in the United States: Women tend to favor the Democratic presidential candidates while men tend to prefer the Republican. The gender gap is depicted in the table below for the 2012 presidential election. These results are based on a sample of 2551 voters and are statistically significant at well below the 0.05 level.

Why do you suppose the gender gap exists? What additional information would you like to have to explore the gap further?

Presidential Preference by Gender

Candidates	Male	Female
B. Obama (Democrat)	45.96%	56.02%
M. Romney (Republican)	54.04%	43.98%
Totals	100.00% (1214)	100.00% (1337)

Source: Jones, Jeffrey. 2102. "Gender Gap in 2012 Vote Is Largest in Gallup's History." Accessed from http://www.gallup.com/poll/158588/gender-gap-2012-vote-largest-gallup-history.aspx

2. Column percentages show us how the independent variable affects the dependent variable. In our example, they show that it's the graduates of accredited programs who are more likely to find work as social workers. We explore column percentages more extensively when we discuss bivariate association in Chapter 11.

The Chi Square Test: An Additional Example

To this point, we have confined our attention to 2 × 2 tables, that is, tables with two rows and two columns. For purposes of illustration, we will work through the computational routines and decision-making process for a larger table. As you will see, larger tables require more computations (because they have more cells), but in all other essentials they are dealt with in the same way as the 2 × 2 table.

A researcher is concerned with the possible effects of marital status on the academic progress of college students. Do married students, with their extra burden of family responsibilities, suffer academically as compared to unmarried students? Is academic performance dependent on marital status? A random sample of 453 students is gathered, and each student is classified by marital status and as a good, average, or poor student, based on grade-point average (GPA). Results are presented in Table 10.7.

For the top-left cell (married students with good GPAs), the expected frequency would be (160 × 175)/453, or 61.8. For the other cell in this row, expected frequency is (160 × 278)/453, or 98.2. In similar fashion, all expected frequencies are computed (being very careful to use the correct row and column marginals) and displayed in Table 10.8.

TABLE 10.7 Grade-Point Average (GPA) by Marital Status

	Marital Status		
GPA	Married	Unmarried	Totals
Good	70	90	160
Average	60	110	170
Poor	45	78	123
Totals	175	278	453

TABLE 10.8 Expected Frequencies for Table 10.7

	Marital Status		
GPA	Married	Unmarried	Totals
Good	61.8	98.2	160
Average	65.7	104.3	170
Poor	47.5	75.5	123
Totals	175.0	278.0	453

TABLE 10.9 Computational Table for Table 10.7

(1)	(2)	(3)	(4)	(5)
f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
70	61.8	8.2	67.24	1.09
90	98.2	−8.2	67.24	0.69
60	65.7	−5.7	32.49	0.50
110	104.3	5.7	32.49	0.31
45	47.5	−2.5	6.25	0.13
78	75.5	2.5	6.25	0.08
453	453.0	0.0		χ^2(obtained) = 2.80

The next step is to solve the formula for χ^2(obtained), being very careful that we are using the proper f_o's and f_e's for each cell. Once again, we will use a computational table (Table 10.9) to organize the calculations and then test the obtained chi square for its statistical significance. Remember that obtained chi square is equal to the total of column 5.

The value of the obtained chi square (2.80) can now be tested for its significance.

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Independent random samples
Level of measurement is nominal

Step 2. Stating the Null Hypothesis.

$$H_0\text{: The two variables are independent.}$$
$$(H_1\text{: The two variables are dependent.})$$

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region.

$$\text{Sampling distribution} = \chi^2 \text{ distribution}$$
$$\text{Alpha} = 0.05$$
$$\text{Degrees of freedom} = (r-1)(c-1) = (3-1)(2-1) = 2$$
$$\chi^2(\text{critical}) = 5.991$$

Step 4. Computing the Test Statistic.

$$\chi^2(\text{obtained}) = 2.80$$

Step 5. Making a Decision and Interpreting the Results of the Test. The test statistic, χ^2(obtained) = 2.80, does not fall in the critical region, which, for alpha = 0.05, $df = 2$, begins at χ^2(critical) of 5.991. Therefore, we fail to reject the null hypothesis. The observed frequencies are not significantly different from the frequencies we would expect to find if the variables were independent and only random chance were operating. Based on these sample results, we can conclude that the academic performance of college students is not dependent on their marital status. Because we failed to reject the null hypothesis, we will not examine column percentages as we did for Table 10.3.

Using SPSS: Crosstabs and Chi Square

In this installment of "Using SPSS," we will use a procedure called **Crosstabs**, which produces bivariate tables, column percentages, and chi square. This procedure also will be used in Chapter 11.

To provide an illustration, we will test to see if there is a statistically significant relationship between gender (*sex*)—the independent variable—and support for abortion "if the family has very low income and can't afford any more children" (*abpoor*), the dependent variable. Is support for legal abortion in this circumstance dependent on gender?

To begin,

1. Click the SPSS icon on your desktop.
2. Load the *GSS2012* database.
   a. Find the **File** command on the far left of the menu bar and then click **File → Open → Data**.
   b. Find the *GSS2012* database and click the file name.
3. From the main menu bar, click **Analyze**, **Descriptive Statistics**, and **Crosstabs**.

4. The "Crosstabs" dialog box appears with the variables listed in a box on the left.
   a. Highlight the dependent variable (*abpoor*) and click the arrow to move the variable name into the "Rows" box.
   b. Next, find the independent variable (*sex*) and move it into the "Columns" box. SPSS will process all combinations of variables in the row and column boxes at one time.
5. Click the **Statistics** button at the right of the window and click the box next to Chi-square.
6. Return to the "Crosstabs" window and click the **Cells** button and select "column" in the "Percentages" box. This will generate column percentages for the table.
7. Return to the "Crosstabs" window and click **OK** and the following bivariate table will be produced:

LOW INCOME—CAN'T AFFORD MORE CHILDREN * RESPONDENTS SEX Cross-tabulation

			RESPONDENTS SEX		
			MALE	FEMALE	Total
LOW INCOME—CAN'T AFFORD MORE CHILDREN	YES	Count	192	216	408
		% within RESPONDENTS SEX	45.9%	44.3%	45.0%
	NO	Count	226	272	498
		% within RESPONDENTS SEX	54.1%	55.7%	55.0%
Total		Count	418	488	906
		% within RESPONDENTS SEX	100.0%	100.0%	100.0%

Each cell displays the cell frequency and column percentage. For example, starting with the upper-left-hand cell, there were 192 respondents who were male and who said "yes" to *abpoor*. These were 45.9% of all men in the sample. For the women in the sample, 216 (44.3%) also supported legal abortion in this circumstance. The column percentages are quite similar, which indicates that *sex* and *abpoor* do not have a significant relationship.

The next output block is labeled "Chi Square Tests," and, to conserve space, we will not reproduce it here. There is a lot of information in this block but we only need to read the top line, or "Pearson Chi-Square," which reports a chi square (obtained) of 0.254 with one degree of freedom. The exact significance of this chi square, reported in the column labeled "Asymp. Sig (2-sided)," is 0.614. This is far above the standard indicator of a significant result (alpha = 0.05), so we may conclude, as we saw with the column percentages, that there is no statistically significant relationship between these variables. Support for legal abortion is not dependent on gender.

The Limitations of the Chi Square Test

Like any other test, chi square has limitations, and you should be aware of several potential difficulties. First, chi square becomes difficult to interpret when the variables have many categories. For example, two variables with five categories each would generate a 5 × 5 table with 25 cells, which would probably be too many combinations of scores to be easily absorbed or understood. As a very rough rule of thumb, the chi square test is easiest to interpret and understand when both variables have four or fewer scores.

Two further limitations of the test are related to sample size. When sample size is small, we can no longer assume that the sampling distribution of all possible sample outcomes is accurately described by the chi square distribution. For chi square, a small sample is defined as one where a high percentage of the cells have expected frequencies (f_e) of 5 or less. Various rules of thumb have been developed to help the researcher decide what constitutes a "high percentage of cells." Probably the safest course is to take corrective action whenever *any* of the cells have expected frequencies of 5 or less.

For 2 × 2 tables, the value of χ^2 (obtained) can be adjusted by applying Yates' correction for continuity, the formula for which is

FORMULA 10.4

$$\chi_c^2 = \sum \frac{(|f_o - f_e| - 0.5)^2}{f_e}$$

where χ_c^2 = corrected chi square

$|f_o - f_e|$ = the absolute values of the difference between observed and expected frequency for each cell. (*NOTE:* Absolute values ignore plus and minus signs).

The correction factor is applied by reducing the absolute value of the term $(f_o - f_e)$ by 0.5 before squaring the difference and dividing by the expected frequency for the cell.

For tables larger than 2 × 2, there is no correction formula for computing χ^2(obtained) for small samples. It may be possible to combine some of the categories of the variables and thereby increase cell sizes. For example, in Table 10.7, the "average" and "poor" categories might be combined. This would reduce the table size to 2 × 2 and maximize the number of cases in the bottom row.

Obviously, categories should be combined only when it is sensible to do so. In other words, meaningful distinctions—like those between average and poor students—should not be erased merely to conform to the requirements of a statistical test.

When you feel that categories cannot be combined to build up cell frequencies and the percentage of cells with expected frequencies of 5 or less is small, it is probably justifiable to continue with the uncorrected chi square test as long as the results are regarded with a suitable amount of caution.

A second potential problem related to sample size occurs with large samples. I pointed out in Chapter 8 that all tests of hypothesis are sensitive to sample size. That is, the probability of rejecting the null hypothesis increases as the number of cases increases, regardless of any other factor. Chi square is especially sensitive to sample size: The value of χ^2(obtained) will increase at the same rate as sample size. If N is doubled, for example, the value of χ^2(obtained) will double. Thus, larger samples may lead to a decision to reject the null hypothesis when the actual relationship is trivial. *(For an illustration of this principle, see problem 10.14.)*

You should be aware of the relationship between sample size and the value of chi square because it once again raises the distinction between statistical significance and theoretical importance. On one hand, tests of significance play a crucial role in research. When we are working with random samples, we must know whether our research results could have been produced by mere random chance.

On the other hand, like any other statistical technique, tests of hypothesis are limited in the range of questions they can answer. Specifically, these tests will tell us whether our results are statistically significant or not. They will not necessarily tell us whether the results are important in any other sense. To deal more directly with questions of importance, we must use an additional set of statistical techniques called *measures of association*. We previewed these techniques in this chapter when we computed column percentages, and measures of association are the subject of Part III of this text.

STATISTICS

IN EVERYDAY LIFE

Who Uses Social Networking Sites?

It will come as no surprise to learn that there is a relationship between age and use of the Internet. Does this relationship apply to use of social networking sites such as Facebook? The table below is based on a nationally representative sample. To conserve space, only percentages are reported but chi square was calculated using cell frequencies.

How Often Do You Visit Social Networking Sites?	Age Groups			
	18–29	30–49	50 and older	Totals
Daily or More	68.2%	57.0%	44.2%	57.5%
Weekly	26.4%	29.1%	37.2%	30.2%
Monthly or Less	5.4%	13.9%	18.6%	12.3%
Totals	100.0% (1167)	100.0% (1469)	100.0% (880)	100.0% (687)

$\chi^2 = 147.82$, $df = 4$, $p < 0.05$

This is a statistically significant relationship and, as expected, younger people are most likely to visit social networking sites. Notice, however, that these sites are popular with all three age groups and that almost 88% of Americans use them at least weekly.

Source: Pew Research Center. 2012. The survey is available in SPSS format at http://pewinternet.org/Shared-Content/Data-Sets/2012/February-2012--Search-Social-Networking-Sites-and-Politics.aspx

Applying Statistics 10.2 Who Reads Horoscopes?

Is there a relationship between interest in astrology and age? Which age group is most likely to consult their horoscope? A random sample of Americans has been asked if they ever read their horoscope or personal astrology report. The results are as follows:

	Age Groups				
Ever Read Horoscope?	18–34	35–50	51–64	65–89	Totals
Yes	54.4%	49.5%	47.9%	45.6%	49.4%
No	45.6%	50.5%	52.1%	54.4%	50.6%
Totals	100.0% (241)	100.0% (277)	100.0% (207)	100.0% (215)	100.0% (996)

$\chi^2 = 3.86$ $df = 3$, $p > 0.05$

Note that results have been reported as in the professional research literature. To conserve space, the table reports only column percentages and the results of the chi square test have been reported beneath the table. The column marginal have been reported in parentheses and the cell frequencies can be reconstructed from that information.

To provide an additional demonstration of the computational routine for chi square, the cell frequencies are as follows:

	Age Groups				
Ever Read Horoscope?	18–34	35–50	51–64	65–89	Totals
Yes	131	137	126	98	492
No	110	140	137	117	504
Totals	241	277	263	215	996

and the computational table would be as follows:

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
131	119.05	11.95	142.85	1.20
137	136.83	0.17	0.03	0.00
126	129.92	–3.92	15.33	0.12
98	106.20	–8.20	67.32	0.63
110	121.95	–11.95	142.85	1.17
140	140.17	–0.17	0.03	0.00
137	133.08	3.92	15.33	0.12
117	108.80	8.20	67.32	0.62
996	996.00	0.00	χ^2(obtained) =	3.86

What can we conclude? With $df = 3$ and $\alpha = 0.05$, χ^2 (critical) is 7.815. The relationship is not significant and the column percentages are quite similar. Interest in astrology is not dependent on age (although the youngest seem slightly more likely to read their horoscope).

Source: 2012 General Social Survey

SUMMARY

1. The chi square test for independence is appropriate for situations in which the variables of interest have been organized into table format. The null hypothesis is that the variables are independent, or that the classification of a case into a particular category on one variable has no effect on the probability that the case will be classified into any particular category of the second variable.
2. Because chi square is nonparametric and requires only nominally measured variables, its model assumptions are easily satisfied. Furthermore, it is computed from bivariate tables, in which the number of rows and columns can be easily expanded, so the chi square test can be used in many situations in which other tests are irrelevant or difficult to apply.
3. In the chi square test, we first find the frequencies that would appear in the cells if the variables were independent (f_e) and then compare those frequencies, cell by cell, with the frequencies actually observed in the

cells (f_o). If the null hypothesis is true, expected and observed frequencies should be quite close in value. The greater the difference between the observed and expected frequencies, the greater the possibility of rejecting the null hypothesis.

4. The chi square test has several important limitations. It is often difficult to interpret when tables have many (more than four or five) dimensions. Also, as sample size (N) decreases, the chi square test becomes less trustworthy, and corrective action may be required. Finally, with very large samples, we may declare relatively trivial relationships to be statistically significant. As is the case with all tests of hypothesis, statistical significance is not the same thing as "importance" in any other sense. As a general rule, statistical significance is a necessary but not sufficient condition for theoretical or practical importance.

SUMMARY OF FORMULAS

FORMULA 10.1 Chi square (obtained):
$$\chi^2(\text{obtained}) = \sum \frac{(f_o - f_e)^2}{f_e}$$

FORMULA 10.2 Expected frequencies:
$$f_e = \frac{\text{Row marginal} \times \text{Column marginal}}{N}$$

FORMULA 10.3 Degrees of freedom, bivariate tables:
$$df = (r - 1)(c - 1)$$

FORMULA 10.4 Yates' correction for continuity:
$$\chi_c^2 = \sum \frac{(|f_o - f_e| - 0.5)^2}{f_e}$$

GLOSSARY

Bivariate table. A table that displays the joint frequency distributions of two variables.

Cells. The cross-classification categories of the variables in a bivariate table.

Chi square (χ^2) test. A nonparametric test of hypothesis for variables that have been organized into a bivariate table.

Column. The vertical dimension of a bivariate table. By convention, each column represents a score on the independent variable.

Column Percentages. Percentages computed within each column of a bivariate table.

Expected frequency (f_e). The cell frequencies that would be expected in a bivariate table if the variables were independent.

Independence. The null hypothesis in the chi square test. Two variables are independent if, for all cases, the classification of a case on one variable has no effect on the probability that the case will be classified in any particular category of the second variable.

Marginals. The row and column subtotals in a bivariate table.

Nonparametric test. A "distribution-free" test. These tests do not assume a normal sampling distribution.

Observed frequency (f_o). The cell frequencies actually observed in a bivariate table.

Row. The horizontal dimension of a bivariate table, conventionally representing a score on the dependent variable.

χ^2(critical). The score on the sampling distribution of all possible sample chi squares that marks the beginning of the critical region.

χ^2(obtained). The test statistic computed from sample results.

PROBLEMS

10.1 The following four 2 × 2 tables, presented without labels or titles, provide simplified opportunities to practice your computational skills. Calculate chi square for each.

HINTS:

- *Calculate the expected frequencies for each cell with Formula 10.2. Double-check to make sure you are using the correct row and column marginals for each cell.*

- *It may be helpful to record the expected frequencies in table format—see Tables 10.4 and 10.8.*
- *Use a computational table to organize the calculation for Formula 10.1—see Tables 10.5 and 10.9.*
- *Follow the step-by-step instructions.*
- *Double-check to make sure you are using the correct values for each cell.*

a.

20	25	45
25	20	45
45	45	90

b.

10	15	25
20	30	50
30	45	75

c.

25	15	40
30	30	60
55	45	100

d.

20	45	65
15	20	35
35	65	100

10.2 **PS** As mentioned in this chapter, many analysts have noted a "gender gap" in U.S. elections, with women more likely to vote for the Democratic candidate. A sample of university faculty has been asked about their political party preference. Do their responses indicate a significant relationship between gender and party preference for this group?

Party Preference	Gender		Totals
	Male	Female	
Democrat	10	15	25
Republican	15	10	25
Totals	25	25	50

a. Is there a statistically significant relationship between gender and party preference?
b. Compute column percentages for the table to clarify the pattern of the relationship. Which gender is more likely to prefer the Democrats?

10.3 **SW** A local politician is concerned that a program for the homeless in her city is discriminating against blacks and other minorities. The following data were taken from a random sample of black and white homeless people.

Received Services?	Race		Totals
	Black	White	
Yes	6	7	13
No	4	9	13
Totals	10	16	26

a. Is there a statistically significant relationship between race and whether the person has received services from the program?
b. Compute column percentages for the table to clarify the pattern of the relationship. Which group was more likely to get services?

10.4 **SOC** A sample of 25 cities have been classified as high or low on their homicide rates and on the number of handguns sold within the city limits.

a. Is there a relationship between these variables? Do cities with higher homicide rates have significantly higher handgun sales? Explain your results in a sentence or two, citing both the results of the chi square test and the patterns of the column percentages.
b. The table is constructed assuming that homicide rate is the independent variable ("higher homicide rates cause people to buy more guns"). Can you think of an argument that would make handgun sales the independent variable? Which is the more logical argument? Why?

Volume of Handgun Sales	Homicide Rate		Totals
	Low	High	
High	8	5	13
Low	4	8	12
Totals	12	13	25

10.5 **PA** Is there a relationship between salary levels and unionization for public employees? The following data represent this relationship for fire departments in a random sample of 100 cities of roughly the same size. Salary data have been dichotomized at the median. Explain your results in a sentence or two, citing both the results of the chi square test and the patterns of the column percentages.

Salary	Status		Totals
	Union	Non-Union	
High	21	29	50
Low	14	36	50
Totals	35	65	100

10.6 **CJ** A local judge has been allowing some individuals convicted of "driving under the influence" to work in a hospital emergency room as an alternative to fines, suspensions, and other penalties. A random sample of offenders has been drawn. Do participants in this program have lower rates of recidivism for this offense? Explain your results in a sentence or two, citing both the results of the chi square test and the patterns of the column percentages.

	Status		
Recidivist?	Nonparticipant	Participant	Totals
Yes	60	123	183
No	55	108	163
Totals	115	231	346

10.7 *SOC* The state Department of Education has rated a sample of local school systems for compliance with state-mandated guidelines for quality.

a. Is there a statistically significant relationship between the quality of a school system and the affluence of the community as measured by per capita income?

b. Compute column percentages. Are high- or low-income communities more likely to have high-quality schools?

	Per Capita income		
Quality	Low	High	Totals
Low	16	8	24
High	9	17	26
Totals	25	25	50

10.8 *SOC* A program of pet therapy has been running at a local nursing home. Are the participants in the program significantly more alert and responsive than nonparticipants? (Remember to compute column percentages to identify the pattern of the relationship.)

	Participation?		
Alertness	Yes	No	Totals
Low	11	18	29
High	23	15	38
Totals	34	33	67

10.9 *SOC* Is there a statistically significant relationship between length of marriage and satisfaction with marriage? The necessary information has been collected from a random sample of 100 respondents drawn from a local community. Write a sentence or two explaining your findings, including the results of the chi square test and pattern of column percentages.

	Length of Marriage (in years)			
Satisfaction	Less than 5	5–10	More than 10	Totals
Low	10	20	20	50
High	20	20	10	50
Totals	30	40	30	100

10.10 *PS* Is there a relationship between college class standing and political orientation? Are juniors and seniors significantly different from freshmen and sophomores on this variable? The following table reports the relationship between these two variables for a random sample of 267 college students. Summarize your findings in a sentence or two, including the results of the chi square test and the pattern of column percentages. Which group is more likely to be liberal? Which group is more likely to be conservative?

	Class Standing		
Political Views	Freshman and Sophomore	Junior and Senior	Totals
Liberal	40	43	83
Moderate	50	50	100
Conservative	44	40	84
Totals	134	133	267

10.11 *SOC* At a large urban college, about half of the students live off campus in various arrangements, and the other half live on campus. Is academic performance dependent on living arrangements? Write a sentence or two explaining your findings, including the results of the chi square test and pattern of column percentages.

	Residential Status			
GPA	Off Campus with Roommates	Off Campus with Parents	On Campus	Totals
Low	22	20	48	90
Moderate	36	40	54	130
High	32	10	38	80
Totals	90	70	140	300

10.12 SOC An urban planning commissioner has created a database describing a sample of the neighborhoods in her city and has developed a scale by which each area can be rated for the "quality of life" (this includes measures of pollution, noise, open space, services available, and so on). She has also asked samples of residents of these areas about their level of satisfaction with their neighborhoods. Is there statistically significant agreement between the commissioner's objective ratings of quality of life and the respondents' self-reports of satisfaction? Write a sentence or two explaining your findings, including the results of the chi square test and pattern of column percentages.

	Quality of Life			
Satisfaction	Low	Moderate	High	Totals
Low	21	15	6	42
Moderate	12	25	21	58
High	8	17	32	57
Totals	41	57	59	157

10.13 SOC Does support for the legalization of marijuana vary by region of the country? The table displays the relationship between the two variables for a random sample of 1020 adult citizens. Is the relationship significant? Describe the pattern of the relationship, using column percentages. Which region is most in support of legalization?

	Region				
Legalize?	North	Midwest	South	West	Totals
Yes	60	65	42	78	245
No	245	200	180	150	775
Totals	305	265	222	228	1020

10.14 SOC A researcher is concerned with the relationship between attitudes toward violence and violent behavior. Do attitudes "cause" behavior? Do people who have positive attitudes toward violence have higher rates of violent behavior? A pretest was conducted on 70 respondents; among other things, the respondents were asked, "Have you been involved in a violent incident of any kind over the past six months?" The researcher established the following relationship:

	Attitude Toward Violence		
Involvement	Favorable	Unfavorable	Totals
Yes	16	19	35
No	14	21	35
Totals	30	40	70

The chi square for this table is 0.23, which is not significant at the 0.05 level (confirm this with your own calculations). Undeterred by this result, the researcher proceeded with the project and gathered a random sample of 7000. In terms of percentage distributions, the results for the full sample were exactly the same as for the pretest:

	Attitude Toward Violence		
Involvement	Favorable	Unfavorable	Totals
Yes	1600	1900	3500
No	1400	2100	3500
Totals	3000	4000	7000

However, the chi square obtained is a very healthy 23.4 (confirm with your own calculations). Why is the full-sample chi square significant when the pretest was not? What happened? Do you think that the second result is important?

10.15 PS A random sample of 748 voters in a large city was asked how they voted in the presidential election of 2012. Calculate chi square and the column percentages for each table below and write a brief report describing the significance of the relationships as well as the patterns you observe.

a. Presidential preference and gender

	Gender		
Preference	Male	Female	Totals
Romney	165	173	338
Obama	200	210	410
Totals	365	383	748

b. Presidential preference and race/ethnicity

	Race/Ethnicity			
Preference	White	Black	Latino	Totals
Romney	289	5	44	338
Obama	249	95	66	410
Totals	538	100	110	748

c. Presidential preference by education

Preference	Education				Totals
	Less than HS	HS Graduate	College Graduate	Post-Graduate Degree	
Romney	30	180	118	10	338
Obama	35	120	218	37	410
Totals	65	300	336	47	748

d. Presidential preference by religion

Preference	Religion					Totals
	Protestant	Catholic	Jewish	None	Other	
Romney	165	110	10	28	25	338
Obama	245	55	20	60	30	410
Totals	410	165	30	88	55	748

Statistical Analysis Using SPSS

10.16 SOC Is there a relationship between gender and fear of walking alone at night (*fear*), support for capital punishment (*cappun*), and job satisfaction (*satjob*)? Are men or women more likely to say that there is a place nearby where they are afraid to walk at night? Which gender is more supportive of the death penalty? Are men or women more satisfied with their jobs (*satjob*)? We will use the **Crosstabs** procedure, with chi square and column percentages to explore this relationship.

- Click the SPSS icon on your desktop.
- Load the *GSS2012* database.
- Click **Analyze Descriptive Crosstabs.**
- Find *fear, cappun,* and *satjob* in the list of variables and move them into the "Rows" box.
- Find *sex* in the list of variables and move it into the "Columns" box.
- Click the **Statistics** button and select **Chi-square**.
- Return to the "Crosstabs" window, click the **Cells** button, and select **Column** in the "Percentages" box.
- Return to the "Crosstabs" window and click **OK**. The results of all three tests will be printed to the "SPSS output" window.

Inspect the column percentages, chi square, and the "Asymp. Sig (2-sided)." Report these results in a few sentences.

10.17 SOC Is there a relationship between education (*degree*) and support for the theory of evolution (*evolved*) and the "big bang" theory of the creation of the universe (*bigbang*)? Would you predict that the more educated are more supportive of these theories? Why?

Use the **Crosstabs** procedure, following the instructions in Problem 10.16. Education is the independent variable, so place *degree* in the "Columns" box and *evolved* and *bigbang* in the "Rows" box. Report the results in a few sentences that take account of the significance of chi square and the column percentages.

YOU ARE THE RESEARCHER

Understanding Political Beliefs

Two projects are presented here, and I urge you to apply your understanding of the chi square test to both. Both projects use the *GSS2012* database. In the first project, you will examine the sources of people's beliefs about some of the most hotly debated topics in U.S. society: assisted suicide, gay marriage, and immigration. In the second, you will compare various independent variables to see which has the most significant relationship with your chosen dependent variable.

Using Crosstabs

In these projects you will use **Crosstabs**, the procedure introduced in the "Using SPSS" installment and in problems 10.16 and 10.17. To run **Crosstabs** for both projects,

- Click **Analyze**, **Descriptive Statistics**, and **Crosstabs.**
- Place your dependent variable(s) in the "Rows" box.
- Place your independent variable(s) in the "Columns" box.
- Click the **Statistics** button, select **Chi-square**, and return to the "Crosstabs" window.
- Click the **Cells** button, select **column** in the "Percentages" box, and return to the "Crosstabs" window. Click **OK**.

Project 1: Explaining Beliefs

In this project, you will analyze beliefs about assisted suicide (*letdie1*), gay marriage (*marhomo*), and immigration (*letin1*). You will select an independent variable, use SPSS to generate chi squares and column percentages, and analyze and interpret your results.

Step 1: Choose an Independent Variable

Select an independent variable that seems likely to be an important cause of people's attitudes about these issues. *Be sure to select an independent variable that has only two to five categories!* If you would like to use an independent variable with more than five scores, use the **recode** command to reduce the number of categories. You might consider gender, level of education (use degree), religion, or age (the recoded version—see Chapter 9) as possible independent variables, but there are many others. Record the variable name and state exactly what the variable measures in the box below.

SPSS Name	What Exactly Does This Variable Measure?

Step 2: Stating Hypotheses

State hypotheses about the relationships you expect to find between your independent variable and each dependent variable. State these hypotheses in terms of which category of the independent variable you expect to be associated with which category of the dependent variable (for example, "I expect that men will be more supportive of the legal right to an abortion for any reason").

Dependent Variable	Hypothesis
letdie1	
marhomo	
letin1	

Step 3: Running Crosstabs

All dependent variables can be placed in the "Rows:" box together. Move the independent variable you selected into the "Columns:" box. Don't forget to click the **Statistics** button to get chi square and the **Cells** button for column percentages.

Step 4: Recording Results

There will be a lot of output and it will be helpful to summarize your results in the following table. You will find the value of chi square, the degrees of freedom, and the statistical significance listed in the first line ("Pearson Chi Square") of the "Chi-Square Tests" block of output. If the value in the column labeled "Asymp. Sig (2-sided)" is less than 0.05, write "Yes" in the right-hand column of the table below.

Dep. Var.	Chi Square	Degrees of Freedom	Significant at 0.05?
letdie1			
marhomo			
letin1			

Step 5: Analyzing and Interpreting Results

Write a short summary of results of each test, in which you

1. Identify the variables, the value and significance of chi square, *N*, and the pattern (if any) of the column percentages. In the professional research literature, you might find the results reported as "For a sample of 949 respondents, there was no significant relationship between gender and support for abortion (chi square = 2.863, $df = 1$, $p > 0.05$). About 46% of the men supported the legal right to an abortion vs. about 44% of the women."
2. Explain whether your hypotheses were supported.

Project 2: Exploring the Impact of Various Independent Variables

In this project, you will examine the relative ability of different independent variables to explain or account for a single dependent variable. You will again use the **Crosstabs** procedure in SPSS to generate chi squares and column percentages and use the value of alpha to judge which independent variable has the most significant relationship with your dependent variable.

Step 1: Choosing Variables

Select a dependent variable. You may use any of the variables from Project 1 in this chapter or select a new dependent variable. *Be sure that your dependent variable has no more than five values or scores.* Use the **recode** command as necessary to reduce the number of categories. Good choices for dependent variables include any measure of attitudes or opinions. *Do not* select characteristics like race, sex, or religion as dependent variables.

Select three independent variables that seem likely to be important causes of the dependent variable you selected. *Be sure your independent variable has no more than five values or scores.* Use the **recode** command as necessary to reduce the number of categories. You might consider gender, level of education (use *degree*), religion, or age (the recoded version—see Chapter 9) as possibilities, but there are many others.

Record the variable names and state exactly what each variable measures in the table below.

SPSS Name	What Exactly Does This Variable Measure?
Dependent Variable	
Independent Variables	

Step 2: Stating Hypotheses

State hypotheses about the relationships you expect to find between your independent variables and the dependent variable. State these hypotheses in terms of which category of the independent variable you expect to be associated with which category of the dependent variable (for example, "I expect that men will be more supportive of the legal right to an abortion for any reason").

Independent Variables (SPSS Names)	Hypothesis
1. ________________	
2. ________________	
3. ________________	

Step 3: Running Crosstabs

Place your dependent variable in the "Rows:" box and all three of your independent variables in the "Columns:" box. Click the **Statistics** button to get chi square and the **Cells** button for column percentages.

Step 4: Recording Results

There will be a lot of output and it will be helpful to summarize your results in the following table. You will find the value of chi square, the degrees of freedom, and the statistical significance listed in the first line ("Pearson Chi Square") of the "Chi-Square Tests" block of output. If the value in the column labeled "Asymp. Sig (2-sided)" is less than 0.05, write "Yes" in the right-hand column of the table below.

Independent Variable	Chi Square	Degrees of Freedom	Significant at 0.05 level?
1. ________________			
2. ________________			
3. ________________			

Step 5: Analyzing and Interpreting Results

Write a short summary of results of each test, in which you

1. Identify the variables being tested, the value and significance of chi square, *N*, and the pattern (if any) of the column percentages.
2. Explain whether your hypotheses were supported.
3. Explain which independent variable had the most significant relationship (lowest value in the "Asymp. Sig 2-tailed" column) with your dependent variable.

Part III Bivariate Measures of Association

Part III covers *measures of association*, statistics that are extremely useful in scientific research and commonly reported in the professional literature. These statistics provide, in a single number, an indication of the strength and (if appropriate) the direction of a bivariate association.

It is important to remember the difference between statistical significance, covered in Part II, and association, the topic of this part. Tests for statistical significance answer a specific question: Were the differences or relationships observed in the sample caused by mere random chance? Measures of association address a different set of questions: How strong is the relationship between the variables? What is the direction or pattern of the relationship?

Thus, the information supplied by measures of association complements that supplied by tests of significance, and association and significance are two different things. It is most satisfying to find results that are *both* statistically significant and strong, but it is common to find mixed or ambiguous results: relationships that are statistically significant but weak, not statistically significant but strong, and so forth.

Chapter 11 introduces the basic ideas behind the analysis of association in terms of bivariate tables and column percentages and presents measures of association for nominal- and ordinal-level variables. Chapter 12 presents Pearson's r, the most important measure of association and the only one designed for interval-ratio-level variables.

11 Bivariate Association for Nominal- and Ordinal-Level Variables

LEARNING OBJECTIVES

By the end of this chapter, you will be able to:

1. Use measures of association to describe and analyze the *importance* (vs. statistical significance) of a bivariate relationship.
2. Define association in the context of bivariate tables.
3. Answer three questions about a bivariate relationship:
   a. Does it exist?
   b. How strong is it?
   c. What is its pattern or direction?
4. Assess the association of variables in a bivariate table by
   a. Calculating and interpreting column percentages.
   b. Computing and interpreting an appropriate measure of association.
5. Compute and interpret Spearman's rho, a measure of association for "continuous" ordinal-level variables.
6. Use SPSS to calculate column percentages for a bivariate table and produce nominal and ordinal measures of association.

USING STATISTICS

The statistical techniques presented in this chapter are used to measure the association between nominal- and ordinal-level variables. Examples of situations in which these techniques are useful include:

1. Is there a relationship between satisfaction and productivity for workers? Are busy workers happy workers?
2. Does the crime rate increase when the economy falters? By how much?
3. Are there relationships between religiosity and support for traditional gender roles, gay marriage, or the death penalty? How strong are these relationships? What patterns do they follow?
4. Do burnout and loss of morale in teachers increase as years of experience increase?

As we have seen over the past several chapters, tests of statistical significance are extremely important in social science research. When we work with random samples rather than populations, tests of significance are indispensable for dealing with the possibility that our research results are the products of mere random chance.

Often, however, significance testing is only the first step in the analysis of research results. These tests do have limitations, and statistical significance is not necessarily the same thing as relevance or importance. Furthermore, all tests of significance are affected by sample size: Tests performed on large samples may result in decisions to reject the null hypothesis when, in fact, the differences or relationships are quite minor.

Beginning with this chapter, we will be working with **measures of association**, statistics which can help us assess the importance or strength of relationships, and test the power and validity of our theories. Scientific theories are almost always stated in cause-and-effect terms: "variable *X* causes variable *Y*." As an example, recall our discussion of the contact hypothesis in Chapter 1. In that theory, the causal (or independent) variable was equal status contacts between groups, and the effect (or dependent) variable was prejudice. The theory asserts that involvement in equal-status-contact situations *causes* prejudice to decline. Measures of association help us trace causal relationships among variables, and they are our most important and powerful statistical tools for documenting, measuring, and analyzing cause-and-effect relationships.

I must point out that measures of association, as useful as they are, cannot *prove* that two variables are causally related. Even if there is a strong (and significant) statistical association between two variables, we cannot necessarily conclude that one variable is a cause of the other. We will explore causation in more detail in Part IV, but for now you should keep in mind that causation and association are two different things. We can use a statistical association between variables as evidence for a causal relationship, but association by itself is not proof that a causal relationship exists.

Another important use for measures of association is prediction. If two variables are associated, we can predict the score of a case on one variable from its score on the other variable. For example, if equal-status contacts and prejudice are associated, we can predict that people who have experienced many such contacts will be less prejudiced than those who have had few or no contacts.

In this chapter you will be introduced to the concept of **association** between variables in the context of bivariate tables. You will also learn how to use percentages and several different measures of association to analyze a bivariate relationship. By the end of this chapter, you will have an array of statistical tools that you can use to analyze the strength and pattern of an association between two variables.

Association and Bivariate Tables

Most generally, two variables are said to be associated if the distribution of one of them changes under the various categories or scores of the other. For example, suppose that an industrial sociologist was concerned with the relationship

TABLE 11.1 Productivity by Job Satisfaction (frequencies)

	Job Satisfaction (*X*)			
Productivity (*Y*)	Low	Moderate	High	Totals
Low	30	21	7	58
Moderate	20	25	18	63
High	10	15	27	52
Totals	60	61	52	173

between job satisfaction and productivity for assembly-line workers. If these two variables are associated, then scores on productivity will change under the different conditions of satisfaction. Highly satisfied workers will have different productivity scores than workers who are low on satisfaction, and levels of productivity will vary by levels of satisfaction.

We can clarify this relationship with a bivariate table. As you recall from Chapter 10, bivariate tables display the scores of cases on two different variables. By convention, the **independent**, or ***X*****, variable** (that is, the variable that is assumed to be the cause) is arrayed in the columns, and the **dependent**, or ***Y*****, variable** in the rows.[1] That is, each column of the table (the vertical dimension) represents a score or category of the independent variable (*X*), and each row (the horizontal dimension) represents a score or category of the dependent variable (*Y*).

Table 11.1 displays a relationship between productivity and job satisfaction for a fictitious sample of 173 factory workers. We focus on the columns to see whether there is an association between the variables. The "within-column" frequency distributions are called the **conditional distributions of *Y***, because they display the distribution of scores on the dependent variable for each condition (or score) of the independent variable.

We can inspect the conditional distributions of *Y* in Table 11.1 by going from column to column. If the conditional distributions change, the variables are associated. The left-hand column shows the conditional distribution of productivity (*Y*) for workers who were low on job satisfaction (*X*): 30 of the 60 were low on productivity, 20 were moderately productive, and 10 were high on productivity. The middle column shows the distribution of *Y* for moderately satisfied workers (21 of the 61 were low on productivity, 25 were moderate, and 15 were high) and the right-hand column displays the distribution of *Y* for highly satisfied workers (only 7 were low on productivity, 18 were moderate, and 27 were high).

So far, we know that the variables in Table 11.1 are associated (because the conditional distributions of *Y* change across the scores of *X*). In the next section, we extend the analysis and show how to develop additional very useful information about the relationship.

[1]In the material that follows, we will often, for the sake of brevity, refer to the independent variable as *X* and the dependent variable as *Y*.

Three Characteristics of Bivariate Associations

To fully investigate a bivariate association, we seek answers to three questions:

1. Does an association exist?
2. How strong is the association?
3. What is the pattern or direction of the association?

We will consider each question separately.

Does an Association Exist? We have already seen that we can detect an association by observing the conditional distributions of Y in a table. In Table 11.1, we know that the two variables are associated to some extent because the conditional distributions of productivity (Y) are different across the various categories of satisfaction (X).

Comparisons from column to column in Table 11.1 are relatively easy to make because the column totals are roughly equal. This will not usually be the case, and it is helpful to compute column percentages (see Chapter 10) to control for varying column totals and to make the association more visible.

The general procedure for detecting association with bivariate tables is to compute percentages within the columns (vertically, or down each column) and then to compare column to column across the table (horizontally, or from column to column). Table 11.2 presents column percentages calculated from the data in Table 11.1. Note that this table reports the row and column marginals in parentheses. Tables with percentages are usually easier to read because changes in the conditional distributions of Y are easier to detect.

In Table 11.2, we can see that the largest cell changes position from column to column. For workers who are low on satisfaction, the single largest cell is in the top row (low on productivity). For the middle column (moderate on satisfaction), the largest cell is in the middle row (moderate on productivity), and, for the right-hand column (high on satisfaction), it is in the bottom row (high on productivity). Even a cursory glance at Table 11.2 reinforces our conclusion that an association does exist between these two variables.

If two variables are not associated, then the conditional distributions of Y will not change across the columns. The distribution of Y would be the same for each condition of X. Table 11.3 illustrates a "perfect nonassociation" between

TABLE 11.2 Productivity by Job Satisfaction (percentages)

	Job Satisfaction (X)			
Productivity (Y)	Low	Moderate	High	Totals
Low	50.0%	34.4%	13.5%	33.5% (58)
Moderate	33.3%	41.0%	34.6%	36.4% (63)
High	16.7%	24.6%	51.9%	30.1% (52)
Totals	100.0% (60)	100.0% (61)	100.0% (52)	100.0% (173)

TABLE 11.3 Productivity by Height (an illustration of no association)

	Height (*X*)		
Productivity (*Y*)	Short	Medium	Tall
Low	33.3%	33.3%	33.3%
Moderate	33.3%	33.3%	33.3%
High	33.3%	33.3%	33.3%
Totals	99.9%	99.9%	99.9%

height and productivity. Table 11.3 is only one of many patterns that indicate "no association." The important point is that the conditional distributions of *Y* are the same. Levels of productivity do not change at all for the various heights; therefore, no association exists between these variables. (*For practice in computing column percentages, see problems 11.1 to 11.9*)

How Strong Is the Association? Once we know there is an association between two variables, we need to know how strong it is. This is essentially a matter of determining the amount of change in the conditional distributions of *Y*. At one extreme, of course, is the case of "no association," where the conditional distributions of *Y* do not change at all (see Table 11.3). At the other extreme is a perfect association, the strongest possible relationship.

A perfect association exists between two variables if each value of the dependent variable is associated with one and only one value of the independent variable.[2] In a bivariate table, the variables would have a perfect relationship if all cases in each column are located in a single cell and there is no variation in *Y* for a given value of *X* (see Table 11.4).

A perfect relationship would be taken as very strong evidence of a causal relationship between the variables, at least for the sample at hand. In fact, the results presented in Table 11.4 would indicate that, for this sample, height is the

TABLE 11.4 Productivity by Height (an illustration of perfect association)

	Height (*X*)		
Productivity (*Y*)	Short	Medium	Tall
Low	0.0%	0.0%	100.0%
Moderate	0.0%	100.0%	0.0%
High	100.0%	0.0%	0.0%
Totals	100.0%	100.0%	100.0%

[2]Each measure of association to be introduced in this and the following chapters incorporates its own definition of a "perfect association," and these definitions vary somewhat, depending on the specific logic and mathematics of the statistic. That is, for different measures computed from the same table, some measures will possibly indicate perfect relationships when others will not. We will note these variations in the mathematical definitions of a perfect association at the appropriate times.

sole cause of productivity. Also, in the case of a perfect relationship, predictions from one variable to the other could be made without error. If we know that a particular worker is short, for example, we could be sure that he or she is highly productive.

Of course, the huge majority of relationships will fall somewhere between the extremes of no association and perfect association, and we need to develop some way of describing these intermediate relationships consistently and meaningfully. For example, Tables 11.1 and 11.2 show an association between productivity and job satisfaction. How could this relationship be described in terms of strength? How close is the relationship to perfect? How far away is it from no association?

We rely on measures of association to give us precise, objective indicators of the strength of a relationship. Virtually all of these statistics are designed so that they have a lower limit of 0.00 and an upper limit of 1.00 ($\pm$1.00 for ordinal and interval-ratio measures of association). A measure of 0.00 indicates no association between the variables (the conditional distributions of Y do not vary), and a measure of 1.00 ($\pm$1.00 in the case of ordinal and interval-ratio measures) indicates a perfect relationship. The exact meaning of values between 0.00 and 1.00 varies from measure to measure, but, for all measures, the closer the value is to 1.00, the stronger the relationship (the greater the change in the conditional distributions of Y).

A Simple Measure of the Strength of an Association. We begin to consider the many measures of association later in this chapter. At this point, we consider the **maximum difference**, a more informal way of assessing the strength of a relationship based on comparing column percentages across the table. To find the maximum difference, compute the column percentages as usual and then skim the table across each of the rows to find the largest difference—in any row—between column percentages. For example, the largest difference in column percentages in Table 11.2 is in the top row between the "Low" column and the "High" column: 50.0% − 13.5% = 36.5%. The maximum difference in the middle row is between "moderates" and "lows" (41.0% − 33.3% = 7.7%), and in the bottom row it is between "highs" and "lows" (51.9% − 16.7% = 35.2%). Both of these values are less than the maximum difference in the top row.

Once you have found the maximum difference in the table, you can use the scale presented in Table 11.5 to describe the strength of the relationship. For instance, we could describe the relationship between productivity and job satisfaction in Table 11.2 as strong.

TABLE 11.5 The Relationship Between the Maximum Difference and the Strength of the Relationship

Maximum Difference	Strength
If the maximum difference is	*The strength of the relationship is*
between 0 and 10 percentage points	Weak
between 11 and 30 percentage points	Moderate
more than 30 percentage points	Strong

Be aware that the relationships between the maximum difference and the descriptive terms (weak, moderate, and strong) in Table 11.5 are arbitrary and approximate. We will get more precise and useful information when we compute and analyze a measure of association.

Also, the maximum difference is easiest to find and most useful for small tables. In large tables, with many (say, more than three) columns and rows, it can be cumbersome to find the high and low percentages, and it is advisable to consider only measures of association as indicators of the strength for these tables. Finally, note that the maximum difference is based on only two values (the high and low column percentages within any row). Like the range (see Chapter 4), this statistic can give a misleading impression of the overall strength of the relationship. *(The maximum difference can be computed for any of the tables in problems 11.1 to11.9.)*

What Is the Pattern and/or the Direction of the Association? To find the pattern of the association, we need to ascertain which values or categories of one variable are associated with which values or categories of the other. We have already remarked on the pattern of the relationship between productivity and satisfaction. Table 11.2 indicates that low scores on satisfaction are associated with low scores on productivity, moderate satisfaction with moderate productivity, and high satisfaction with high productivity.

When one or both variables in a bivariate table are measured at the nominal level, we can discuss the pattern of the relationship *only*.[3] However, when both variables are at least ordinal in level of measurement, the association may also be described in terms of direction. The direction of the association can be either positive or negative.

In a **positive association**, the variables vary in the same direction: They will increase or decrease together. High scores on one variable are associated with high scores on the other variable, and low scores on one variable are associated with low scores on the other. The cases will tend to fall along a diagonal from the upper left of a bivariate table to the lower right. Table 11.6 displays, with fictitious data, a positive relationship between education and use of public libraries.

TABLE 11.6 Library Use by Education (an illustration of a positive relationship)

	Education (*X*)		
Library Use (*Y*)	Low	Moderate	High
Low	60%	20%	10%
Moderate	30%	60%	30%
High	10%	20%	60%
Totals	100%	100%	100%

[3]The scores of variables measured at the nominal level have no numerical order (by definition). Therefore, associations including nominal-level variables, while they may have a pattern, cannot have a direction.

TABLE 11.7 Amount of Television Viewing by Education (an illustration of a negative relationship)

	Education (*X*)		
TV Viewing (*Y*)	Low	Moderate	High
Low	10%	20%	60%
Moderate	30%	60%	30%
High	60%	20%	10%
Totals	100%	100%	100%

As education increases (as you move from left to right across the table), library use also increases (the percentage of "high" users increases). The association between job satisfaction and productivity, as displayed in Tables 11.1 and 11.2, is also a positive association.

In a **negative association**, the variables vary in opposite directions. High scores on one variable are associated with low scores on the other, and increases in one variable are accompanied by decreases in the other. Table 11.7 displays a negative relationship, again with fictitious data, between education and television viewing. The amount of television viewing decreases as education increases. In other words, as you move from left to right across the top of the table (as education increases), the percentage of heavy viewers decreases.

Measures of association for ordinal and interval-ratio variables are designed so that they will be positive for positive associations and negative for negative associations. Thus, a measure of association preceded by a plus sign indicates a positive relationship between the two variables, with the value $+1.00$ indicating a perfect positive relationship. A negative sign indicates a negative relationship, with -1.00 indicating a perfect negative relationship. *(For practice in determining the pattern of an association, see problems 11.1 to 11.4, 11.5a and c, and 11.13. For practice in determining the direction of a relationship, see problems 11.5b, 11.6 to 11.9, and 11.14.)*

Association and Statistical Significance

Tables 11.1 and 11.2 show that there is a strong, positive relationship between job satisfaction and productivity. If the 173 cases in the table were a random sample, it would be necessary to also test the relationship for statistical significance. In fact, the obtained chi square for the table is 24.2, which is significant at the 0.05 level. This means that job satisfaction and productivity are (probably) related in the population.

Usually, we want to argue for a causal relationship between our variables, and we are in the strongest position to do so when we find an association that is *both* strong and statistically significant. Remember, however, that association and significance are two different things and it is possible for a relationship to be strong but insignificant, significant but weak, and so forth.

Association and Causation

In Tables 11.1 and 11.2, the association between productivity and satisfaction is strong, positive, and significant, and it might be tempting to conclude that satisfaction *causes* productivity ("happy workers are busy workers"). This conclusion is consistent with the evidence, but remember that association, by itself, does not prove causation. In fact, although we have assumed that satisfaction is the independent variable in this relationship, we could have argued the reverse causal relationship ("busy workers are happy workers"). The tables are consistent with both causal arguments.

Applying Statistics 11.1 Reading Bivariate Tables and Avoiding Errors of Interpretation

Percentages are among the humblest of statistics, but they can be misread and misused. To avoid problems when reading bivariate tables, always

- arrange tables with the independent variable in the columns and
- calculate percentages within the columns, so that the percentages in each column add up to 100.

The relationship between the variables, if any, can be ascertained be comparing the column percentages. Remember to "calculate percentages *down* and compare *across*."

To illustrate, consider the relationship between support for affirmative action and racial or ethnic group, shown in the table below for a nationally representative sample. How do we know that race or ethnicity is the independent variable and should be placed in the columns? In this case, the decision is straightforward. A person's group membership could be a cause of his or her attitudes and opinions, but the reverse cannot be true: A person's opinion cannot be a cause of his or her race or ethnicity.

Support for Affirmative Action by Racial or Ethnic Group

	Racial or Ethnic Group			
Affirmative Action	White	Black	Hispanic	Totals
Support	139 (15.0%)	76 (43.2%)	8 (14.8%)	223 (19.3%)
Oppose	788 (85.0%)	100 (56.8%)	46 (85.2%)	934 (80.7%)
Totals	927 (100.0%)	176 (100.0%)	54 (100.0%)	1157 (100.0%)
Chi square = 76.26, $df = 2$, $p < 0.05$				

Source: 2012 General Social Survey.

The pattern in the table supports the view that these variables are associated. The relationship is statistically significant and the conditional distributions change across the table. Blacks are the most supportive of affirmative action, and Hispanics and whites are about equally opposed. The maximum difference between the columns is about 28 percentage points, which means that the relationship is moderate to strong.

These results should seem straightforward. But, what if we had mistakenly switched the position of the variables and placed support for affirmative action in the columns? If we then computed percentages down and compared across, we could make some serious errors of analysis. Consider the incorrectly calculated percentages in the following table:

(*continued*)

Racial or Ethnic Group by Support of Affirmative Action (with percentages calculated incorrectly)

	Affirmative Action		
Racial or Ethnic Group	Support	Oppose	Totals
White	62.3%	84.4%	80.1%
Black	34.1%	10.7%	15.2%
Hispanic	3.6%	4.9%	4.7%
Totals	100.0%	100.0%	100.0%

If we looked *only* at the "Support" column of this table (as people sometimes do), we might conclude that whites are more supportive of affirmative action than other groups. But the "Oppose" column shows that whites are also the huge majority (84.4%) of those who *oppose* the policy. How can this be?

This table incorrectly treats race or ethnicity as the dependent variable and the patterns in the table simply reflect the fact that whites vastly outnumber the other groups: There are 5 times as many whites as blacks in the sample and 17 times as many whites as Hispanics. A quick, unreflective look at this table could easily lead to false conclusions about the relationship between these variables.

Even professional researchers sometimes compute percentages in the wrong direction or ask a question about the relationship incorrectly, and you should always check to make sure that the analysis agrees with the patterns in the table.

You should also be aware that tables are sometimes constructed with the independent variable in the rows. In such a case, you should compute percentages within the rows and compare from row to row.

STATISTICS IN EVERYDAY LIFE

Discrimination in Football?

Even after professional football was integrated in the middle of the 20th century, racial discrimination continued on a number of levels. Perhaps the most obvious example was that, for many years, there was a perfect relationship between race and the quarterback position. This central leadership role was reserved for whites, a practice that reflected widespread stereotypes regarding the intelligence and decision-making ability of blacks.

Today, racial discrimination in professional sports is much weaker than in the past and minority-group athletes face much less discrimination in their competition for any position, including quarterback. In fact, African Americans are vastly overrepresented in professional football: They are 66% of all NFL players but only 13% of the general population. However, on opening day of the 2013 season, only 9 of the 32 starting quarterbacks (or 28%) were African American,* a pattern that may reflect lingering—if more subtle—prejudices and stereotypes. What data would you need to investigate these patterns further?

**Source:* Lapchik, R., 2013. *The 2013 Racial and Gender Report Card: National Football League*. Available at http://www.tidesport.org/RGRC/2013/2013_NFL_RGRC.pdf

Measures of Association: Introduction

Column percentages provide important information about bivariate associations and should always be computed and analyzed. However, they can be awkward and cumbersome to use, particularly for larger tables. In contrast, measures of association summarize the overall strength (and direction for ordinal-level variables) of a bivariate association in a single number, a much more compact and convenient format for interpretation and discussion.

There are many measures of association, but we will focus on a few of the most widely used. We will cover these statistics by the level of measurement for which they are most appropriate, first covering measures appropriate for nominal variables and then dealing with measures for ordinal-level variables. Finally, in Chapter 12, we will consider Pearson's *r*, a measure of association or correlation for interval-ratio-level variables. To analyze relationships for variables that are at different levels (for example, one nominal-level variable and one ordinal-level variable), we generally use the measures of association appropriate for the lower of the two levels of measurement.

Measures of Association for Nominal Variables

There are two types of measures of association for nominal-level variables, one based on chi square (see Chapter 10) and the other on a logic called "proportional reduction in error" or PRE. We consider these two categories separately.

Chi Square–Based Measures of Association: Phi and Cramer's *V*

There are several chi square–based measures of association, but we will consider only two. The first, called **phi** ($\boldsymbol{\phi}$), is appropriate for 2 × 2 tables (that is, tables with two rows and two columns) and is easy to calculate. As shown in Formula 11.1, phi is simply the square root of chi square divided by *N*.

FORMULA 11.1

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

For tables larger than 2 × 2, the upper limit of phi can exceed 1.00 and this can make the statistic difficult to interpret. For larger tables, a more general form of the statistic called Cramer's *V* is commonly used.

FORMULA 11.2

$$V = \sqrt{\frac{\chi^2}{N(\min r - 1, c - 1)}}$$

where $(\min r - 1, c - 1)$ = the minimum value of number of rows (r) minus 1 or number of columns (c) minus 1

To calculate *V*, follow the instructions in the "One Step at a Time" box. Cramer's *V* has an upper limit of 1.00 for a table of any size and will be the same value as phi if the table has either two rows or two columns.

Both phi and Cramer's *V* can be interpreted as indices that measure the strength of the association between two variables. Table 11.8 provides guidelines for the interpretation of the strength of nominal-level measures of association. As with Table 11.5, the descriptive terms are intended as general guides only.

To illustrate the computation of *V*, consider Table 11.9, which shows the relationship between membership in various types of student organizations and academic

TABLE 11.8 Guidelines for Interpreting the Strength of the Relationship for Nominal-Level Measures of Association

Measure of Association	Strength
If the value is	*The strength of the relationship is*
between 0.00 and 0.10	Weak
between 0.11 and 0.30	Moderate
greater than 0.30	Strong

TABLE 11.9 Academic Achievement (GPA) by Club Membership

	Membership			
Achievement	Greek	Other	None	Totals
Low	4 (17.4%)	4 (15.4%)	17 (65.4%)	25 (33.3%)
Moderate	15 (65.2%)	6 (23.1%)	4 (15.4%)	25 (33.3%)
High	4 (17.4%)	16 (61.5%)	5 (19.2%)	25 (33.3%)
Totals	23 (100.0%)	26 (100.0%)	26 (100.0%)	75 (99.9%)

achievement for a sample of college students. The obtained chi square for this table is 32.14, which is significant at the 0.05 level. The table has the same number of rows and columns, so we may use either $(r - 1)$ or $(c - 1)$ in the denominator. In either case, the value of the denominator is N multiplied by $(3 - 1)$, or 2. Cramer's V is

$$V = \sqrt{\frac{\chi^2}{N(\min r - 1, c - 1)}}$$

$$V = \sqrt{\frac{32.14}{(75)(2)}}$$

$$V = \sqrt{\frac{32.14}{150}}$$

$$V = \sqrt{0.21}$$

$$V = 0.46$$

A Cramer's V of 0.46 indicates a strong association between club membership and academic achievement. From the column percentages, we can see that members of fraternities and sororities tend to be moderate, members of other organizations tend to be high, and nonmembers tend to be low in academic achievement.

ONE STEP AT A TIME Computing Cramer's *V*

Step	*Operation*
1.	Find the number of rows (r) and number of columns (c) in the table. Subtract 1 from the lesser of these two numbers to find ($\min r - 1, c - 1$).
2.	Multiply the value you found in step 1 by N.
3.	Divide the value of chi square by the value you found in step 2.
4.	Take the square root of the quantity you found in step 3. The resulting value is V.

ONE STEP AT A TIME	Interpreting Cramer's *V*
Step	***Operation***
1.	See Table 11.8 for general guidelines.
2.	Use the column percentages to identify the pattern of the relationship.

The Limitations of Phi and Cramer's *V*

One limitation of these measures is that they are only general indicators of the strength of the relationship. Of course, the closer the measure is to 0.00, the weaker the relationship and the closer to 1.00, the stronger the relationship. Values between 0.00 and 1.00 can be described as weak, moderate, or strong, according to the terms displayed in Table 11.8, but they have no direct or meaningful interpretation. On the other hand, phi and *V* are easy to calculate (once the value of chi square has been obtained) and have been commonly used indicators of the importance of an association.[4] *(For practice in computing Cramer's V, see problems 11.1 to 11.4, 11.5a and c, and 11.13. Remember that for tables with either two rows or two columns, phi and Cramer's V will have the same value.)*

A Measure of Association Based on Proportional Reduction in Error

There are several nominal-level measures of association based on the logic of **proportional reduction in error (PRE)**. We will focus on **lambda** (symbolized by the Greek letter λ).

The logic of these measures is based on two different predictions about the scores of cases on the dependent variable. In one prediction, we ignore information about the independent variable and, in the second, we take the independent variable into account. Lambda tells us how much knowledge of the independent variable improves our predictions.

For lambda, we first predict the category into which each case will fall on the dependent variable (Y) while ignoring the independent variable (X). Since we would be predicting blindly in this case, we would often predict the value of a case on the dependent variable incorrectly.

For the second prediction, we take the independent variable into account. If the two variables are associated, this additional information will reduce our errors of prediction—that is, we should misclassify fewer cases. Furthermore, the stronger the association between the variables, the greater the reduction in errors. In the case of a perfect association, we would make no errors at all when predicting the score on Y from the score on X. When there is no association between the variables, on the other hand, knowledge of the independent variable will not improve the accuracy of our predictions.

[4]Two other chi square–based measures of association, T^2 and C (the contingency coefficient) are sometimes reported in the literature. Both of these measures have serious limitations. T^2 has an upper limit of 1.00 only for tables with an equal number of rows and columns, and the upper limit of C varies, depending on the dimensions of the table. These characteristics make these measures more difficult to interpret and thus less useful than phi or Cramer's V.

TABLE 11.10 Height by Gender (fictitious data)

	Gender		
Height	Male	Female	Totals
Tall	44	8	52
Short	6	42	48
Totals	50	50	100

To calculate lambda, we find two quantities: first, the number of prediction errors made while ignoring the independent variable (E_1) and, second, the number of prediction errors made while taking the independent variable into account (E_2). These two sums are then compared to derive the statistic. The "One Step at a Time" box states the computational procedures for tables of any size.

An example will clarify the computation of lambda. We will illustrate with Table 11.10, a 2×2 table that displays a relationship between height and gender.

First, we find E_1, the number of prediction errors made while ignoring the independent variable (gender), by working with the row marginals and making a prediction for *all* cases. We could predict that all subjects are tall, a prediction that will result in 48 errors. That is, for this prediction, all 100 cases would be placed in the top row. Only 52 of the cases actually belong in this row, so this prediction would result in $(100 - 52)$, or 48, errors. If we had predicted that all subjects were short, on the other hand, we would have made 52 errors $(100 - 48 = 52)$. We will take the *lesser* of these two numbers and refer to this quantity as E_1, for the number of errors made while ignoring the independent variable. So $E_1 = 48$.

Next, we predict the score on Y (height) again, but this time we take X (gender) into account by moving from column to column. Because each column is a category of X, we thus use X in making our predictions. For each column, we predict that all cases will fall in the cell with the largest frequency. For males, we predict that all 50 cases will be tall and make 6 errors $(50 - 44 = 6)$. For females, our prediction is that all are short, and 8 errors will be made. By moving from column to column, we have made a total of 14 errors of prediction, a quantity we will label E_2 ($E_2 = 6 + 8 = 14$).

If the variables are associated, we will make fewer errors under the second procedure than under the first and E_2 will be smaller than E_1. In our example, E_2 is indeed less than E_1, so gender and height are associated. Our errors were reduced from 48 to only 14. To find the *proportional* reduction in error, use Formula 11.3:

FORMULA 11.3

$$\lambda = \frac{E_1 - E_2}{E_1}$$

For the sample problem, the value of lambda would be

$$\lambda = \frac{E_1 - E_2}{E_1}$$

$$\lambda = \frac{48 - 14}{48}$$

$$\lambda = \frac{34}{48}$$

$$\lambda = 0.71$$

ONE STEP AT A TIME **Computing Lambda**

Step	*Operation*
These instructions are stated in general terms and will work with tables of any size.	
1.	To find E_1, subtract the largest row subtotal (marginal) from N.
2.	To find E_2, start with the far left-hand column and subtract the largest cell frequency in the column from the column total. Repeat this step for all columns in the table.
3.	Add up all the values you found in step 2. The result is E_2.
4.	Subtract E_2 from E_1.
5.	Divide the quantity you found in step 4 by E_1. The result is lambda.

ONE STEP AT A TIME **Interpreting Lambda**

Step	*Operation*
1.	See Table 11.8 for general guidelines.
2.	For a PRE interpretation, multiply the value of lambda by 100. This percentage expresses the extent to which predictions of the dependent variable are improved by taking the independent variable into account.
3.	Use the column percentages to identify the pattern of the relationship.

The value of lambda ranges from 0.00 to 1.00. Of course, a value of 0.00 means that the variables are not associated at all (E_1 is the same as E_2), and a value of 1.00 means that the association is perfect (E_2 is 0, and scores on the dependent variable can be predicted without error from the independent variable).

Unlike phi or V, the value of lambda between the extremes of 0.00 and 1.00 has a precise meaning: It tells us how much the independent variable (X) helps us to predict (or, more loosely, understand) the dependent variable (Y). When multiplied by 100, the value of lambda indicates the strength of the association in terms of the percentage reduction in error. Thus, the lambda calculated for Table 11.10 could be interpreted by saying that knowledge of gender improves our ability to predict height by 71%. That is, we are 71% better off knowing gender when attempting to predict height.

The Limitations of Lambda

Lambda has two characteristics that should be stressed. First, lambda is asymmetric. This means that the value of the statistic will vary, depending on which variable is taken as independent. Thus, you should exercise some caution in the designation of an independent variable. If you consistently follow the convention of arraying the independent variable in the columns, the asymmetry of the statistic should not be confusing.

Second, when one of the row totals is much larger than the others, lambda can be misleading. It can be 0.00 even when other measures of association are greater than 0.00 and the conditional distributions for the table indicate that there is an association between the variables. This means that great caution should be exercised in the interpretation of lambda when the row marginals are very unequal. In fact, in the case of very unequal row marginals, a chi square–based measure of association would be the preferred measure of association. (*For practice in computing lambda, see problems 11.1 to 11.4, 11.5a and c, and 11.13.*)

Applying Statistics 11.2 Using Cramer's *V* and Lambda

In Applying Statistics 11.1, we used the maximum difference to determine that the relationship between racial or ethnic group and support for affirmative action was moderate to strong in strength. Does Cramer's *V* confirm that description? For convenience, the table is reproduced here.

Support for Affirmative Action by Racial or Ethnic Group

	Racial or Ethnic Group			
Affirmative Action	White	Black	Hispanic	Totals
Support	139 (15.0%)	76 (43.2%)	8 (14.8%)	223 (19.3%)
Oppose	788 (85.0%)	100 (56.8%)	46 (85.2%)	934 (80.7%)
Totals	927 (100.0%)	176 (100.0%)	54 (100.0%)	1157 (100.0%)
Chi square = 76.26, $df = 2$, $p < 0.05$				

Source: 2012 General Social Survey

$$V = \sqrt{\frac{\chi^2}{N(\min r - 1, c - 1)}}$$

$$V = \sqrt{\frac{76.26}{(1157)(1)}}$$

$$V = \sqrt{\frac{76.26}{1157}}$$

$$V = \sqrt{0.06}$$

$$V = 0.25$$

Based on Table 11.8, *V* is consistent with the maximum difference and the relationship is moderate to strong.

Would we get the same result with Lambda? To calculate the statistics, we first find E_1 and E_2:

$$E_1 = 1157 - 934 = 223$$

For whites, $E_2 = 927 - 788 = 139$
For blacks, $E_2 = 176 - 100 = 76$
For Hispanics, $E_2 = 54 - 46 = 8$

$$E_2 = 223$$

Lambda is

$$\lambda = \frac{E_1 - E_2}{E_1} = \frac{223 - 223}{223} = 0.00$$

Contrary to the maximum difference and *V*, Lambda indicates no relationship between the variables. How can this be? Remember that lambda can be misleading when row totals are very different, as is the case here. The other measures of association convey a more reasonable summary of the relationship and would be the preferred statistics in this case.

STATISTICS IN EVERYDAY LIFE

The Problem with Congress

It will come as no surprise to learn that Americans are frustrated with the federal government and Congress in particular. Where do people place the blame? Do they believe that the entire system of government is broken or is it the current members of Congress who are to blame? A recent poll of a random sample of Americans found surprising agreement across political lines.

(continued)

STATISTICS
IN EVERYDAY LIFE

(*continued*)

Is the Problem with Congress the System or the Members?*

	Political Party			
	Republican	Democrat	Independent	Totals
The Members	234 (58.1%)	270 (57.1%)	312 (56.0%)	816 (56.9%)
The System	169 (41.9%)	203 (42.9%)	245 (44.0%)	617 (43.1%)
	403 (100.0%)	473 (100.0%)	557 (100.0%)	1433 (100.0%)

**Actual Question:* Which comes closer to your view of Congress these days?
a. The political system can work fine, it's the members that are the problem.
b. Most members have good intentions, it's the political system that is broken.

$\chi^2 = 0.40$, $df = 2$, $p > 0.05$ Cramer's $V = 0.02$ Lambda $= 0.00$

Source: Pew Research Center, Jan., 2013. Accessed at http://www.people-press.org/2013/01/31/majority-says-the-federal-government-threatens-their-personal-rights/

There is no relationship between these variables. Chi square is not statistically significant and the measures of association are virtually zero. Regardless of political party, Americans agree that the current members of congress, not the system itself, are the source of their frustrations.

Measures of Association for Collapsed Ordinal Variables

There are two types of ordinal-level variables. Some have only a few (no more than five or six) scores, and we will refer to these as *collapsed ordinal variables.* A number of measures of association have been invented for use with these variables but, rather than attempt to cover them all, we will focus on a statistic called **gamma (G)**.

The other type of ordinal variable has many possible scores and resembles interval-ratio–level variables. We will call these *continuous ordinal variables*. A scale that measures an attitude with many different items could produce this type of variable. We will cover **Spearman's rho (r_s)**, which is designed to measure the association between variables of this type, in the next section.

Ordinal-level measures of association are more sophisticated than lambda and Cramer's V but it is important to remember that they still address the same three basic questions about a bivariate association:

- Are the variables associated?
- How strong is the association?
- What is the direction of the association?

Gamma

Gamma can be interpreted using the proportional reduction in error (PRE) logic introduced for lambda. The value of gamma, when expressed as a percentage, tells us the extent to which knowledge of one variable improves our ability to predict the other variable.

There are two important differences between gamma and lambda. First, the scores of ordinal-level variables have order and can be arrayed from high to low. Thus, gamma tells us the direction of the relationship—positive or negative—in addition to its strength.

Second, with gamma, the PRE logic is based on comparing the rank order of pairs of cases on the two variables. If a case ranks higher than another case on one variable, does it also rank higher on the other variable? Gamma compares the number of pairs of cases ranked in the same order on both variables (N_s) with the number of cases ranked in different order (N_d).

Computing Gamma. See the "One Step at a Time" box for instructions on computing gamma for tables of any size. To illustrate the computational routine, we will use a 2 × 2 table. Larger tables require more computational steps but follow the same logic.

Assume that a researcher is concerned about the causes of "burnout" (that is, demoralization and loss of commitment) among elementary school teachers and wonders whether there is a relationship between burnout (Y) and years of service (X). Table 11.11 presents a relationship between "length of service" and "burnout" for a fictitious sample of 100 teachers.

How many pairs of cases are ranked in the same order in this table? Suppose Mr. Jones is a new teacher (scores low on years of service, X) who has a lot of enthusiasm (scores low on burnout, Y). He would be placed in the upper-left cell of Table 11.11.

Now suppose that Ms. Kelly is nearing the end of her career (scores high on X) and is feeling very discouraged (scores high on Y). She would find herself in the lower-right cell.

If we formed a pair of cases with these two individuals, they would rank in the same order on both variables: On X, Ms. Kelly would rank higher than Mr. Jones (she has taught for more years) and on Y, she would also rank higher (she is more burned out). In fact, any pair of cases formed between the upper-left cell and the lower-right cell would contribute to N_s. So, to find the total number of pairs of cases ranked the same on both variables in a 2 × 2 table, multiply the number of cases in the upper-left cell by the number of cases in the lower-right cell. For Table 11.11,

$$N_s = (25)(32) = 800$$

TABLE 11.11 Burnout by Length of Service (fictitious data)

	Years of Service (X)		
Burnout (Y)	Low	High	Totals
Low	25 (*52.1%*)	20 (*38.5%*)	45 (*45.0%*)
High	23 (*47.9%*)	32 (*61.5%*)	55 (*55.0%*)
Totals	48 (*100.0%*)	52 (*100.0%*)	100 (*100.0%*)

What about cases ranked in different order? Imagine Mr. Gordon, a teacher who has many years of experience (ranks high on *X*) and retains his enthusiasm for the work (ranks low on *Y*). He would be placed in the upper-right cell of Table 11.11.

Finally, imagine Ms. Griffin, a new teacher (who ranks low on *X*) who has already lost her interest in teaching (ranks high on *Y*). She would be sorted into the lower-left cell of the table and, if we formed a pair of cases with these two, they would rank differently on the variables. On *X*, Mr. Gordon ranks higher than Ms. Griffin because he has taught longer but, on *Y*, Ms. Griffin ranks higher because she is more burned out.

Any pair of cases formed between the upper-right cell and the lower-left cell would contribute to N_d, the total number of pairs of cases ranked differently. To find total N_d for a 2×2 table, multiply the number of cases in the upper-right cell by the number of cases in the lower-left cell. For Table 11.11,

$$N_d = (20)(23) = 460$$

The formula for Gamma is

FORMULA 11.4

$$G = \frac{N_s - N_d}{N_s + N_d}$$

where N_s = the total number of pairs of cases ranked in the same order on both variables
N_d = the total number of pairs of cases ranked in different order on the variables

For Table 11.11, gamma would be

$$G = \frac{N_s - N_d}{N_s + N_d}$$

$$G = \frac{800 - 460}{800 + 460}$$

$$G = \frac{340}{1260}$$

$$G = +0.27$$

Interpreting Gamma. Remember that gamma is an ordinal measure and tells us two different things: the strength of the association and its direction. We will deal with these separately.

In terms of the strength of the association, we can use the logic of PRE and say that a gamma of +0.27 means that we would reduce our errors by 27% (0.27×100) when predicting the order of pairs of cases on one variable from the order of pairs of cases on the other (as opposed to predicting order while ignoring the other variable). Also, we can use Table 11.12 to interpret the strength of gamma in a format similar to Tables 11.5 and 11.8. As before, the descriptive terms are arbitrary and intended as general guidelines only. A gamma of +0.27 is weak to moderate in strength.

TABLE 11.12 Guidelines for Interpreting the Strength of the Relationship for Ordinal-Level Measures of Association

Measure of Association	Strength
If the value is	*The strength of the relationship is*
between 0.00 and 0.30	Weak
between 0.31 and 0.60	Moderate
greater than 0.60	Strong

In terms of direction, this is a positive relationship: Burnout increases as length of service increases. Knowing the rankings of two teachers on length of service (Kelly is higher on length of service than Jones) will help us predict their ranking on burnout (we would predict that Kelly will also be higher than Jones on burnout).

We should also use the column percentages to help us analyze the direction of the relationship. In a positive relationship, high scores would cluster with high scores and low scores with low. Hence, the cases would tend to fall in a diagonal from upper left to lower right, as they do in Table 11.11. In a negative relationship, high scores on one variable would be associated with low scores on the other, and cases would tend to fall in a diagonal from lower left to upper right. See Tables 11.6 and 11.7 for examples of positive and negative relationships.

Finally, note that strength and direction are two different things. That is, a gamma of −0.35 is exactly as strong as a gamma of +0.35 but is opposite in direction.

Unlike lambda, gamma is a symmetrical measure of association. Its value will be the same regardless of which variable is taken as independent. *(To practice computing and interpreting gamma, see problems 11.5b, 11.6 to 11.9, and 11.14.)*

ONE STEP AT A TIME Computing Gamma

Step	***Operation***
	Make sure the table is arranged so that columns increase in value from left to right and rows increase from top to bottom.
1	To compute N_s, start with the upper-left-hand cell. Multiply the number of cases in the cell by the *total* number of cases in *all* cells below and to the right. Repeat this process for every cell in the table. Remember that cells that have no cells below and to the right cannot contribute to N_s.
2	To compute N_d, start with the upper-right-hand cell. Multiply the number of cases in the cell by the *total* number of cases in *all* cells below and to the left. Repeat this process for every cell in the table. Remember that cells that have no cells below and to the left cannot contribute to N_s.
	To solve Formula 11.4:
3	Subtract N_d from N_s.
4	Add N_s and N_d.
5	Divide the value you found in Step 3 by the value you found in Step 4. This value is Gamma.

ONE STEP AT A TIME **Interpreting Gamma**

Step	***Operation***
To interpret strength:	
1.	Use Table 11.12 to describe strength in general terms.
2.	Multiply gamma by 100. This value represents the percentage improvement in predicting the dependent variable by taking the independent variable into account.
To interpret direction:	
1.	The sign of gamma indicates direction. However, *be careful when interpreting direction with ordinal variables*. Coding for these variables is arbitrary and a positive gamma may actually indicate a negative relationship and vice versa. See the box "Interpreting Direction with Ordinal-Level Variables."
2.	Always consider the column percentages. If the percentages tend to fall in the diagonal from upper-left to lower-right, the relationship is positive. If they tend to fall in the diagonal from lower-left to upper-right, the relationship is negative.

Applying Statistics 11.3 Interpreting Direction with Ordinal-Level Variables

When determining the direction of a relationship, you might be tempted to look only at the sign of the measure of association. This is certainly understandable but can lead to confusion and mistakes because the coding schemes for ordinal-level variables are arbitrary and a higher score may mean "more" or "less" of the variable being measured. For example, if we measured social class as upper, middle, or lower, we could assign scores to the categories in either of two ways:

A	B
1. Upper	3. Upper
2. Middle	2. Middle
3. Lower	1. Lower

Coding scheme B might seem preferable (because higher scores go with higher class position), but *both* schemes are perfectly legitimate, and the direction of a relationship will change depending on which scheme is selected. Using scheme B, we would find a positive gamma between social class and (for example) education: As education increased, so would class.

With scheme A, however, the same relationship would appear to be negative because the numerical scores (1, 2, 3) are coded in reverse order: The highest social class is assigned the lowest score, and so forth. If you didn't check the coding scheme, you might conclude that class decreases as education increases, when, actually, the opposite is true.

Unfortunately, this source of confusion cannot be avoided when working with ordinal-level variables. Coding schemes will always be arbitrary for these variables, so you need to exercise additional caution when interpreting the direction of ordinal-level variables.

Using SPSS: Crosstabs and Gamma

In this installment of "Using SPSS," we will once again use the **Crosstabs** procedure, but this time to produce measures of association. We will examine the relationship between *degree* and happiness (*happy*). Are more-educated people happier? Both variables are ordinal in level of measurement, but we will request nominal measures of association in addition to gamma.

To begin,

1. Click the SPSS icon on your desktop.
2. Load the *GSS2012* database.
   a. Find the **File** command on the far left of the menu bar and then click **File → Open → Data**.
   b. Find the *GSS2012* database and click the file name.
3. From the main menu bar, click **Analyze, Descriptive Statistics**, and **Crosstabs**.
4. In the "Crosstabs" dialog box:
   a. Find *happy* and move it to the "Rows" box.
   b. Find *degree* and move it into the "Columns" box.
5. Click the **Statistics** button and select **Chi-square, Phi and Cramer's *V*, Lambda**, and **Gamma**.
6. Return to the "Crosstabs" window and click the **Cells** button and select **column** in the "Percentages" box.
7. Return to the "Crosstabs" window and click **OK**.
8. Note: The output table has been edited to fit in this space. The statistics we requested are reported under the table, as you would see in the professional research literature.

Happiness by Degree

	R's Highest Degree					
Happy:	LT HS	HS	Jr. Coll.	BA	Grad.	Totals
Very	31.4%	29.2%	34.8%	35.9%	34.9%	31.8%
Pretty	45.9%	58.2%	50.0%	54.9%	57.9%	55.1%
Not Too	22.7%	12.6%	15.2%	9.2%	7.2%	13.1%
Totals	100.0% (220)	100.0% (692)	100.0% (112)	100.0% (273)	100.0% (152)	100.0% (1449)

$\chi^2 = 32.35$, $df = 8$, $p < 0.05$. Lambda = 0.00; Cramer's $V = 0.11$; Gamma = −0.12

As usual, SPSS produces a lot of information, which we will examine step-by-step.

- Chi square is significant ($p < 0.05$). This is important since we are working with a random sample. We can be confident that these variables are also associated in the population.
- The column percentages change, so these variables are associated. Looking at the top row ("very happy"), happiness tends to increase as education increases and the most educated respondents (people with BA and graduate degrees) are most likely to report being very happy.
- The measures of association also indicate that the variables are related. Cramer's *V* and Gamma suggest that the strength of this association is weak to moderate. (We can disregard the lambda of 0.00 since it is contradicted by the other statistics and the pattern of column percentages.)

- Gamma is negative but the column percentages show that this relationship is actually positive ("happiness increases as education increases"). How can this be? The scores on *happy* are "reversed" and lower scores mean greater happiness. (See Appendix G or click **Utilities → Variables** on the SPSS menu bar to see the coding scheme). See the box "Interpreting Direction with Ordinal-level Variables" and be careful when interpreting direction.

What can we conclude? There is a statistically significant, weak to moderate, positive relationship between happiness and education. Happiness tends to increase with education and the most educated tend to be the happiest.

STATISTICS
IN EVERYDAY LIFE

Generation Gaps

In a Chapter 9 "Statistics in Everyday Life" feature, we mentioned generation gaps in opinions on controversial issues such as gay marriage and raised the issue of the source of these gaps: Are the differences caused by aging or do they reflect the different eras in which the generations were born?

Sometimes, it's easy to figure out the source of the gap. Consider the use of modern technology. Younger generations are "native-born" to personal computers, the Internet, and smart phones: Their world has always included these technologies. Older generations are "newcomers" to the digital world and less involved in its wonders.

For example, according to a recent study, younger people use social media at much higher rates and are far more involved in the latest technologies, including posting "selfies" on social networking sites such as Facebook, as shown in the table below. This relationship is statistically significant and, according to gamma, positive and strong. The positive sign for gamma means that column percentages tend to fall on the diagonal from upper left to lower right: The youngest respondents were the most likely to have shared a "selfie."

Once again, note that interpreting direction with ordinal-level variables can be confusing. This is a "positive" relationship because of the coding scheme: The higher score on the row variable (2) is associated with the higher scores on age group (3 and 4).

Percentage Saying They Have Shared a "Selfie" on a Social Networking Site by Age Group (percentages)

	Age Group			
Selfie?	1	2	3	4
	18–33	*34–49*	*50–68*	*69 and older*
(1) Yes	55%	24%	9%	4%
(2) No	45%	76%	91%	96%
Totals	100% (617)	100% (351)	100% (576)	100% (246)

$\chi^2 = 407.54$, $df = 3$, $p < 0.05$. Gamma $= +0.87$

Source: Pew Research Center. 2014. "Millennials in Adulthood: Detached from Institutions, Networked with Friends." p. 6. Available at http://www.pewsocialtrends.org/files/2014/03/2014-03-07_generations-report-version-for-web.pdf

A Measure of Association for Continuous Ordinal Variables: Spearman's Rho (r_s)

Some ordinal-level variables have a broad range of scores and resemble interval-ratio-level variables. The scores of these variables may be collapsed into a few broad categories (such as high, moderate, and low), organized into a bivariate table, and analyzed with gamma. However, collapsing scores may obscure or lose some important distinctions between cases.

For example, suppose a researcher wished to test the claim that jogging is beneficial, not only physically but also psychologically. Do joggers have an enhanced sense of self-esteem? To deal with this issue, 10 female joggers are evaluated on two scales, both of which range from 0 to 20. The first scale measures involvement in jogging and the second measures self-esteem. Scores are reported in Table 11.13.

The scores could be collapsed and, perhaps, split into high and low categories, but this raises several potential difficulties. First, the scores seem continuous, and there are no obvious or natural division points in the distribution that would allow us to clearly distinguish between high scores and low ones. Second, and more important, grouping these cases into broader categories loses information. That is, if both Wendy and Debbie were placed in the category "high" on involvement, the fact that they had different scores would be lost. If these differences are important and meaningful, then we should choose a measure of association that allows us to retain as much detail in the scores as possible.

Spearman's rho (**r_s**) is a measure of association for ordinal-level variables that have a broad range of many different scores and few ties between cases on either variable. As is appropriate for ordinal-level variables, Spearman's rho is computed from the ranks of the scores, not the scores themselves. First, the cases are ranked from high to low on each variable, and then the ranks are used to produce the final measure. Table 11.14 displays the original scores and the rankings of the cases on both variables.

Computing Spearman's Rho. To rank the cases, first find the highest score on each variable and assign it rank 1. Wendy has the high score on X (18) and is thus ranked number 1. Debbie, on the other hand, is highest on Y and is ranked first on that variable. All other cases are then ranked in descending order.

TABLE 11.13 Scores on Involvement in Jogging and Self-Esteem

Jogger	Involvement in Jogging (X)	Self-Esteem (Y)
Wendy	18	15
Debbie	17	18
Phyllis	15	12
Stacey	12	16
Evelyn	10	6
Tricia	9	10
Christy	8	8
Patsy	8	7
Marsha	5	5
Lynn	1	2

TABLE 11.14 Computing Spearman's Rho

	Involvement (X)	Rank	Self-Image (Y)	Rank	D	D^2
Wendy	18	1	15	3	–2	4
Debbie	17	2	18	1	1	1
Phyllis	15	3	12	4	–1	1
Stacey	12	4	16	2	2	4
Evelyn	10	5	6	8	–3	9
Tricia	9	6	10	5	1	1
Christy	8	7.5	8	6	1.5	2.25
Patsy	8	7.5	7	7	0.5	0.25
Marsha	5	9	5	9	0	0
Lynn	1	10	2	10	0	0
					$\Sigma D = 0$	$\Sigma D^2 = 22.5$

If cases have the same score, assign them the average of the ranks they would have used up had they not been tied. Christy and Patsy have identical scores of 8 on involvement. Had they not been tied, they would have used up ranks 7 and 8. The average of these two ranks is 7.5, and this average is assigned to all tied cases. (For example, if Marsha had also had a score of 8, three ranks—7, 8, and 9—would have been used, and all three tied cases would have been assigned rank 8.)

The formula for Spearman's rho is

FORMULA 11.5

$$r_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

where ΣD^2 = the sum of the squares of the differences in ranks

See the "One Step at a Time" box for detailed instructions on computing Spearman's rho. For our sample problem,

$$r_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

$$r_s = 1 - \frac{6(22.5)}{10(100 - 1)}$$

$$r_s = 1 - \frac{135}{990}$$

$$r_s = 1 - 0.14$$

$$r_s = 0.86$$

Interpreting Spearman's Rho. Spearman's rho is an index of the strength of association between the variables and can be interpreted using the guidelines presented in Table 11.12. The direction of the relationship is indicated by the sign of the statistic. A positive value means that the cases tend to be ranked in the same order on the two variables, and a negative value means that they tend to be ranked in reverse order.

The Spearman's rho of 0.86 in our example problem indicates a strong, positive relationship between these two variables. The respondents who were highly involved in jogging also ranked high on self-image. These results are supportive of claims regarding the psychological benefits of jogging.

If the value of rho is squared, the statistic can be interpreted with the logic of PRE. Rho squared (r_s^2) represents the proportional reduction in errors of prediction when predicting rank on one variable from rank on the other variable, as compared to predicting rank while ignoring the other variable. In our example, $r_s = 0.86$ and $r_s^2 = 0.74$. Thus, our errors of prediction would be reduced by 74% if we used the rank on jogging to predict rank on self-image. (*For practice in computing and interpreting Spearman's rho, see problems 11.10 to 11.12.*)

ONE STEP AT A TIME — Computing Spearman's Rho

Step	***Operation***
1.	Set up a computing table like Table 11.14 to organize the computations.
2.	In the far left-hand column, list the cases in order, with the case with the highest score on the independent variable (*X*) stated first.
3.	In the next column, list the rank of each case on *X*, beginning with rank 1 for the highest score. If any cases have the same score, assign them the average of the ranks they would have used up had they not been tied.
4.	In the next column, list the scores on *Y* and then, in a new column, rank the cases from high to low on *Y*. Assign the rank of 1 to the case with the highest score on *Y*, and assign any tied cases the average of the ranks they would have used up had they not been tied.
5.	For each case, subtract the rank on *Y* from the rank on *X* and write the difference (*D*) in a new column. The sum of this column must be zero. If it is not, you have made a mistake and need to re-compute.
6.	Square the value of each *D* and record the result in a new column. Add this column up to find ΣD^2, and substitute this value into the numerator of Formula 11.5.

To solve Formula 11.5:

1.	Multiply ΣD^2 by 6.
2.	Square *N* and subtract 1 from the result.
3.	Multiply the quantity you found in step 2 by *N*.
4.	Divide the quantity you found in step 1 by the quantity you found in step 3.
5.	Subtract the quantity you found in step 4 from 1. This is r_s.

ONE STEP AT A TIME — Interpreting Spearman's Rho

Step	***Operation***

To interpret strength:

1.	Use Table 11.12 to describe strength in general terms.
2.	Square the value of rho and multiply by 100. This value represents the percentage improvement in our predictions of the dependent variable by taking the independent variable into account.

To interpret direction:

1.	Use the sign of r_s. Be careful when interpreting direction with ordinal-level variables. Remember that coding schemes for these variables are arbitrary and that a positive r_s may mean that the actual relationship is negative, and vice versa.

Applying Statistics 11.4 Spearman's Rho

Five cities have been rated on an index that measures the quality of life. Also, the percentage of the population that has moved into each city over the past year has been determined. Have cities with higher quality-of-life scores attracted more new residents? The following table below summarizes the scores, ranks, and differences in ranks for each of the five cities.

City	Quality of Life	Rank	% New Residents	Rank	D	D^2
A	30	1	17	1	0	0
B	25	2	14	3	−1	1
C	20	3	15	2	1	1
D	10	4	3	5	−1	1
E	2	5	5	4	1	1
					$\Sigma D = 0$	$\Sigma D^2 = 4$

Spearman's rho for these variables is

$$r_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

$$r_s = 1 - \frac{(6)(4)}{5(25 - 1)}$$

$$r_s = 1 - \frac{24}{120}$$

$$r_s = 1 - 0.20$$

$$r_s = 0.80$$

These variables have a strong, positive association. The higher the quality-of-life score, the greater the percentage of new residents. The value of r_s^2 is 0.64($0.80^2 = 0.64$), which indicates that we will make 64% fewer errors when predicting rank on one variable from rank on the other, as opposed to ignoring rank on the other variable.

STATISTICS IN EVERYDAY LIFE

Ranking Haunted Cities

Rankings such as those used to compute Spearman's rho have become a staple of the media and the Internet and there are infinite lists of everything imaginable: best places to live, cities with the worst traffic, most important news stories, most popular celebrities, and so forth. These "Top Ten" lists are bound to be arbitrary and subjective but can also provide useful insights and interesting information. A search of the Internet provides a multitude of examples, one of the most interesting of which was a list of the most haunted cities in the United States:

Rank	City	Comment
1	New Orleans, LA	Haunted mansions, above-ground graveyards, voodoo, murder, screams in the night—New Orleans has it all.
2	Galveston, TX	Perhaps 8000 died in the Great Storm of 1900, still the worst natural disaster in U.S. history.
3	Gettysburg, PA	Almost 50,000 died during the three-day battle.
4	Key West, FL	Pirates, rumrunners, and Robert the Haunted Doll!
5	Savannah, GA	Streets filled with shadows; creepy legends abound.
6	Charleston, SC	A rich and tragic history, giving rise to many tales and "experiences."
7	San Francisco, CA	Disasters, haunted houses, and Alcatraz.

STATISTICS
IN EVERYDAY LIFE

(*continued*)

Rank	City	Comment
8	Chicago, IL	The great fire, gangsters, and “Resurrection Mary.”
9	Salem, MA	Witches!
10	Miami, FL	Haunted hotels (“strange noises on the 13th floor”), the Miami Circle, and gangsters.

Source: http://www.hauntedamericatours.com/toptenhaunted/toptenhauntedcities/

This particular list was compiled by a company that sells ghost tours, so its scientific validity might be debatable (to say the least). Still, are there some sociological variables that might correlate with this ranking? What relationships suggest themselves? What types of social conditions would breed ghost stories and interest in the “supernatural”?

BECOMING A CRITICAL CONSUMER: Bivariate Tables in the Professional Literature

As was the case with ANOVA, it is unlikely that you will encounter reports using chi square, bivariate tables, or measures of association in everyday life or in the popular media. Therefore, we will focus on the presentation of bivariate tables in the professional research literature. These tables will usually—but not always—be constructed with the independent variable in the columns and the dependent variable in the rows. Be careful to check the format to see if these conventions have been followed and be prepared to reorient your analysis (or redraw the table) if the independent variable is in the rows. Above all, you should check to see that the percentages have been calculated in the correct direction. Even skilled professionals occasionally calculate percentages incorrectly and misinterpret the data.

Once you have assured yourself that the table is properly presented, you can apply the analytical techniques developed in this chapter. By comparing the conditional distributions of the dependent variable, you can ascertain for yourself if the variables are associated and check the strength, the pattern, and (for tables with ordinal-level variables) the direction of the association. You may then compare your conclusions with those of the researchers. Finally, note the measures of association and compare the interpretation presented by the researchers with the guidelines presented in this chapter. What does the measure of association say about the strength of the relationship?

SUMMARY

1. Measures of association provide information that is complementary to tests of significance.
2. We have a variety of tools to analyze the association between variables, including bivariate tables and various statistics. We use these tools to develop answers to three questions regarding the existence, strength, and pattern or direction of relationships.
3. Phi, Cramer’s *V*, and lambda are measures of association for nominal variables and each is appropriate for a specific situation. These statistics indicate *only* the strength of the relationship. The pattern of the relationship can be determined by inspecting column percentages.
4. Gamma and Spearman’s rho are measures of association for collapsed and continuous ordinal-level variables, respectively. They indicate both the strength and direction of the relationship.

SUMMARY OF FORMULAS

FORMULA 11.1	Phi:	$\phi = \sqrt{\frac{\chi^2}{N}}$
FORMULA 11.2	Cramer's *V:*	$V = \sqrt{\frac{\chi^2}{N(\text{min } r - 1, c - 1)}}$
FORMULA 11.3	Lambda:	$\lambda = \frac{E_1 - E_2}{E_1}$
FORMULA 11.4	Gamma:	$G = \frac{N_s - N_d}{N_s + N_d}$
FORMULA 11.5	Spearman's rho:	$r_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$

GLOSSARY

Association. The relationship between two (or more) variables. Two variables are said to be associated if the distribution of one variable changes for the various categories or scores of the other variable.

Column percentages. Percentages computed within each column of a bivariate table.

Conditional distribution of *Y*. The distribution of scores on the dependent variable for a specific score or category of the independent variable when the variables have been organized into table format.

Cramer's *V*. A chi square–based measure of association for nominally measured variables that have been organized into a bivariate table with any number of rows and columns.

Dependent variable. In a bivariate relationship, the variable that is taken as the effect.

Gamma (*G*). A measure of association for "collapsed" ordinal variables in bivariate table format.

Independent variable. In a bivariate relationship, the variable that is taken as the cause.

Lambda. A proportional reduction in error (PRE) measure for nominally measured variables that have been organized into a bivariate table.

Maximum difference. A way to assess the strength of an association between variables that have been organized into a bivariate table. The maximum difference is the largest difference between column percentages for any row of the table.

Measures of association. Statistics that quantify the strength of the association between variables.

N_d. In a bivariate table, the number of pairs of cases ranked in different order on two variables.

N_s. In a bivariate table, the number of pairs of cases ranked in the same order on two variables.

Negative association. A bivariate relationship where the variables vary in opposite directions. As one variable increases, the other decreases, and high scores on one variable are associated with low scores on the other.

Phi (ϕ). A chi square–based measure of association for nominally measured variables that have been organized into a bivariate table with two rows and two columns (a 2 × 2 table).

Positive association. A bivariate relationship where the variables vary in the same direction. As one variable increases, the other also increases, and high scores on one variable are associated with high scores on the other.

Proportional reduction in error (PRE). A logic used in the design of certain measures of association. A PRE measure indicates how much knowledge of the independent variable improves our predictions of the dependent variable.

Spearman's rho (r_s). A measure of association for "continuous" ordinal variables.

***X*.** Symbol used for any independent variable.

***Y*.** Symbol used for any dependent variable.

PROBLEMS

11.1 *SOC* The administration of a local college campus wants to increase the mandatory student fee to finance an upgrading of the football program. A survey of a sample of faculty regarding this proposal has been completed. Is there any association between support for raising fees and the gender, discipline, or tenured status of the faculty? Use column percentages, the maximum difference, and an appropriate measure of association to describe the strength and pattern of these associations.

a. Support for raising fees by gender

	Gender		
Support?	Males	Females	Totals
For	12	7	19
Against	15	13	28
Totals	27	20	47

b. Support for raising fees by discipline

	Discipline		
Support?	Liberal Arts	Science and Business	Totals
For	6	13	19
Against	14	14	28
Totals	20	27	47

c. Support for raising fees by tenured status

	Status		
Support?	Tenured	Nontenured	Totals
For	15	4	19
Against	18	10	28
Totals	33	14	47

11.2 *SOC* A researcher has conducted a survey on sexual attitudes for a sample of 317 teenagers. The respondents were asked whether they considered premarital sex "always wrong" or "Not always wrong." The following tables summarize the relationship between responses to this item and several other variables. For each table, assess the strength and pattern of the relationship and write a paragraph interpreting these results.

a. Attitudes toward premarital sex by gender

	Gender		
Premarital Sex	Female	Male	Totals
Always wrong	90	105	195
Not always wrong	65	57	122
Totals	155	162	317

b. Attitudes toward premarital sex by courtship status

	Ever Gone "Steady?"		
Premarital Sex	No	Yes	Totals
Always wrong	148	47	195
Not always wrong	42	80	122
Totals	190	127	317

11.3 *GER* A survey of senior citizens who live in either a housing development specifically designed for retirees or an age-integrated neighborhood has been conducted. Is type of living arrangement related to sense of social isolation? Analyze the relationship with column percentages, the maximum difference, and an appropriate measure of association. Summarize your findings in a paragraph.

	Living Arrangement		
Sense of Isolation	Housing Development	Integrated Neighborhood	Totals
Low	80	30	110
High	20	120	140
Totals	100	150	250

11.4 *SOC* Below are the four dependent variables cross-tabulated against gender. Analyze these relationships with column percentages, the maximum difference, and an appropriate measure of association. Summarize the results of your analysis in a paragraph that describes the strength and pattern of each relationship.

a. Support for the legal right to an abortion by gender

	Gender		
Right to Abortion?	Males	Females	Totals
Yes	310	418	728
No	432	618	1050
Totals	742	1036	1778

b. Support for capital punishment by gender

Capital Punishment?	Gender		
	Males	Females	Totals
Favor	908	998	1906
Oppose	246	447	693
Totals	1154	1445	2599

c. Approve of suicide for people with an incurable disease by gender

Right to Suicide?	Gender		
	Males	Females	Totals
Yes	524	608	1132
No	246	398	644
Totals	770	1006	1776

d. Support for sex education in public schools by gender

Sex Education?	Gender		
	Males	Females	Totals
Favor	685	900	1585
Oppose	102	134	236
Totals	787	1034	1821

11.5 *PS* Below are some of the tables you analyzed in problem 10.17. You already have the results of the chi square tests. Now, analyze these relationships in terms of association, using an appropriate measure of association. Write a paragraph summarizing your results.

a. Presidential preference and race/ethnicity

Preference	Race/Ethnicity			
	White	Black	Latino	Totals
Romney	289	5	44	338
Obama	249	95	66	410
Totals	538	100	110	748

b. Presidential preference by education

Preference	Education				
	Less than HS	HS Graduate	College Graduate	Post-Grad. Degree	Totals
Romney	30	180	118	10	338
Obama	35	120	218	37	410
Totals	65	300	336	47	748

c. Presidential preference by religion

Preference	Religion					
	Protestant	Catholic	Jewish	None	Other	Totals
Romney	165	110	10	28	25	338
Obama	245	55	20	60	30	410
Totals	410	165	30	88	55	748

11.6 *SOC* A small sample of non-English-speaking immigrants to the United States has been interviewed about their level of assimilation. Is the pattern of adjustment affected by length of residence in the United States? For each table compute gamma and summarize the relationship in terms of strength and direction.

a. English fluency

English Fluency	Length of Residence		
	Less Than Five Years (Low)	More Than Five Years (High)	Totals
Low	20	10	30
High	5	15	20
Totals	25	25	50

b. Total family income

Income	Length of Residence		
	Less Than Five Years (Low)	More Than Five Years (High)	Totals
Low	18	8	26
High	7	17	24
Totals	25	25	50

c. Extent of contact with country of origin

Contact	Length of Residence		
	Less Than Five Years (Low)	More Than Five Years (High)	Totals
Rare	5	20	25
Frequent	20	5	25
Totals	25	25	50

11.7 *PA* Various supervisors in the city government of Shinbone, Kansas, have been rated on the extent to which they practice authoritarian styles of leadership and decision making. The efficiency of each department has also been rated, and the results are summarized in the following table. Use column percentages, the maximum difference, and an appropriate measure of association to describe this association. What is the direction of the association?

	Authoritarianism		
Efficiency	Low	High	Totals
Low	10	12	22
High	17	5	22
Totals	27	17	44

11.8 *PA* All applicants to the Police Academy are given an aptitude test, but the test has never been evaluated to see if test scores are in any way related to job performance. Describe the strength and direction of the relationship in a sentence or two. Should the test continue to be administered? Why or why not?

	Test Scores			
Efficiency	Low	Moderate	High	Totals
Low	12	6	7	25
Moderate	9	11	9	29
High	5	9	10	24
Totals	26	26	26	78

11.9 *SW* A sample of children has been observed and rated for symptoms of depression. Their parents have been rated for level of alcohol use. Is there any relationship between these variables? Write a few sentences stating your conclusions.

Symptoms of Depression	Level of Alcohol Use			
	Low	Moderate	High	Totals
Few	7	8	9	24
Some	15	10	18	43
Many	8	12	3	23
Totals	30	30	30	90

11.10 *SOC* Eleven neighborhoods have been rated by an urban sociologist on a "quality-of-life" scale (which includes measures of affluence, availability of medical care, and recreational facilities) and a social cohesion scale. The results are presented here in scores. Higher scores indicate higher quality of life and greater social cohesion. Are the two variables associated? What is the strength and direction of the association? Summarize the relationship in a sentence or two. *(HINT: Don't forget to square the value of Spearman's rho for a PRE interpretation.)*

Neighborhood	Quality of Life	Social Cohesion
Beaconsdale	17	8.8
Brentwood	40	3.9
Chesapeake Shores	47	4.0
College Park	90	3.1
Denbigh Plantation	35	7.5
Kingswood	52	3.5
North End	23	6.3
Phoebus	67	1.7
Riverview	65	9.2
Queens Lake	63	3.0
Windsor Forest	100	5.3

11.11 *SOC* Following are the scores of a sample of 15 nations on a measure of ethnic diversity (the higher the number, the greater the diversity) and a measure of economic inequality (the higher the score, the greater the inequality). Are these variables related? Are ethnically diverse nations more economically unequal?

Nation	Diversity	Inequality
India	91	29.7
South Africa	87	58.4
Kenya	83	57.5
Canada	75	31.5
Malaysia	72	48.4
Kazakhstan	69	32.7
Egypt	65	32.0
United States	63	41.0
Sri Lanka	57	30.1
Mexico	50	50.3
Spain	44	32.5
Australia	31	33.7
Finland	16	25.6
Ireland	4	35.9
Poland	3	27.2

11.12 *SOC* Twenty ethnic, racial, or national groups were rated by a random sample of white and black students on a Social Distance Scale. Lower scores represent less social distance and less prejudice. How similar are these rankings?

Group	Average Score	
	White Students	Black Students
1 White Americans	1.2	2.6
2 English	1.4	2.9
3 Canadians	1.5	3.6
4 Irish	1.6	3.6
5 Germans	1.8	3.9
6 Italians	1.9	3.3
7 Norwegians	2.0	3.8
8 American Indians	2.1	2.7
9 Spanish	2.2	3.0
10 Jews	2.3	3.3
11 Poles	2.4	4.2
12 Black Americans	2.4	1.3
13 Japanese	2.8	3.5
14 Mexicans	2.9	3.4
15 Koreans	3.4	3.7
16 Russians	3.7	5.1
17 Arabs	3.9	3.9
18 Vietnamese	3.9	4.1
19 Turks	4.2	4.4
20 Iranians	5.3	5.4

Statistical Analysis Using SPSS

11.13 *SOC* In problem 10.16, you used **Crosstabs** to analyze the relationship between gender and fear of walking alone at night (*fear*), support for capital punishment (*cappun*), and job satisfaction (*satjob*). Now, we will add measures of association to the analysis.

- Click the SPSS icon on your desktop.
- Load the *GSS2012* database.
- Click **Analyze → Descriptive → Crosstabs**.
- Find *fear, cappun,* and *satjob* in the list of variables and move them into the "Rows" box
- Find *sex* in the list of variables and move it to the "Columns" box
- Click the **Statistics** button select chi square, Phi and Cramer's V, and Lambda.
- Return to the "Crosstabs" window, click the **Cells** button and select **column** in the "Percentages" box.
- Return to the "Crosstabs" window and click **OK.** The results of the test will be printed to the SPSS output window.

Examine the column percentages as well as the measures of association and report these results in a few sentences.

11.14. *SOC* Analyze the relationship between education (degree) and fear of walking alone at night (*fear*), support for capital punishment (*cappun*), and frequency of prayer (*pray*). Since degree is an ordinal-level variable, we will use gamma as our measure of association.

- Click the SPSS icon on your desktop.
- Load the *GSS2012* database.
- Click **Analyze → Descriptive → Crosstabs**.
- Find *fear, cappun,* and *pray* in the list of variables and move them into the "Rows" box
- Find *degree* in the list of variables and move it to the "Columns" box
- Click the **Statistics** button select **Chi square** and **Gamma**.
- Return to the "Crosstabs" window, click the **Cells** button and select **column** in the "Percentages" box.
- Return to the "Crosstabs" window and click **OK.** The results of the test will be printed to the "SPSS output" window.

Examine the column percentages as well as the measures of association and report these results in a few sentences.

11.15 *SOC and PS* In this problem, you will use Spearman's rho to analyze the relationship between two variables in the *Intl-POP* database. The first variable (*rights*) measures the extent of civil and political rights in a nation and the second (*corruption*) measures the level of corruption in the public sector. Are these variables related? Are nations with stronger individual rights less corrupt? Take *rights* as the independent variable for this problem.

Both of these variables are ordinal in level of measurement, so Spearman's rho is an appropriate statistic to analyze the strength of the relationship. To do this, we will use a new procedure called **Correlate**, which we will also use in the next chapter.

Before proceeding, take careful note of the coding scheme of the variables (see Appendix G or click **Utilities → Variables** from the SPSS menu). Do high scores mean high levels of corruption and strong civil liberties? Keep the coding scheme in mind as you analyze the relationship.

- Click the SPSS icon on your desktop.
- Load the *Intl-POP* database.
- Click **Analyze → Correlate → Bivariate**
- Find *corruption* and *rights* in the list of variables and move them into the "Variables" box

- In the box labeled "Correlation Coefficients" find and click the button next to "Spearman."
- Click **OK.**

The output will show the relationship between the variables in a box labeled "Correlations." Look in the first row and you will see a cell reporting the relationship between corruption and itself ($r_s = 1.000$). The r_s we are seeking is in the next cell to the right. Take note of the strength and direction of the relationship and write a brief summary. How strong is this relationship? By what percentage do we improve our predictions of *corruption* if we use *rights*? What is the direction of the relationship? *Be careful when interpreting direction!*

YOU ARE THE RESEARCHER

Understanding Political Beliefs, Part II

At the end of Chapter 10, you investigated possible causes of people's beliefs on several controversial issues. Now you will extend your analysis by using **Crosstabs** to get the statistics presented in this chapter. There are two projects and you should complete both.

Project 1: Explaining Beliefs

In this project, you will again analyze beliefs about assisted suicide (*letdie1*), gay marriage (*marhomo*), and immigration (*letin1*). You will select an independent variable *other than the one you used in Chapter 10* and use SPSS to generate chi square, column percentages, and measures of association.

Step 1: Choose an Independent Variable

Select an independent variable that seems likely to be an important cause of people's attitudes about these issues. Be sure to select an independent variable that has *only* 2–5 categories and use the recode command if necessary. You might consider gender, level of education (use *degree*), religion, or age (the recoded version—see Chapter 9) as possibilities, but there are many others. Record the variable name and state exactly what the variable measures in the table below. Also note the level of measurement of the variable.

SPSS Name	What Exactly Does This Variable Measure?	Level of Measurement

Step 2: Stating Hypotheses

State hypotheses about the relationships you expect to find between your independent variable and each of the dependent variables. State these hypotheses in terms of which category of the independent variable you expect to be associated with which category of the dependent variable (for example, "I expect that men will be more supportive of the legal right to an abortion for any reason").

Dependent Variable	Hypothesis
letdie 1	
marhomo	
letin 1	

Step 3: Running Crosstabs

Click **Analyze** → **Descriptive Statistics** → **Crosstabs** and place the dependent variables in the "Rows:" box and the independent variable you selected in the "Columns:" box. Click the **Statistics** button to get chi square and the measures of association and the **Cells** button for column percentages.

Step 4: Recording Results

These commands will generate a lot of output and it will be helpful to summarize your results in the following table. For the chi square test, write "Yes" in the right-hand column of the table if the value of "Asymp. Sig (2-sided)" is less than 0.05.

Dependent Variable	Phi or Cramer's V	Lambda	Gamma	Chi Square Significant at 0.05?
letdie1				
marhomo				
letin1				

Step 5: Analyzing and Interpreting Results

Choose *one* measure of association for each relationship. Your choice should be based primarily on level of measurement but should take account of other statistical characteristics as well.

Write a short summary of results for each dependent variable in which you identify the variables being tested, the results of the chi square test (see Chapter 10), and the strength and pattern or direction of the relationship. It is probably best to characterize the relationship in general terms and then cite the statistical values in parentheses. For example, we might summarize the relationship between education (*degree*) and happiness (*happy*) by saying, "The relationship between education and happiness was significant and weak to moderate in strength (chi square = 32.25, $df = 8$, $p < 0.00$, $V = 0.11$, Gamma = −0.12). Happiness tends to increase as education increases." You should also note whether or not your hypotheses were supported.

Project 2: Exploring the Impact of Various Independent Variables

In this project, you will examine the relative ability of a variety of independent variables to explain or account for a single dependent variable. You will again use the **Crosstabs** procedure in SPSS to generate statistics and use the alpha levels and measures of association to judge which independent variable has the most important relationship with your dependent variable.

Step 1: Choosing Variables

Select a dependent variable. You may use any of the variables from Project 1 or select a new dependent variable. *Be sure that your dependent variable has no more than five values or scores*. Good choices for dependent variables include any measure of attitudes or opinions. *Do not* select characteristics like race, sex, or religion as dependent variables.

Select three independent variables that seem likely to be important causes of the dependent variable you selected. *Be sure that your independent variable has no more than five or six categories*. You might consider gender, level of education (use *degree*), religion, or age (the recoded version—see Chapter 9) as possibilities, but there are many others.

Record the variable names and state exactly what each variable measures and the level of measurement in the table below.

SPSS Name	What Exactly Does This Variable Measure?	Level of Measurement
Dependent Variable		
Independent Variables		

Step 2: Stating Hypotheses

State hypotheses about the relationships you expect to find between your independent variables and the dependent variable. State these hypotheses in terms of which categories of the variables you expect to be associated with each other.

Independent Variables (SPSS Names)	Hypothesis
1. _____________	
2. _____________	
3. _____________	

Step 3: Running Crosstabs

Place your dependent variable in the “Rows:” box and all three of your independent variables in the “Columns:” box. Click the **Statistics** button to get chi square and the measures of association. Click the **Cells** button for column percentages.

Step 4: Recording Results

Your output will consist of three tables, and it will be helpful to summarize your results in the following table.

Independent Variable	Phi or Cramer’s *V*	Lambda	Gamma	Chi Square Significant at 0.05?
1.				
2.				
3.				

Step 5: Analyzing and Interpreting Results

Choose *one* measure of association for each relationship. Your choice should be based primarily on level of measurement but should take account of other statistical characteristics as well.

Write a short summary of results of each test using the same format as in Project 1. Remember to explain whether or not your hypotheses were supported. Finally, assess which of the independent variables had the most important relationship with your dependent variable. Use the alpha level and the value of the measures of association to make this judgment.

12 Association Between Variables Measured at the Interval-Ratio Level

LEARNING OBJECTIVES

By the end of this chapter, you will be able to:

1. Interpret a scatterplot.
2. Calculate and interpret slope (*b*), *Y* intercept (*a*), and Pearson's *r* and r^2.
3. Find and explain the least-squares regression line and use it to predict values of *Y*.
4. Explain the concepts of total, explained, and unexplained variance.
5. Use regression and correlation techniques to analyze and describe a bivariate relationship in terms of the three questions introduced in Chapter 11.
6. Use SPSS to find correlations and conduct regression analysis.

Using Statistics

The statistical techniques presented in this chapter are used to measure the strength and direction of the association between variables measured at the interval-ratio level. Examples of research situations in which these techniques are useful include the following:

1. A family sociologist is studying the division of labor between dual-wage-earner married couples and wonders whether husbands increase their contribution to housework as the number of children in the family increases.
2. A criminologist is studying the relationship between poverty and crime and has gathered information on both variables for 542 counties across the United States. Does the crime rate increase as poverty increases? How strong is the relationship?
3. A demographer is interested in the relationship between fertility (average number of births) and education for females. She has information from 128 nations and wonders whether fertility decreases as average number of years of schooling for women increases. How strong is the association?

In this chapter, we will consider statistical techniques for analyzing the association, or correlation, between variables measured at the interval-ratio level.[1] The techniques presented here are also commonly used on ordinal-level variables, especially when they are "continuous" or have many values (see Chapter 11).

[1]The term *correlation* is commonly used instead of *association* when discussing the relationship between interval-ratio variables. We will use the two terms interchangeably.

As we shall see, these statistics are rather different in their logic and computation from the measures of association covered in Chapter 11. Let me stress at the outset, therefore, that we are still asking the same three questions:

- Is there a relationship between the variables?
- How strong is the relationship?
- What is the direction of the relationship?

It will quickly become apparent that the mathematics in this chapter can be challenging and time-consuming. I use a very small data set to illustrate the calculations and, of course, you should familiarize yourself with these mathematical routines. However, researchers today use computerized statistical packages like SPSS to calculate these statistics and, after working through our simplified example problem, we will do the same. Likewise, the end-of-chapter problems offer some problems with small data sets that you can use to be sure you understand the mathematical routines, but most problems use SPSS.

Scatterplots and Regression Analysis

The usual first step in analyzing a relationship between interval-ratio variables is to construct and examine a **scatterplot**. Like bivariate tables, these graphs allow us to identify several important features of the relationship. An example will illustrate how to construct and interpret scatterplots.

Suppose a researcher is interested in analyzing how dual-wage-earner families (that is, families where both husband and wife have jobs outside the home) cope with housework. Specifically, the researcher wonders whether the number of children in the family is related to the amount of time the husband contributes to housekeeping chores. The relevant data for a sample of 12 families are displayed in Table 12.1.

Constructing Scatterplots

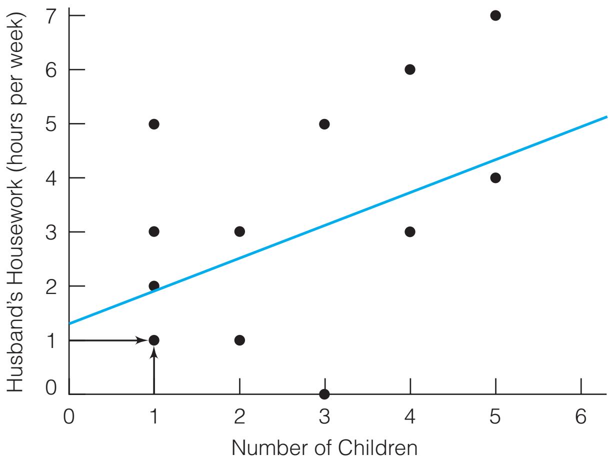
A scatterplot, like a bivariate table, has two dimensions. The scores of the independent (X) variable are arrayed along the horizontal axis, and the scores of the dependent (Y) variable along the vertical axis. Each dot on the scatterplot represents a case in the sample, and the dot's location represents the scores of the case on both variables.

Figure 12.1 shows a scatterplot displaying the relationship between "number of children" and "husband's housework" for the sample of 12 families presented in Table 12.1. Family A has a score of 1 on the X variable (number of children) and 1 on the Y variable (husband's housework) and is represented by the dot above the score of 1 on the X axis and directly to the right of the score of 1 on the Y axis. All 12 cases are similarly represented by dots on Figure 12.1. Also note that, as always, the scatterplot is clearly titled and both axes are labeled.

TABLE 12.1 Number of Children and Husband's Contribution to Housework (fictitious data)

Family	Number of Children	Hours per Week Husband Spends on Housework
A	1	1
B	1	2
C	1	3
D	1	5
E	2	3
F	2	1
G	3	5
H	3	0
I	4	6
J	4	3
K	5	7
L	5	4

FIGURE 12.1 Husband's Housework by Number of Children



Adding a "Freehand" Regression Line

The overall pattern of the dots or cases summarizes the nature of the relationship between the two variables. The clarity of the pattern can be enhanced by drawing a straight line through the cluster of dots such that the line touches every dot or comes as close to doing so as possible. We will soon consider a precise technique for fitting this line to the pattern of the dots but, for now, an "eyeball" approximation will suffice. This summarizing line, called the **regression line**, has already been added to the scatterplot.

Using Scatterplots to Answer the Three Questions

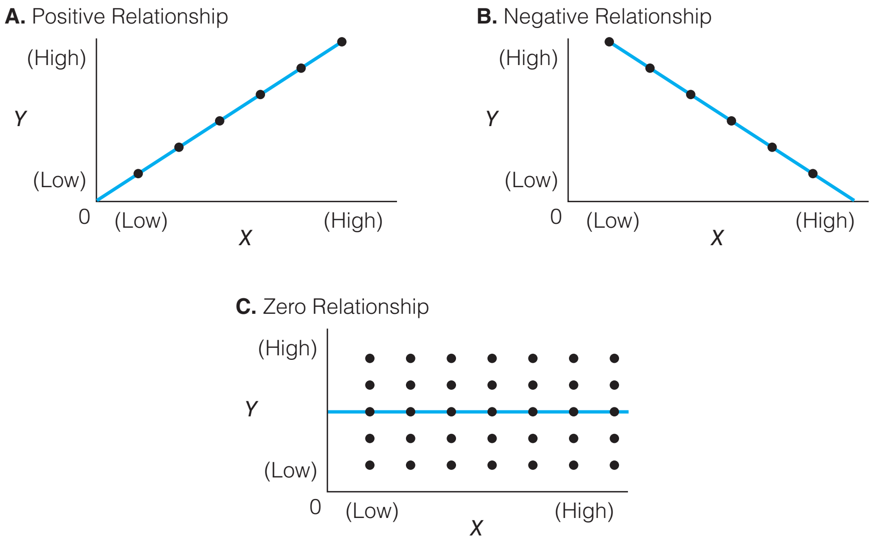
Scatterplots, even when they are crudely drawn, can be used for a variety of purposes. They provide at least impressionistic information about the existence, strength, and direction of the relationship and can also be used to check the relationship for linearity (that is, how well the dots approximate a straight line). Finally, the scatterplot can be used to predict the score of a case on one variable from the score of that case on the other variable. We examine these uses first in terms of the three questions we first asked in Chapter 11. Then, we will investigate matters of linearity and prediction.

- **Does a Relationship Exist?** The basic definition of an association was stated in Chapter 11: Two variables are associated if the distributions of Y (the dependent variable) change for the various conditions of X (the independent variable). In Figure 12.1, scores on X (number of children) are arrayed along the horizontal axis. The dots above each score on X are the scores (or conditional distributions) of Y. That is, the dots represent scores on Y for each value of X.

  There is a relationship between these variables, because the conditional distributions of Y (the dots above each score on X) change as X changes. The existence of an association is further reinforced by the fact that the regression line lies at an angle to the X axis. If these two variables had not been associated, the conditional distributions of Y would not have changed, and the regression line would have been parallel to the horizontal axis.
- **How Strong Is the Relationship?** The strength of the bivariate association can be judged by observing the spread of the dots around the regression line. In a perfect association, all dots would lie on the regression line. The more the dots are clustered around the regression line, the stronger the association.
- **What Is the Direction of the Relationship?** The direction of the relationship can be detected by observing the angle of the regression line. Figure 12.1 shows a positive relationship: As X (number of children) increases, husband's housework (Y) also increases. Husbands in families with more children tend to do more housework. If the relationship had been negative, the regression line would have sloped in the opposite direction to indicate that high scores on one variable were associated with low scores on the other.

To summarize these points, Figure 12.2 shows a perfect positive and a perfect negative relationship and a "zero relationship," or "nonrelationship," between two variables.

FIGURE 12.2 Perfect Positive, Perfect Negative, and Zero Relationships



Using Scatterplots to Check for Linearity

Correlation analysis with interval-ratio-level variables assumes that the variables have an essentially **linear relationship**. In other words, the observation points or dots in the scatterplot must form a pattern that can be approximated with a straight line. Significant departures from linearity would require the use of statistical techniques beyond the scope of this text. Examples of some common curvilinear relationships are presented in Figure 12.3.

If the scatterplot shows that the variables have a nonlinear relationship, the techniques described in this chapter should be used with great caution or not at all. Checking for the linearity of the relationship is perhaps the most important reason for examining the scatterplot before proceeding with the statistical

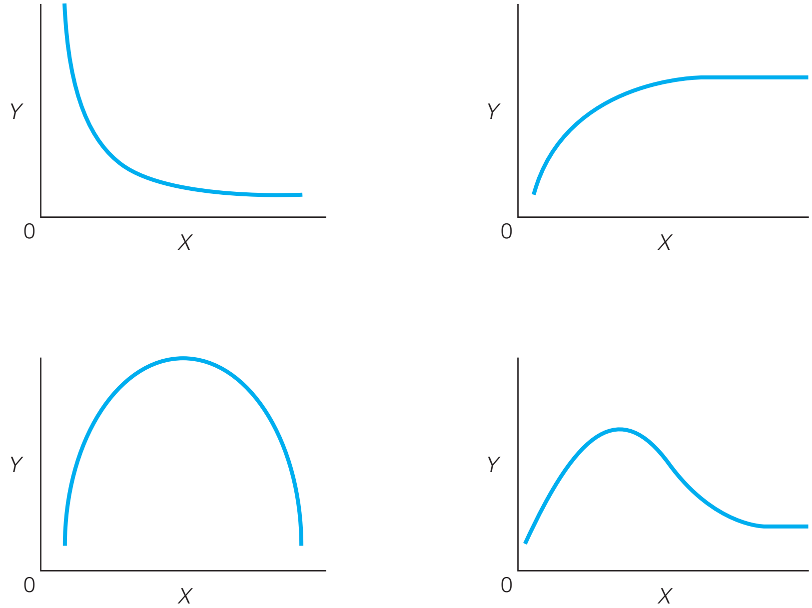
STATISTICS
IN EVERYDAY LIFE

Who Does the Housework?

The data used in our example are fictitious. Studies of the actual division of labor in households find large differences between the contributions of husbands and wives, regardless of number of children. For example, one study* investigated patterns of household labor and family satisfaction across 31 nations and found that women contributed significantly more time and effort to domestic chores, even when they were more involved in the paid labor force. The researchers also found that, although men in "traditional" relationships (where the man is the sole breadwinner) were the most satisfied with family life, satisfaction in all relationships depended heavily on both partners making contributions to housework and decision making.

*Forste, Renata and Fox, Kiira. 2012. "Household Labor, Gender Roles, and Family Satisfaction: A Cross-National Comparison." *Journal of Comparative Family Studies*. 43: 613–631.

FIGURE 12.3 Some Nonlinear Relationships



analysis. If the relationship is nonlinear, you might need to treat the variables as if they were ordinal rather than interval-ratio in level of measurement.

Using SPSS to Produce a Scatterplot

We can illustrate the features of a scatterplot with SPSS. We will use the *States* database and our dependent variable will be *TeenBirthRate* (births per 1000 females age 15–19 in 2011). The independent variable will be *edspend* (spending on education, grades K–12, 2011). Is there a correlation between support for education and the teen birth rate? Would you expect that states that spent more on education would have lower rates of teen births?

Follow these steps to get a scatterplot with the regression line:

1. Load the *States* database.
   a. Click **File** → **Open** → **Data**.
   b. Find the *States.sav* database.
2. From the menu bar across the top of the SPSS window, click **Graphs**, **Legacy Dialog**, and **Scatter/Dot**.
3. On the "Scatter/Dot" window, click **Simple Scatter** and then click **Define**.
4. The "Simple Scatterplot" window will appear.
   a. Find *TeenBirthRate* in the list of variables and click the arrow to move the variable name to the "Y Axis:" box.

b. Find *edspend* in the list of variables and click the arrow to move the variable name to the "X Axis:" box.

c. Click **OK** and the scatterplot will be sent to the SPSS output window.

5. To add the regression line,

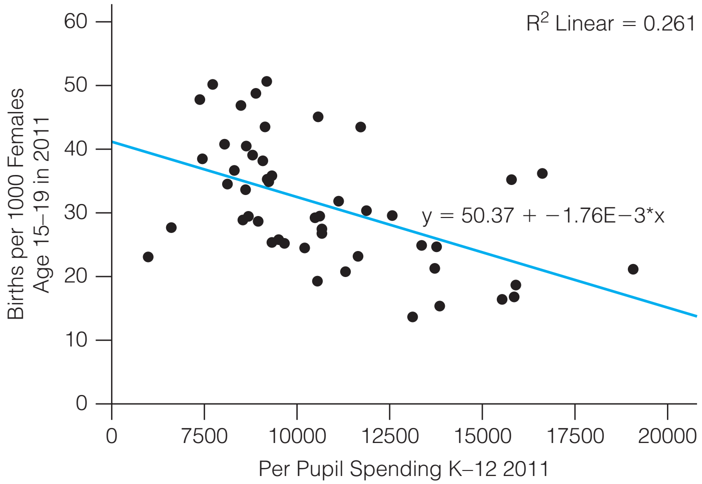
a. Double-click on the graph and the "Chart Editor" window will appear.

b. Click the **Elements** command at the top of the window and, in the drop-down menu, click **Fit Line at Total**.

c. A new window will open. In the "Fit Method" panel, click **Linear** and then click **Close**.

d. Close the "Chart Editor" window to return to the Output window of SPSS.

The scatterplot will look like this:

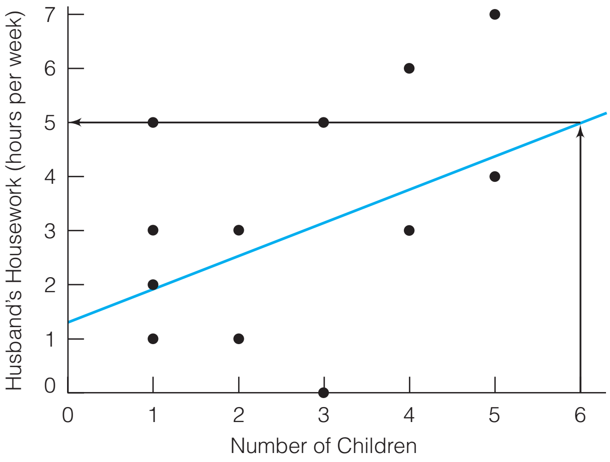


The dots are fairly widely spread around the regression line, which falls from left to right, so this is a moderate negative relationship. As support for education in a state increases, the teen birth rate decreases. We will identify and discuss the other elements in this graph (the equation in the center of the graph and the notation "R^2 Linear = 0.261" at the upper right) later in the chapter. *(For practice in constructing and interpreting scatterplots, see Problem 12.6)*

Using Scatterplots for Prediction

A final use of the scatterplot is to predict scores of cases on one variable from their score on the other. To illustrate, suppose that, based on the relationship between number of children and husband's housework displayed in Figure 12.1, we wish to predict the number of hours of housework a husband with a family of six children would do each week. The sample has no families with six children, but

FIGURE 12.4 Predicting Husband's Housework



if we extend the axes and regression line in Figure 12.1 to incorporate this score, a prediction is possible. Figure 12.4 reproduces the scatterplot and illustrates how the prediction would be made.

The predicted score on Y—which is symbolized as $\boldsymbol{Y'}$ ("Y-prime") to distinguish predictions of Y from actual Y scores—is found by first locating the relevant score on X ($X = 6$ in this case) and then drawing a straight line from that point to the regression line. From the regression line, another straight line parallel to the X axis is drawn across to the Y axis. The predicted Y score (Y') is found at the point where the line crosses the Y axis. In our example, we would predict that, in a dual-wage-earner family with six children, the husband would devote about 5 hours per week to housework.

The Regression Line

Predicting the score on Y with scatterplots and a freehand regression line can be crude and impressionistic. A better way to predict would be to find the straight line that most accurately describes the relationship between the two variables. How can we find this line?

Recall that our criterion for the freehand regression line was that it touch all the dots or come as close to doing so as possible. Also, recall that the dots above each value of X can be thought of as conditional distributions of Y, the dependent variable. Within each conditional distribution of Y, the mean is the point around which the variation of the scores is minimized. In Chapter 3, we noted that the mean of any distribution of scores is the point around which the variation of the scores, as measured by squared deviations, is minimized:

$$\sum (X_i - \overline{X})^2 = \text{minimum}$$

TABLE 12.2 Conditional Means of *Y* (husband's housework) for Various Values of *X* (number of children)

Number of Children (*X*)	Husband's Housework (*Y*)	Conditional Mean of *Y*
1	1, 2, 3 ,5	2.75
2	3, 1	2.00
3	5, 0	2.50
4	6, 3	4.50
5	7, 4	5.50

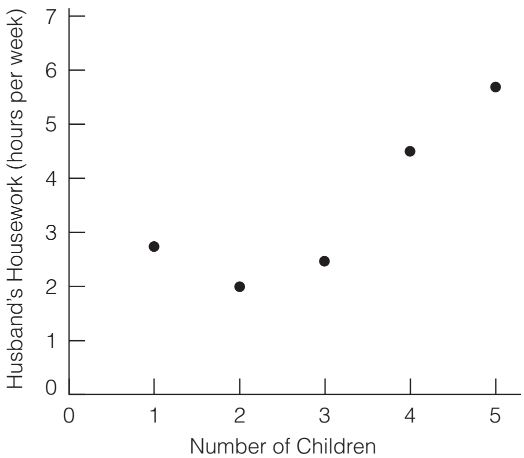
Thus, if the regression line were drawn so that it touches each **conditional mean of *Y***, it would be the straight line that comes as close as possible to all the scores.

Conditional means are found by summing all *Y* values for each value of *X* and then dividing by the number of cases. For example, four families had one child ($X = 1$), and the husbands of these four families devoted 1, 2, 3, and 5 hours to housework. The conditional mean of *Y* for $X = 1$ is 2.75 ($11/4 = 2.75$). Husbands in families with one child worked an average of 2.75 hours per week doing housekeeping chores. Conditional means of *Y* are computed in the same way for each value of *X*. They are displayed in Table 12.2 and plotted in Figure 12.5.

Let us quickly remind ourselves of the reason for these calculations. We are seeking the single best-fitting regression line for summarizing the relationship between *X* and *Y*, and we have seen that a line drawn through the conditional means of *Y* will minimize the spread of the observation points. It will come as close to all the scores as possible and will therefore be the single best-fitting regression line.

Now, a line drawn through the points on Figure 12.5 (the conditional means of *Y*) will be the best-fitting line we are seeking, but you can see from the scatterplot that the line will not be straight. In fact, only rarely (when there

FIGURE 12.5 Conditional Means of *Y*



is a perfect relationship between X and Y) will conditional means fall in a perfectly straight line.

Because we still must meet the condition of linearity, let us revise our criterion and define the regression line as the unique straight line that touches all conditional means of Y or comes as close to doing so as possible. Formula 12.1 defines this "least-squares" regression line, or the single straight regression line that best fits the pattern of the data points.

FORMULA 12.1

$$Y = a + bX$$

Where: Y = score on the dependent variable
a = the Y intercept, or the point where the regression line crosses the Y axis
b = the slope of the regression line, or the amount of change produced in Y by a unit change in X
X = score on the independent variable

The formula introduces two new concepts. First, the ***Y* intercept (*a*)** is the point at which the regression line crosses the vertical, or Y, axis. Second, the **slope (*b*)** of the least-squares regression line is the amount of change produced in the dependent variable (Y) by a unit change in the independent variable (X). Think of the slope of the regression line as a measure of the effect of the X variable on the Y variable: The greater the effect of X on Y, the higher the value of the slope (b). If the two variables are unrelated, the least-squares regression line would be parallel to the X axis, and b would be 0.00 (the line would have no slope).

With the least-squares formula (Formula 12.1), we can predict values of Y in a much less arbitrary and impressionistic way than through mere eyeballing. This will be so, remember, because the least-squares regression line is the single straight line that best fits the data and comes as close as possible to all of the conditional means of Y. Before seeing how predictions of Y can be made, however, we must first calculate a and b.

Computing and Interpreting the Regression Coefficients (*a* and *b*)

In this section, we cover the computation and interpretation of the coefficients in the equation for the regression line: the slope (b) and the Y intercept (a). We begin with the computation of the slope because the value of b is needed to compute a.

Computing the Slope (*b*). The formula for the slope is

FORMULA 12.2

$$b = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

The numerator of this formula is called the *covariation* of X and Y. It is a measure of how X and Y vary together, and its value will reflect both the direction and the strength of the relationship. The denominator is the sum of the squared deviations around the mean of X.

TABLE 12.3 Computation of the Slope (*b*)

1	2	3	4	5	6
X	$X - \overline{X}$	Y	$Y - \overline{Y}$	$(X - \overline{X})(Y - \overline{Y})$	$(X - \overline{X})^2$
1	−1.67	1	−2.33	3.89	2.79
1	−1.67	2	−1.33	2.22	2.79
1	−1.67	3	−0.33	0.55	2.79
1	−1.67	5	1.67	−2.79	2.79
2	−0.67	3	−0.33	0.22	0.45
2	−0.67	1	−2.33	1.56	0.45
3	0.33	5	1.67	0.55	0.11
3	0.33	0	−3.33	−1.10	0.11
4	1.33	6	2.67	3.55	1.77
4	1.33	3	−0.33	−0.44	1.77
5	2.33	7	3.67	8.55	5.43
5	2.33	4	0.67	1.56	5.43
32	−0.04	40	0.04	18.32	26.68

$$\overline{X} = \frac{32}{12} = 2.67$$

$$\overline{Y} = \frac{40}{12} = 3.33$$

The calculations necessary for computing the slope should be organized into a computational table, as in Table 12.3, which has a column for each of the quantities needed to solve the formula. The data are from the dual-wage-earner family sample (see Table 12.1).

The first column of the table lists the original X scores for each case, and the second column shows the deviations of these scores around their mean. The third and fourth columns repeat this information for the Y scores and the deviations of the Y scores. Column 5 shows the covariation of the X and Y scores. The entries in this column are found by multiplying the deviation of the X score (column 2) by the deviation of the Y score (column 4) for each case. Finally, the entries in column 6 are found by squaring the value in column 2 for each case. See the "One Step at a Time" box for detailed instructions on computing the slope.

Table 12.3 gives us all the quantities we need to solve Formula 12.2. Substitute the total of column 5 in Table 12.3 in the numerator and the total of column 6 in the denominator:

$$b = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\Sigma(X - \overline{X})^2}$$

$$b = \frac{18.32}{26.68}$$

$$b = 0.69$$

ONE STEP AT A TIME **Computing the Slope (*b*)**

Step	***Operation***
To compute the slope (b), solve Formula 12.2:	
1.	Use a computing table like Table 12.3 to help organize the computations. List the scores of the cases on the independent variable (X) in column 1.
2.	Compute the mean of X by dividing the total of column 1 by the number of cases.
3.	Subtract the mean of X from each X score and list the results in column 2. *NOTE: The sum of column 2 must be zero (except for rounding error). If this sum is not zero, you have made a mistake in computations.*
4.	List the score of each case on Y in column 3. Compute the mean of Y by dividing the total of column 3 by the number of cases.
5.	Subtract the mean of Y from each Y score and list the results in column 4. *NOTE: The sum of column 4 must be zero (except for rounding error). If this sum is not zero, you have made a mistake in computations.*
6.	For each case, multiply the value in column 2 by the value in column 4. Place the result in column 5. Find the sum of this column.
7.	Square each value in column 2 and place the result in column 6. Find the sum of this column.
8.	Divide the sum of column 5 by the sum of column 6. The result is the slope (b).

Interpreting the Slope (*b*). The value of the slope tells us the size of the change in Y for every unit change in X. In this case, the slope means that husbands increase their contribution to housework by 0.69 hour for each additional child.

Computing the *Y* Intercept (*a*). Once the slope has been calculated, finding the intercept (a) is relatively easy. We computed the means of X and Y while calculating the slope, and we enter these figures into Formula 12.3.

FORMULA 12.3

$$a = \bar{Y} - b\bar{X}$$

For our sample problem, the value of a would be

$$a = \bar{Y} - b\bar{X}$$
$$a = 3.33 - (0.69)(2.67)$$
$$a = 3.33 - 1.84$$
$$a = 1.49$$

Thus, the least-squares regression line will cross the Y axis at the point where Y equals 1.49.

Now that we have values for the slope and the Y intercept, we can state the full least-squares regression line for our sample data:

$$Y = a + bX$$
$$Y = 1.49 + (0.69)X$$

ONE STEP AT A TIME | **Computing the *Y* Intercept (*a*)**

Step	*Operation*
	To compute the Y intercept (a), solve Formula 12.3:
1.	Multiply the slope (*b*) by the mean of *X*.
2.	Subtract the value you found in step 1 from the mean of *Y*. This value is *a*, or the *Y* intercept.

Predicting Scores on *Y* with the Least-Squares Regression Line

The regression formula can be used to estimate, or predict, scores on *Y* for any value of *X*. Previously, we used the freehand regression line to predict a score on *Y* (husband's housework) for a family with six children ($X = 6$). Our prediction was that, in families of six children, husbands would contribute about 5 hours per week to housekeeping chores. By using the least-squares regression line, we can see how close our impressionistic, eyeball prediction was.

$$Y' = a + bX$$
$$Y' = 1.49 + (0.69)(6)$$
$$Y' = 1.49 + 4.14$$
$$Y' = 5.63$$

Based on the least-squares regression line, we would predict that in a dual-wage-earner family with six children, the husband would devote 5.63 hours a week to housework. What would our prediction of the husband's housework be for a family of seven children ($X = 7$)?

Note that our predictions of *Y* scores are "educated guesses." We will be unlikely to predict values of *Y* exactly except in the (rare) case where the bivariate relationship is perfect and perfectly linear. Note also, however, that the accuracy of our predictions will increase as relationships become stronger. This is because the dots are more clustered around the least-squares regression line in stronger relationships. *(For practice in computing b and a using the regression line to predict scores on Y from scores on X, see problems 12.1 to 12.3.)*

ONE STEP AT A TIME | **Using the Regression Line to Predict Scores on *Y***

Step	*Operation*
1.	Choose a value for *X*. Multiply this value by the value of the slope (*b*).
2.	Add the value you found in step 1 to the value of *a*, the *Y* intercept. The resulting value is the predicted score on *Y* (*Y′*).

The Correlation Coefficient (Pearson's *r*)

The slope of the least-squares regression line (b) is a measure of the effect of X on Y and it will increase as the effect of X increases. However, b does not vary between 0 and 1.00 and is therefore awkward to use as a measure of association. Instead, researchers rely heavily (almost exclusively) on a statistic called **Pearson's *r***, or the correlation coefficient, to measure association between interval-ratio variables. Like the ordinal measures of association discussed in Chapter 11, Pearson's r varies from 0.00 to ± 1.00, with 0.00 indicating no association and $+1.00$ and -1.00 indicating perfect positive and perfect negative relationships, respectively. The formula for Pearson's r is

FORMULA 12.4

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\left[\sum(X - \bar{X})^2\right]\left[\sum(Y - \bar{Y})^2\right]}}$$

Note that the numerator of this formula is the covariation of X and Y, as was the case with Formula 12.2. A computing table like Table 12.4, with a column added for the sum of $(Y - \bar{Y})^2$, is strongly recommended as a way of organizing the quantities needed to solve this equation (see Table 12.4).

For our sample problem involving dual-wage-earner families, the quantities displayed in Table 12.4 can be substituted directly into Formula 12.4:

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\left[\sum(X - \bar{X})^2\right]\left[\sum(Y - \bar{Y})^2\right]}}$$

$$r = \frac{18.32}{\sqrt{(26.68)(50.68)}}$$

$$r = \frac{18.32}{\sqrt{1352.14}}$$

$$r = \frac{18.32}{36.77}$$

$$r = 0.50$$

ONE STEP AT A TIME **Computing Pearson's *r***

Step	*Operation*
1.	Add a column to the computing table you used to compute the slope (b). Square the value of $(Y - \bar{Y})$ and record the result in column 7.
2.	Find the sum of column 7.
3.	Multiply the sum of column 6 by the sum of column 7.
4.	Take the square root of the value you found in step 3.
5.	Divide the quantity you found in step 4 into the sum of column 5, or the sum of the cross products. The result is Pearson's r.

TABLE 12.4 Computation of Pearson's *r*

1	2	3	4	5	6	7
X	$X-\bar{X}$	Y	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
1	−1.67	1	−2.33	3.89	2.79	5.43
1	−1.67	2	−1.33	2.22	2.79	1.77
1	−1.67	3	−0.33	0.55	2.79	0.11
1	−1.67	5	1.67	−2.79	2.79	2.79
2	−0.67	3	−0.33	0.22	0.45	0.11
2	−0.67	1	−2.33	1.56	0.45	5.43
3	0.33	5	1.67	0.55	0.11	2.79
3	0.33	0	−3.33	−1.10	0.11	11.09
4	1.33	6	2.67	3.55	1.77	7.13
4	1.33	3	−0.33	−0.44	1.77	0.11
5	2.33	7	3.67	8.55	5.43	13.47
5	2.33	4	0.67	1.56	5.43	0.45
32	−0.04	40	0.04	18.32	26.68	50.68

An r value of 0.50 indicates a moderately strong, positive linear relationship between the variables. As the number of children in the family increases, the hourly contribution of husbands to housekeeping duties also increases. *(For practice in computing and interpreting Pearson's r, see problems 12.1 to 12.5. For practice in using SPSS to produce and interpret Pearson's r, see problems 12.7 to 12.12.)*

Interpreting the Correlation Coefficient: r^2

Pearson's r is an index of the strength of the linear relationship between two variables. While a value of 0.00 indicates no linear relationship and a value of ±1.00 indicates a perfect linear relationship, values between these extremes have no direct interpretation. We can, of course, describe relationships in terms of how closely they approach the extremes (for example, coefficients approaching 0.00 can be described as "weak" and those approaching ±1.00 as "strong"), but this description is somewhat subjective. Also, we can use the guidelines stated in Table 11.12 for ordinal-level measures of association to describe strength in terms of weak, moderate, and strong. Remember, of course, that these labels are just guidelines and will not be appropriate or useful in all possible research situations.

The Coefficient of Determination

Fortunately, we can use a more direct interpretation of r that involves calculating an additional statistic called the **coefficient of determination**. This statistic, which is simply the square of Pearson's r (r^2), can be interpreted using a logic similar to that of proportional reduction in error (PRE). As you recall, the logic of PRE measures of association is to predict the value of the dependent variable under two different conditions. First, Y is predicted while ignoring the information supplied by X; second, the independent variable is taken into account. With r^2, both the method of prediction and the construction of the final statistic require the introduction of some new concepts.

Predicting *Y* Without *X*

When working with variables measured at the interval-ratio level, the predictions of the Y scores under the first condition (while ignoring X) will be the mean of the Y scores. Given no information on X, this prediction strategy will be optimal because we know that the mean of any distribution is closer to all the scores than any other point in the distribution. Recall the principle of minimized variation introduced in Chapter 3 and expressed as

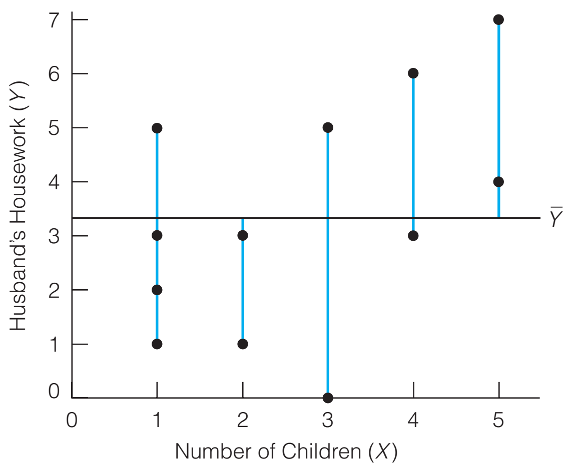
$$\Sigma(X_i - \overline{X})^2 = \text{minimum}$$

The scores of any variable vary less around the mean than around any other point. If we predict the mean of Y for every case, we will make fewer errors of prediction than if we predict any other value for Y.

Of course, we will still make many errors in predicting Y even if we faithfully follow this strategy. The amount of error is represented in Figure 12.6, which displays the relationship between number of children and husband's housework, with the mean of Y noted. The vertical lines from the actual scores to the predicted scores represent the amount of error we would make when predicting Y while ignoring X.

We can define the extent of our prediction error under the first condition (while ignoring X) by subtracting the mean of Y from each actual Y score and squaring and summing these deviations. The resultant sum, which can be noted as $\Sigma(Y - \overline{Y})^2$, is called the **total variation** in Y. We now have a visual representation (Figure 12.6) and a method for calculating the error we incur by predicting Y without knowledge of X. As we shall see, we do not actually need to calculate the total variation to find the value of the coefficient of determination, r^2.

FIGURE 12.6 Predicting *Y* Without *X* (dual-career families)



Predicting *Y* with *X*

Our next step will be to determine the extent to which knowledge of X improves our ability to predict Y. If the two variables have a linear relationship, then predicting scores on Y from the least-squares regression equation will incorporate knowledge of X and reduce our errors of prediction. So, under the second condition, our predicted Y score for each value of X will be

$$Y' = a + bX$$

Figure 12.7 displays the data from the dual-career families with the regression line drawn in. The vertical lines from each data point to the regression line represent the amount of error in predicting Y that remains even after X has been taken into account.

Explained, Unexplained, and Total Variation

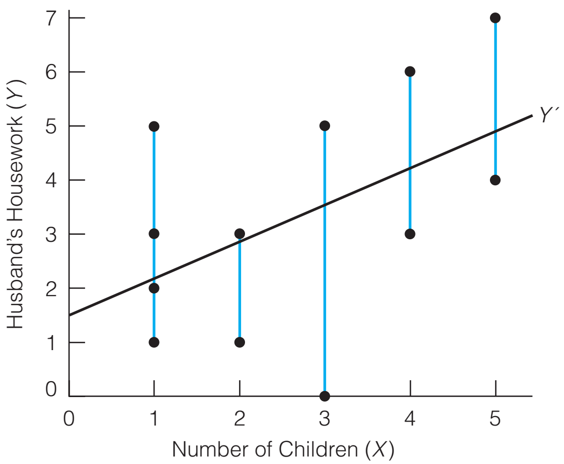
We can precisely define the reduction in error that results from taking X into account. Specifically, two different sums can be found and then compared with the total variation of Y to construct a statistic that will indicate the improvement in prediction.

The first sum, called the **explained variation**, represents the improvement in our ability to predict Y when taking X into account. This sum is found by subtracting $\bar{Y}$ (our predicted Y score without X) from the score predicted by the regression equation (or the Y score predicted with knowledge of X) for each case and then squaring and summing these differences. These operations can be summarized as $\sum(Y' - \bar{Y})^2$ and the resultant figure could then be compared with the total variation in Y to ascertain the extent to which our knowledge of X improves our ability to predict Y. Specifically, it can be shown mathematically that

FORMULA 12.5

$$r^2 = \frac{\sum(Y' - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

FIGURE 12.7 Predicting *Y* with *X* (dual-career families)



Thus, the coefficient of determination, or r^2, is the proportion of the total variation in Y attributable to, or explained by, X. Like other PRE measures, r^2 indicates precisely the extent to which X helps us predict, understand, or explain Y.

We refer to the improvement in predicting Y with X as the explained variation. The use of this term suggests that some of the variation in Y will be "unexplained," or not attributable to the influence of X. In fact, the vertical lines in Figure 12.7 represent the **unexplained variation**, or the difference between our best prediction of Y with X and the actual scores. The unexplained variation is thus the scattering of the actual scores around the regression line; it can be found by subtracting the predicted Y score from the actual Y score for each case and then squaring and summing the differences. These operations can be summarized as $\sum(Y - Y')^2$, and the resultant sum would measure the amount of error in predicting Y that remains even after X has been taken into account. The proportion of the total variation in Y unexplained by X can be found by subtracting the value of r^2 from 1.00. Unexplained variation is usually attributed to the influence of some combination of other variables, measurement error, and random chance.

As you may have recognized by this time, the explained and unexplained variations bear a reciprocal relationship with one another. As one increases in value, the other decreases. Also, the stronger the linear relationship between X and Y, the greater the value of the explained variation and the lower the unexplained variation. In the case of a perfect relationship ($r = \pm 1.00$), the unexplained variation would be 0 and r^2 would be 1.00. This would indicate that X explains, or accounts for, all the variation in Y and that we could predict Y from X without error.

On the other hand, when X and Y are not linearly related ($r = 0.00$), the explained variation would be 0 and r^2 would be 0.00. In such a case, we would conclude that X explains none of the variation in Y and does not improve our ability to predict Y.

Relationships intermediate between these two extremes can be interpreted in terms of how much X increases our ability to predict, or explain, Y. For the dual-career families, we calculated an r of 0.50. Squaring this value yields a coefficient of determination of 0.25 ($r^2 = 0.25$), which indicates that number of children (X)

ONE STEP AT A TIME Interpreting the Strength and Direction of Pearson's *r*

To interpret the strength of Pearson's r:

You can use Table 11.12 to describe strength in general terms but the preferred method would be to square the value of r and multiply by 100. This value represents the percentage of variation in Y that is explained by X. In our example problem, the number of children explained 25% ($r^2 = 0.25$) of the variation in husband's housework.

To interpret the direction of the relationship:

If r has a plus sign (or if there is no sign), the relationship is positive and the variables change in the same direction. If r has a minus sign, the relationship is negative and the variables change in opposite directions. The relationship in our example problem was positive: as the number of children increased, the number of hours husbands contributed to housework also increased.

explains 25% of the total variation in husband's hours of housework (Y). When predicting the number of hours per week that husbands in such families would devote to housework, we will make 25% fewer errors by basing the predictions on number of children and predicting from the regression line, as opposed to ignoring the X variable and predicting the mean of Y for every case. Also, 75% of the variation in Y is unexplained by X and presumably due to some combination of the influence of other variables, measurement error, and random chance. *(For practice in the interpretation of r^2, see any of the problems at the end of this chapter.)*

Applying Statistics 12.1 Computing and Interpreting the Regression Coefficients and Pearson's *r*

Information has been collected on unemployment rate (X) and number of civil disturbances (Y)—including riots and strikes—for five cities. Are these variables associated? The data are presented in the following table. Columns have been added for all sums necessary for the computation of the slope (b) and Pearson's r.

City	(X)	$(X-\bar{X})$	(Y)	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
A	14	4	12	5	20	16	25
B	13	3	10	3	9	9	9
C	10	0	8	1	0	0	1
D	8	−2	5	−2	4	4	4
E	5	−5	0	−7	35	25	49
	50	0	35	0	68	54	88

$\bar{X} = 10$
$\bar{Y} = 7$

The slope (b) is

$$b = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2}$$

$$b = \frac{68}{54}$$

$$b = 1.26$$

A slope of 1.26 means that for every unit change in X (for every increase of 1 in the unemployment rate), there was a change of 1.26 units in Y (the number of civil disturbances increased by 1.26).

The Y intercept (a) is

$$a = \bar{Y} - b\bar{X}$$
$$a = 7 - (1.26)(10)$$
$$a = 7 - 12.60$$
$$a = -5.60$$

The least-squares regression equation is

$$Y = a + bX = -5.60 + (1.26)X$$

The correlation coefficient is

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{[\Sigma(X-\bar{X})^2][\Sigma(Y-\bar{Y})^2]}}$$

$$r = \frac{68}{\sqrt{(54)(88)}}$$

$$r = \frac{68}{\sqrt{4752}}$$

$$r = \frac{68}{68.93}$$

$$r = 0.99$$

These variables have a strong (nearly perfect), positive association. The number of civil disturbances increases as the unemployment rate increases. The coefficient of determination, r^2, is $(0.99)^2$, or 0.98. This indicates that 98% of the variance in civil disturbances is explained by the unemployment rate.

Using SPSS to Conduct a Regression Analysis and Calculate Pearson's *r* and r^2

As you would expect, SPSS performs the calculations required for a regression analysis quickly and effortlessly. We will illustrate using the International Population database (*Intl-POP*) supplied with this text and the **Regression** program, one of two programs we will use in this chapter and in Chapter 13 to investigate correlations. We will focus on the relationship between life expectancy (*LifeExp*) and gross national income per capita (*GNICap*), a measure of affluence.

To begin,

1. Click the SPSS icon on your desktop.
2. Load the *Intl-POP* database.
   a. Find the **File** command on the far left of the menu bar and then click **File → Open → Data**.
   b. Find the *Intl-POP* database file supplied with this text. You can download this file from the website for this text if you haven't already.
3. From the main menu bar, click **Analyze**, **Regression**, and **Linear**.
4. In the "Linear Regression" dialog box, find *LifeExp* and move it to the "Dependent:" window. Next, find *GNIcap* and move it to the "Independent(s):" window. Click **Continue** to return to the "Linear Regression" dialog box.
5. Click **OK** and the output will be printed to the SPSS Output screen.

The **Regression** program produces a lot of output, only some of which is relevant for us. The "Model Summary" box reports an r of 0.750 under R and an r^2 of 0.562 under R Square. There is a strong, positive relationship between these variables and level of affluence (*GNIcap*) explains 56% of the variance in life expectancy for these nations.

In the "Coefficients" box, the Y intercept (a) is reported in the first row of the column labeled B as the Constant and the slope (b) is reported in the second. The regression line for this relationship is

$$Y = 61.252 + (.001)X$$

We would predict that a nation with a low per capita gross national income of \$8,000 ($X = 8{,}000$) would have a life expectancy of 69.25 years. A nation with twice the per capita gross national income of \$16,000 ($X = 16{,}000$) would have a life expectancy of 77.25 years.

The Correlation Matrix

Social science research projects include many variables, and the data analysis phase of a project often begins with the examination of a **correlation matrix**, a table that shows the relationships between all possible pairs of variables. The correlation matrix gives a quick, easy-to-read overview of the interrelationships in the data set and may suggest strategies or "leads" for further analysis. These tables are commonly included in the professional research literature, and it will be useful to have some experience reading them.

Using SPSS to Produce a Correlation Matrix

In this installment of Using SPSS, we will produce and interpret a correlation matrix, using the *States* database supplied with this text. We will use a procedure called **Correlate**, which is somewhat easier to use than the **Regression** procedure but does not produce the regression coefficients (a and b). We used this procedure in Problem 11.15 to find Spearman's rho.

To begin,

1. Click the SPSS icon on your desktop.

2. Load the *States* database.

  a. Find the **File** command on the far left of the menu bar and then click **File → Open → Data**.

  b. Find the *States* database file supplied with this text. You can download this file from the website for this text if you haven't already.

3. From the main menu bar, click **Analyze**, **Correlate**, and **Bivariate**. Pearson's r will be preselected.

4. In the "Bivariate Correlations" dialog box, find *BirthRate, College*, *FamPoor09*, and *TeenBirthRate* and move them to the "Variables" window.

5. Click **OK** and the output will be printed to the SPSS Output screen.

The matrix presents three pieces of information for each bivariate relationship: Pearson's r, the statistical significance or "Sig. (2-tailed)" of the relationship, and N or the number of cases. Since the 50 states are not a random sample and the number of cases is always 50, we can focus only on the correlations. An abbreviated form of the matrix is presented in Table 12.5.

Table 12.5 A Correlation Matrix Showing the Relationship Between Four Variables

	1	2	3	4
	Birth Rate	Education	Poverty	Teen Births
1. Birth Rate	1.00	−0.19	0.10	0.40
2. Education	−0.19	1.00	−0.71	−0.76
3. Poverty	0.10	−0.71	1.00	0.81
4. Teen Births	0.40	−0.76	0.81	1.00

Key: "Birth Rate" is number of births per 1000 population.
"Education" is percentage of the population with a college degree or more.
"Poverty" is percentage of families below the poverty line in 2009.
"Teen Births" is births per 1000 females age 15–19.

Analyzing a Correlation Matrix

The matrix uses variable names as rows and columns, and the cells in the table show the bivariate correlation (Pearson's *r*) for each combination of variables. Note that the row headings duplicate the column headings. To read the table, begin with birth rate, the variable in the far left-hand column (column 1) and top row (row 1). Read down column 1 or across row 1 to see the correlations of this variable with all other variables, including the correlation of birth rate with itself (1.00) in the top left-hand cell. To see the relationships between other variables, move from column to column or row to row.

Note that the diagonal from upper left to lower right of the matrix contains the correlation of each variable with itself. Values along this diagonal will always be 1.00 and, since this information is not useful, it could easily be deleted from the output.

Also note that the cells below and to the left of the diagonal duplicate the cells above and to the right. For example, look at the second cell down (row 2) in column 1. This cell displays the correlation between birth rate and percentage of the population that is college educated, as does the cell in the top row (row 1) of column 2. In other words, the cells below and to the left of the diagonal are mirror images of the cells above and to the right of the diagonal. Commonly, research articles in the professional literature will delete the redundant cells in order to make the table more readable.

What does this matrix tell us? Starting at the upper left of the table (column 1), we can see that birth rate has a weak to moderate negative relationship with the level of college education. This means that birth rate decreases as education increases: States with a higher percentage of college educated populations have lower birth rates. Continuing down column 1, we see that birth rate has a weak positive relationship with poverty (birth rates increase as poverty increases) and a moderate positive relationship with teen births (states with higher birth rates tend to have higher rates of teen births).

To assess the other relationships in the data set, move from column to column and row to row, one variable at a time. For each subsequent variable, there will be one fewer cell of new information. For example, consider education (the percentage of the state's population that has at least a college degree), the variable in column 2 and row 2. We have already noted its weak to moderate negative relationship with birth rate, and, of course, we can ignore the correlation of the variable with itself. This leaves only two new relationships, which can be read by moving down column 2 or across row 2. Education has a strong negative relationship with poverty (the higher the percentage of the population that is college educated, the lower the rate of poverty), and a strong negative relationship with teen births (more educated states tend to have fewer teen births).

For poverty, the variable in column 3, there is only one new relationship: a strong positive correlation with teen births (the higher the poverty, the higher the rate of teen births). For column 4 (teen births), there are no new relationships to consider: We have already examined all correlations between this variable and all other variables.

Correlation and Causation

As we have noted on several occasions, correlation and causation are two different things. We can see some strong relationships in Table 12.5 and it is tempting to conclude that some of these variables are causing others. For example, it is probably not surprising to see a strong relationship between education and poverty but the question of causation is not clear. Do low levels of education cause poverty or does poverty cause low levels of education? Which variable is cause and which is effect?

How can we build a case for a causal relationship between variables like poverty and education? What evidence is needed? The first consideration is the value of the measure of association. The weaker the measure, the more difficult it is to argue for a causal relationship. By this criterion, the strong correlation between poverty and education is consistent with the argument that a causal relationship exists.

The second consideration is time order: To argue for causation, we must be able to demonstrate that the independent variable occurs before the dependent variable. For example, we might be able to demonstrate that rising levels of poverty over a period of years was followed by a decline in the number of people enrolled in college: This also would be consistent with the argument that a causal relationship exists.

The final step in building a case for a causal relationship is to demonstrate that no third variable affects the bivariate relationship. To continue with our example, we must be able to demonstrate that levels of education and poverty would be related regardless of other economic or cultural changes (e.g., the unemployment rate, changing family size, crime rates, or a host of other variables). Satisfying this criterion requires multivariate analysis, so we will defer a discussion until the next chapter, although we briefly consider the matter again in the "Becoming a Critical Consumer" box at the end of this chapter.

In conclusion, remember that correlation is a necessary but not sufficient condition for causation. If a correlation is zero or very weak, the relationship is almost certainly not causal. Even with very strong correlations, however, we must deal with several other considerations before we can argue that one variable causes another.

Correlation and Statistical Significance

Remember, also, that correlation and statistical significance are two different things. When we are working with randomly-selected samples, like the 2012 GSS, we must evaluate the probability that the bivariate relationship exists in the population as well as in the sample. There is a test of significance for Pearson's r but, rather than consider it, I will simply point out that SPSS automatically tests for significance every time we run the **Correlate** or **Regression** programs.

As I noted earlier, the **Correlate** program presents the statistical significance as "Sig. (2-tailed)" in the second row of each cell of the correlation matrix.

STATISTICS
IN EVERYDAY LIFE

Teen Pregnancies

Table 12.5 shows a strong relationship between poverty and teen births: the poorer the state, the higher the rate of births to teens. Besides the association with poverty, the rate of teen pregnancy tends to be strongly correlated with a number of other variables. For example, teens are less likely to get adequate prenatal care and babies born to teen moms are more likely to be underweight and have other health problems. Hence, it is good news to hear that the rates of teen pregnancy have been declining regularly for the past 20 years,* although it is still higher in the United States than in many other developed nations. In 1990, there were about 60 births for every 1000 females age 15–19. In 2014, the rate had fallen dramatically, to about 26. Does this decline reflect increased abstinence, more access to birth control, or some other factors? How could you find out?

**Source:* Office of Adolescent Health. 2014. "Trends in Teen Pregnancy and Childbearing." Department of Health and Human Services. Accessed at http://www.hhs.gov/ash/oah/adolescent-health-topics/reproductive-health/teen-pregnancy/trends.html

We can ignore this information if, like the *States* database, the sample has not been randomly selected from a larger population. On the other hand, if we are working with a probability sample like the *GSS2012*, we would need to include this information in our analysis.

For the **Regression** program, we can use the ANOVA test to evaluate statistical significance when appropriate. This information is presented in the ANOVA box, in the column labeled "Sig."

With either program, remember that even very strong relationships may not be statistically significant and vice versa. Generally, when working with probability samples, we seek relationships that are strong, in the predicted direction, and statistically significant but it is common to get ambiguous or mixed results.

Correlation, Regression, Level of Measurement, and Dummy Variables

Correlation and regression are very powerful and useful techniques, so much so that they are often used to analyze relationships between ordinal-level variables. This practice is generally not a problem, especially with ordinal variables that are "continuous" (see Chapter 11). However, this flexibility does not extend to nominal-level variables such as religious denomination or gender. The scores of nominal variables are not numbers and have no mathematical quality. We might represent a Protestant with a score of "2" and a Catholic with a score of "1," but the former score is not "twice as much" as the latter. The scores of nominal-level variables are labels, not numbers, and it makes no sense to compute a slope or to discuss positive or negative relationships for these variables.

This is an unfortunate situation. Many of the variables that are most important in everyday social life—gender, marital status, race, and ethnicity—are nominal in level of measurement and cannot be included in a regression equation or a correlational analysis, two of the most powerful and sophisticated tools available for social science research.

Fortunately, researchers have developed a way to change the scoring on nominal-level variables by creating **dummy variables**. Dummy variables can be of any level of measurement, including nominal, and have exactly two categories, one coded as 0 and the other coded as 1. Treated this way, nominal-level variables, such as gender (for example, with males coded as 0 and females coded as 1), race (with whites coded as 0 and nonwhites as 1), and religious denomination (Catholics coded as 0 and non-Catholics coded as 1), are commonly included in regression equations.

Using SPSS to Produce a Regression Analysis with a Dummy Variable

To illustrate, we will use SPSS to analyze the relationship between region and teen pregnancy. Suppose that we suspect that the southern states had a higher rate of teen pregnancy. In Table 12.5, we saw that "Teen Births" had a strong negative relationship with education and a strong positive relationship with poverty. What is the relationship with region? We can find out by using region as a dummy variable, coding the South as 0 and all non-Southern states as 1. We will use the **Regression** procedure since, unlike **Correlate**, it produces the regression coefficients (*a* and *b*).

To begin,

1. Click the SPSS icon on your desktop.
2. Load the *States* database.
   a. Find the **File** command on the far left of the menu bar and then click **File → Open → Data**.
   b. Find the *States* database file supplied with this text.
3. From the main menu bar, click **Analyze, Regression**, and **Linear**.
4. On the "Linear Regression" screen, place *TeenBirthRate* in the "Dependent" window and *SthDUMMY* in the "Independent" window.
5. Click **Statistics** and check **Descriptives**. Click **Continue** to return to the "Linear Regression" screen.
6. Click **OK** and results will be printed to the Output screen in several separate boxes, only some of which are relevant for our purposes. To conserve space, the output boxes will not be reproduced here.

7. The "Correlations" box displays the correlation between the variables ($r = -0.64$)
8. The "Coefficients" box displays the value of *a* and *b* under "Unstandardized Coefficients." Look in the column labeled B. The value in the first row is *a* and the entry in the second row is *b*.

The output shows a strong negative relationship ($r = -0.64$) between the region dummy variable and teen births. The r^2 is 0.41, which means that region explains 41% of the variation in teen births.

Turning to the regression equation, we can use the values for *a* and *b* in the "Coefficients" output box to write the regression equation:

$$Y = a + bX$$

$$Y = 41.09 + (-13.65)X$$

"Teen Birth Rate" is the dependent (or *Y*) variable and region (*X*) is the independent variable. The regression line crosses the vertical axis of the scatterplot at the point where $Y = 41.09$. The value for the slope ($b = -13.65$) indicates a negative relationship: As region "increases" (or moves toward the higher score associated with non-Southern states), the percentage of teen births tends to decrease. In other words, the states outside the South averaged lower rates of births to teens. Note that the *sign* of the slope (*b*) would have been positive had we reversed the coding scheme and labeled "South" as 1 and non-South as 0, but the *value* of *b* would have stayed exactly the same. The coding scheme for dummy variables is arbitrary, and, as with ordinal-level variables, the researcher needs to be clear about what the values of a dummy variable indicate. (*For experience in working with dummy variables, see problem 12.8.*)

Applying Statistics 12.2 Fertility and Women's Education

What happens to birth rate when women are more educated? Do more-educated women choose adult roles other than wife and mother and reduce their number of children? We can get an idea about these relationships with the International Population database (*Intl-POP.sav*) supplied with this text.

The variables are

- Total fertility rate (*TFR*), the average number of children born to a women during her lifetime, and
- The level of schooling for females is measured by *EducFemales*, the ratio of the number of females enrolled in any post-secondary educational programs to the number of females in the appropriate age group (generally 18–22). The higher this number, the greater the extent of education for females in the nation. This number can exceed 100% if the total enrollment of all age groups exceeds the number of females in the designated age group (e.g., 18–22).

Start SPSS, load the *Intl-POP* database, and use the **Correlate** command (see "Using SPSS: Producing a Correlation Matrix"). Enter the variable names in the "Correlate" window and click **OK**.

The strong negative relationship ($r = -0.78$) between the variables shows that nations with more educated women have lower fertility rates. This is consistent with the idea that the number of children decreases as the level of education for women increases.

BECOMING A CRITICAL CONSUMER: Correlation, Causation, and Cancer

Causation—how variables affect each other—is a central concern in everyday life as well as in science. For example, you might hear a news commentator say that a downturn in the economy will lead to higher crime rates, or that higher gasoline prices will cause people to change their driving habits. How can we judge the credibility of arguments that one variable causes another?

We have already noted the three criteria that are used to argue for a causal relationship (a reasonably strong measure of association, time order, and the effect of third variables). How can we apply these to the suggested associations we encounter in everyday life?

Consider the relationship between smoking tobacco and cancer. Today, virtually everyone knows that one causes the other but this information was not part of the common wisdom just a few generations ago. As recently as the 1950s, about half of all men smoked, and smoking was equated with sophistication and mature adulthood, not illness and disease. Since that time, medical research has established the causal case and the connection has been widely broadcast. The effect has been dramatic: now, less than 20% of adults smoke.

Measures of association like Pearson's *r* played an important role in building the case for causation. Both the strength and the direction of the statistic consistently pointed to a meaningful correlation between smoking and disease.

To demonstrate the time order between the variables, the best, most powerful studies followed groups of individuals over long periods of time. The studies started with a large number of smokers in good health. If the smokers developed cancer later in the study, cancer must be the dependent variable. That is, given the time order, smoking could cause cancer but the reverse could not be true: getting cancer could not have caused people to start smoking.

By itself, time order doesn't prove that the relationship is causal. Other variables might have been involved and might explain both why people smoke and why they get cancer. For example, maybe very anxious people smoke to "calm their nerves" and are more likely to contract cancer because of their anxiety, not because of their smoking. If this were the case, the relationship between smoking and cancer would be spurious or false: what looks like a bivariate, causal relationship is actually caused by anxiety.

How did the researchers respond to this possibility? The most convincing studies use various means to control for and examine the effect of third variables—like anxiety, family history, gender, race, and so forth—and were able to statistically eliminate the possibility that these factors affected the bivariate relationship. We will examine some of the multivariate statistics used by these studies in the next chapter.

Whenever you encounter a claim that one variable causes another, think through these questions: How strong is the measure of association? Is it certain the independent variable occurred first? What other variables might affect the relationship? Have they been accounted for? The greater the extent to which a relationship satisfies these criteria, the stronger the case for a causal relationship.

SUMMARY

This summary is based on the example used throughout the chapter.

1. We began with a question: Is the number of children in dual–wage-earner families related to the number of hours per week husbands devote to housework? We presented the observations in a scatterplot (Figure 12.1), and our visual impression was that the variables were associated in a positive direction. The pattern formed by the observation points in the scatterplot could be approximated with a straight line; thus, the relationship was roughly linear.
2. The least-squares regression line ($Y = a + bX$) is the single straight line that best fits the data because it minimizes the variation in Y. In our example, we found a slope (b) of 0.69, which indicates that each additional child (a unit change in X) is accompanied by an increase of 0.69 hour of housework per week for the husbands. We also predicted, based on this formula, that in a dual–wage-earner family with six children ($X = 6$), husbands would contribute 5.63 hours of housework a week ($Y = 5.63$ for $X = 6$).

3. Pearson's r is a statistic that measures the overall linear association between X and Y. Our impression from the scatterplot of a substantial positive relationship was confirmed by the computed r of 0.50. We also saw that this relationship yields an r^2 of 0.25, which indicates that 25% of the total variation in Y (husband's housework) is accounted for, or explained, by X (number of children).
4. We acquired a great deal of information about this bivariate relationship. We know the strength and direction of the relationship and have also identified the regression line that best summarizes the effect of X on Y. We know the amount of change we can expect in Y for a unit change in X. In short, we have a greater volume of more precise information about this association between interval-ratio variables than we ever did about associations between ordinal or nominal variables. This is possible, of course, because the data generated by interval-ratio measurement are more precise and flexible than those produced by ordinal or nominal measurement techniques.

SUMMARY OF FORMULAS

FORMULA 12.1	Least-squares regression line:	$Y = a + bX$
FORMULA 12.2	Slope (b):	$b = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$
FORMULA 12.3	Y intercept (a):	$a = \bar{Y} - b\bar{X}$
FORMULA 12.4	Pearson's r:	$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\left[\sum(X - \bar{X})^2\right]\left[\sum(Y - \bar{Y})^2\right]}}$
FORMULA 12.5	Coefficient of determination:	$r^2 = \frac{\sum(Y' - \bar{Y})}{\sum(Y - \bar{Y})} = \frac{\text{Explained Variation}}{\text{Total Variation}}$

GLOSSARY

Coefficient of determination (r^2). The proportion of all variation in Y that is explained by X. Found by squaring the value of Pearson's r.

Conditional mean of Y. The mean of all scores on Y for each value of X.

Correlation matrix. A table that shows the correlation coefficients between all possible pairs of variables.

Dummy variable. A nominal-level variable dichotomized so that it can be used in regression analysis. A dummy variable has two scores, one coded as 0 and the other as 1.

Explained variation. The proportion of all variation in Y that is attributed to the effect of X. Equal to $\sum(Y' - \bar{Y})^2$.

Linear relationship. A relationship between two variables in which the observation points (dots) in the scatterplot can be approximated with a straight line.

Pearson's r. A measure of association for variables that have been measured at the interval-ratio level.

Regression line. The single best-fitting straight line that summarizes the relationship between two variables. Regression lines are fitted to the data points by the least-squares criterion, whereby the line touches all conditional means of Y or comes as close to doing so as possible.

Scatterplot. A graph that depicts the relationship between two variables.

Slope (b). The amount of change in one variable per unit change in the other; b is the symbol for the slope of a regression line.

Total variation. The spread of the Y scores around the mean of Y. Equal to $\sum(Y - \bar{Y})^2$.

Unexplained variation. The proportion of the total variation in Y that is not accounted for by X. Equal to $\sum(Y - Y')^2$.

Y intercept (a). The point where the regression line crosses the Y axis.

Y'. Symbol for predicted score on Y.

PROBLEMS

The datasets in the first five problems are unrealistically small to allow you to practice the mathematical routines required to conduct correlational and regression analysis. Most problems use SPSS.

12.1 PS Why does voter turnout vary from election to election? For municipal elections in five different cities, information has been gathered on the percentage of citizens who voted (the dependent variable) and three different independent variables: unemployment rate, average years of education for the city, and the percentage of all political ads that used "negative campaigning" (personal attacks, negative portrayals of the opponent's record, etc.). For each relationship between turnout (Y) and the independent variables,

a. Compute the slope (b) and find the Y intercept (a). (*HINT: Remember to compute* b *before computing* a. *A computing table such as Table 12.3 is highly recommended.*)
b. State the least-squares regression line and predict the voter turnout for a city in which the unemployment rate was 12%, a city in which the average years of schooling was 11, and an election in which 90% of the ads were negative.
c. Compute r and r^2. (*HINT: A computing table such as Table 12.4 is highly recommended. If you constructed one for computing* b, *you already have most of the quantities you will need to solve for* r.)
d. Describe the strength and direction of each relationship in a sentence or two. Which of the independent variables had the strongest effect on turnout?

City	Turnout (Y)	Unemployment Rate (X_1)	Average Years of School (X_2)	Percentage of Negative Ads (X_3)
A	55	5	11.9	60
B	60	8	12.1	63
C	65	9	12.7	55
D	68	9	12.8	53
E	70	10	13.0	48

12.2 SOC Occupational prestige scores for a sample of fathers and their oldest son and oldest daughter are presented in the following table.

Family	Father's Prestige	Son's Prestige	Daughter's Prestige
A	80	85	82
B	78	80	77
C	75	70	68
D	70	75	77
E	69	72	60
F	66	60	52
G	64	48	48
H	52	55	57

Analyze the relationship between father's and son's prestige and the relationship between father's and daughter's prestige. For each relationship,

a. Compute the slope (b) and find the Y intercept (a).
b. State the least-squares regression line. What prestige score would you predict for a son whose father had a prestige score of 72? What prestige score would you predict for a daughter whose father had a prestige score of 72?
c. Compute r and r^2.
d. Describe the strength and direction of the relationships in a sentence or two. Does the occupational prestige of the father have an impact on his children? Does it have the same impact for daughters as it does for sons?

12.3 GER The residents of a housing development for senior citizens have completed a survey in which they indicated how physically active they are and how many visitors they receive each week. Are these two variables related for the ten cases reported here? Compute r and r^2. Find the least-squares regression line. What would be the predicted number of visitors for a person whose level of activity was a 5? How about a person who scored 18 on level of activity?

Case	Level of Activity	Number of Visitors
A	10	14
B	11	12
C	12	10
D	10	9
E	15	8
F	9	7
G	7	10
H	3	15
I	10	12
J	9	2

12.4 *PS* The following variables were collected for a sample of ten precincts during the last national election. Compute r and r^2 for each combination of variables and write a paragraph interpreting the relationship between these variables. Take voter turnout as the dependent variable.

Precinct	Percentage Democrat	Percentage Minority	Percentage Voter Turnout
A	33	9	36
B	78	15	60
C	62	18	64
D	50	10	56
E	45	12	55
F	56	8	52
G	85	20	25
H	25	0	42
I	13	5	89
J	33	9	88

12.5 *SOC* The basketball coach at a small local college believes that his team plays better and scores more points in front of larger crowds. The number of points scored and attendance for all home games last season are reported here. Do these data support the coach's argument?

Game	Points Scored	Attendance
1	54	378
2	57	350
3	59	320
4	80	478
5	82	451
6	75	250
7	73	489
8	53	451
9	67	410
10	78	215
11	67	113
12	56	250
13	85	450
14	101	489
15	99	472

Statistical Analysis Using SPSS

12.6 *SOC* Load the *Intl-Pop* data set and get scatterplots for the independent variable *GNIcap* (Gross National Income per capita, a measure of affluence) and each of these dependent variables: *BirthRate* (birth rate or number of births per 1000 population), *DthRate* (death rate or number of deaths per 1000 population), and *LifeExp* (life expectancy or the average number of years of life a newborn infant can expect to live).

- Click **Graphs**, **Legacy Dialog**, and **Scatter/Dot**.
- On the "Scatter/Dot" window, click **Simple Scatter** and then click **Define**.
- On the "Simple Scatterplot" window,
  - Find *BrthRate* in the list of variables and move the variable name to the "Y Axis:" box.
  - Find *GNIcap* and click the arrow to move the variable name to the "X Axis:" box.
  - Click **OK** and the scatterplot will be sent to the SPSS output window.
- To add the regression line to the graph,
  - Click anywhere on the graph and the "Chart Editor" window will appear.
  - Click the **Elements** command at the top of the window and, in the drop-down menu, click **Fit Line at Total**.
  - A new window will open. In the "Fit Method" panel, click **Linear** and then click **Close**.
  - Close the "Chart Editor" window to return to the Output window of SPSS.
- Repeat the **Graphs**, **Legacy Dialog**, and **Scatter/Dot** commands and, when you get to the "Simple Scatterplot" window, substitute *DthRate* for *BirthRate*. Repeat the instructions above to get a scatterplot for *DthRate* and *GNIcap*.
- Repeat the **Graphs**, **Legacy Dialog**, and **Scatter/Dot** commands once more and substitute *LifeExp* for *DthRate* in the "Simple Scatterplot" window. Repeat the instructions above to get a scatterplot for *LifeExp* and *GNIcap*.

a. Describe each of these graphs in terms of the direction and strength of the relationship they display. Use the regression line and the spread of the dots around the line to make these determinations.

b. Do any of these scatterplots violate the assumption of linearity? How?

12.7 *SOC* Load the *GSS2012* data set and use the **Correlate** command to get correlations between *age* (the independent or *X* variable) and four dependent (or *Y*) variables: church attendance (*attend*), number of children (*childs*), self-rated social position (*rank*), and hours of TV watching each day (*tvhours*).

- Click the SPSS icon on your desktop.
- Load the *GSS2012* database.
- From the main menu bar, click **Analyze**, **Correlate**, and **Bivariate**.

- In the "Bivariate Correlations" dialog box, find *age*, *attend*, *childs*, *tvhours*, and *rank* in the variable list on the left and move them to the "Variables" window on the right.
- Click **OK** and the output will be printed to the SPSS Output screen.

Focus on the relationships with *age* and record the values for *r* for each relationship below. Calculate r^2 and record those values as well.

	attend		*childs*		*prestige*		*tvhours*	
	r	r^2	r	r^2	r	r^2	r	r^2
age								

a. Write a summary of these relationships in which you note the strength and direction of each relationship with age. What percentage of the variance in each dependent variable was explained by age?

b. Which relationships were statistically significant at the 0.05 level? To find out, look at the second line in each cell of the correlation matrix ("Sig. 2-tailed"). Values below 0.05 are significant.

12.8 *SOC* Load the *States* dataset and use the **Correlate** command to get correlations between *SthDUMMY* (South/Non-South), *College* (percentage of the state's population with at least a college degree), *TrafDths11* (the rate of traffic fatalities in 2011), *Internet* (the percentage of households using the Internet), and *InfantMort* (the death rate for people less than one year of age 1).

- Click the SPSS icon on your desktop.
- Load the *States* database.
- From the main menu bar, click **Analyze**, **Correlate**, and **Bivariate**.
- In the "Bivariate Correlations" dialog box, find *SthDUMMY*, *College*, *TrafDths11*, *Internet*, and *InfantMort* in the variable list on the left and move them to the "Variables" window on the right.
- Click **OK** and the output will be printed to the SPSS Output screen.

Take region (*SthDUMMY*) and education (*College*) as independent variables and the other three variables as dependent. Record the values of *r* for each relationship below. Calculate r^2 and record those values as well.

	TrafDths11		*Internet*		*InfantMort*	
	R	r^2	r	r^2	r	r^2
SthDUMMY						
College						

a. Write a summary of these relationships in which you note the strength and direction of each relationship. What percentage of the variance in each dependent variable was explained by region? By education?

b. Do you think these relationships are causal? How? Do one of the following:
- Pick one of the six relationships in the table and explain why you think it is causal. Or,
- Pick one relationship and explain why you think it's a case of correlation but not causation.

12.9 *SOC/CJ* Use the *States* dataset and the **Correlate** command to explore some of the correlates of crime. Use two measures of crime in 2012 as dependent variables: the homicide rate (*Hom12*) and the rate of auto theft (*Carthft12*). Examine the relationships of these variables with three possible independent variables: population growth rate (*PopGrow*), population density (*PopDense*), and unemployment rate (*Unemplymnt*)

- Click the SPSS icon on your desktop.
- Load the *States* database.
- From the main menu bar, click **Analyze**, **Correlate**, and **Bivariate**.
- In the "Bivariate Correlations" dialog box, find *Hom12*, *Carthft12*, *PopGrow*, *PopDense*, and *Unemplymnt* and move them to the "Variables" window.
- Click **OK** and the output will be printed to the SPSS Output screen.

Record the values of *r* for each relationship. Calculate r^2 and record those values in the table as well.

	Hom12		*CarThft12*	
	r	r^2	r	r^2
PopGrow				
PopDense				
Unemplymnt				

a. Write a summary of these relationships in which you note the strength and direction of each relationship. What percentage of the variance in each measure of crime variable was explained by population growth rate? By population density? By unemployment?

b. Which of these relationships seems causal? Why?

12.10 *SOC/CJ* Follow up on problem 12.9 by finding the regression coefficients for the relationship between homicide rate and unemployment. Use the *States* data set and the **Regression** command.

- Load the *States* database.
- From the main menu bar, click **Analyze**, **Regression**, and **Linear**.
- On the "Linear Regression" screen, place *Hom12* in the "Dependent" window and *Unemplymnt* in the "Independent" window.
- Click **OK** and results will be printed to the Output screen.
- The "Coefficients" box displays the value of *a* and *b* under "Unstandardized Coefficients." Look in the column labeled B. The value in the first row is *a* and the entry in the second row is *b*.

a. State the regression equation with the values for *a* and *b*.

b. Use the equation to predict homicide rates for states with unemployment rates of 2% ($X = 2$), 5% ($X = 5$), 10% ($X = 10$), and 15% ($X = 15$). What happens to homicide rate as unemployment increases? Does this mean that unemployment *causes* homicide? Why or why not?

12.11 **SOC** Use the *Intl-Pop* dataset and the **Correlate** command to explore some of the correlates of cell phone use around the globe. Use *CellPhones* (the number of cell phone subscribers per 100 population), *GNIcap* (Gross National Income per capita—this is a measure of affluence), and *Urban* (percentage of population living in cities).

- Click the SPSS icon on your desktop.
- Load the *Intl-Pop* database.
- From the main menu bar, click **Analyze**, **Correlate**, and **Bivariate**.
- In the "Bivariate Correlations" dialog box, find *CellPhones*, *GNIcap*, and *Urban* and move them to the "Variables" window.
- Click **OK** and the output will be printed to the SPSS Output screen.

Record the values of *r* for each relationship. Calculate r^2 and record those values in the table as well.

	GNIcap		*Urban*	
	r	r^2	*r*	r^2
CellPhones				

Write a summary of these relationships in which you note the strength and direction of each relationship. What percentage of the variance in *CellPhones* was explained by *GNIcap*? By *Urban*?

12.12 **SOC** Follow up on problem 12.11 by finding the regression coefficients for the relationship between cell phone use and urbanization. Use the *Intl-Pop* dataset and the **Regression** command.

- Load the *Intl-Pop* database.
- From the main menu bar, click **Analyze**, **Regression**, and **Linear**.
- On the "Linear Regression" screen, place *CellPhones* in the "Dependent" window and *Urban* in the "Independent" window.
- Click **OK** and results will be printed to the Output screen.
- The "Coefficients" box displays the value of *a* and *b* under "Unstandardized Coefficients." Look in the column labeled B. The value in the first row is *a* and the entry in the second row is *b*.

a. State the regression equation with the values for *a* and *b*.

b. Use the equation to predict cell phone use for nations that are 25% urbanized ($X = 25$), 50% urbanized ($X = 50$), 75% urbanized ($X = 75$), and 99% urbanized ($X = 99$). What happens to cell phone use as urbanization increases?

12.13 **SOC/PS** What are the correlates of voter turnout? Use the *States* database with *Voters* (percentage of the population voting for president in 2012) as the dependent variable and *College* (percent of the population with a college degree) and *MdHHInc* (median household income—a measure of affluence) as independent variables

- Click the SPSS icon on your desktop.
- Load the *States* database.
- From the main menu bar, click **Analyze**, **Correlate**, and **Bivariate**.
- In the "Bivariate Correlations" dialog box, find *College*, *MdHHInc*, and *Voters* and move them to the "Variables" window.
- Click **OK** and the output will be printed to the SPSS Output screen.

Record the values of *r* for each relationship. Calculate r^2 and record those values in the table as well.

	Voters	
	r	r^2
College		
MdHHInc		

a. Write a summary of these relationships in which you note the strength and direction of each relationship. What percentage of the variance in voter turnout was explained by the education variable? By the affluence variable?

b. Which of these relationships seems causal? Why?

YOU ARE THE RESEARCHER

Who Watches TV? Who Succeeds in Life?

Two projects are presented to help you apply the statistical skills developed in this chapter. In the first, you will analyze the correlates of time spent watching TV. You will select four independent variables and assess their impact on *tvhours* (number of hours per day spent watching TV). In the second project, you will choose either *income06 or rank* as your dependent variable. Both variables measure social class standing and you will select independent variables you believe might be associated with an individual's level of success.

Project 1: Who Watches TV?

In this exercise, your dependent variable will be *tvhours*. Scores on this variable range from zero to a high of 24 hours (all day!), with most cases clustered around 1 to 4 hours of TV watching per day.

Step 1: Choosing Independent Variables

Select four variables from the 2012 GSS that you think might be important causes of *tvhours*. Your independent variables cannot be nominal in level of measurement unless you recode the variable into a dummy variable (see below). You may use any interval-ratio or ordinal-level variable.

Dummy variables. To include a dummy variable in your analysis, recode the variable so that it has only two values: 0 and 1. Some possibilities include recoding sex so that males = 0 and females = 1, *racecen1* so that whites = 0 and nonwhites = 1, or *relig* so that Protestants = 0 and non-Protestants = 1. See Chapter 9 for instructions on recoding.

Once you have selected your variables, list them in the table below and describe exactly what they measure.

SPSS Variable Name	What Exactly Does This Variable Measure?

Step 2: Stating Hypotheses

State hypotheses about the relationships you expect to find between your independent variables and *tvhours*. State these hypotheses in terms of the strength and direction of the relationships you expect to find. For example, you might hypothesize that time spent watching TV will increase strongly as age increases.

SPSS Variable Name	Hypothesis

Step 3: Running Bivariate Correlations

Click **Analyze → Bivariate → Correlate** and place all variables in the "Variables:" box. Click **OK** to get your results.

Step 4: Recording Results

Use the table below to summarize your results. Enter the ***r*** for each independent variable in each cell. Calculate r^2 and record those values as well. As you read the correlation matrix, ignore correlations of variables with themselves and any redundant information.

	Independent variables							
	1.______		2.______		3.______		4.______	
	r	r^2	r	r^2	r	r^2	r	r^2
tvhours								

Step 5: Analyzing and Interpreting Results

Write a short summary of results for each independent variable. Your summary needs to identify the variables being tested and the strength and direction of the relationship. It is probably best to characterize the relationship in general terms and then cite the statistical values in parentheses. Be sure to note whether or not your hypotheses were supported. *Be careful when interpreting direction* and refer back to the coding scheme to make sure you understand the relationship.

Project 2: Who Succeeds?

For this exercise, choose either income (*income06*) or the respondents' ranking of their social position (*rank*) as your dependent variable. Most Americans would regard these variables as measures of success in life. What are the correlates and antecedents of affluence and prestige?

Step 1: Choosing Independent Variables

Select four variables that you think might be important causes of your dependent variable. An obvious choice is education (use *educ*, the interval-ratio version, not *degree*, the ordinal version). Remember that independent variables *cannot* be nominal in level of measurement unless you recode the variable into a dummy variable (see below). You may use any interval-ratio or ordinal-level variable.

Dummy variables. To include a dummy variable in your analysis, recode the variable so that it has only two values: 0 and 1. Some possibilities include recoding *sex* so that males = 0 and females = 1, *racecen1* so that whites = 0 and nonwhites = 1, or *relig* so that Protestants = 0 and non-Protestants = 1. See Chapter 9 for instructions on recoding.

Once you have selected your variables, list them in the table below and describe exactly what they measure.

SPSS Variable Name	What Exactly Does This Variable Measure?

Step 2: Stating Hypotheses

State hypotheses about the relationships you expect to find between your independent variables and the dependent variable you selected. State these hypotheses in terms of the direction of the relationship you expect to find. For example, you might hypothesize that income will increase as age increases.

SPSS Variable Name	Hypothesis

Step 3: Running Bivariate Correlations

Click **Analyze** ⟶ **Correlate** ⟶ **Bivariate** and place all variables in the "Variables:" box. Click **OK** to get your results.

Step 4: Recording Results

Use the table below to summarize your results. Enter the r for each independent variable in each cell. Calculate r^2 and record those values as well. As you read the correlation matrix, ignore correlations of variables with themselves and any redundant information.

	Independent variables							
income06 or rank	1.__________		2.__________		3.__________		4.__________	
	r	r^2	r	r^2	r	r^2	r	r^2

Step 5: Analyzing and Interpreting Results

Write a short summary of results for each independent variable. Your summary needs to identify the variables being tested and the strength and direction of the relationship. It is probably best to characterize the relationship in general terms and then cite the statistical values in parentheses. Be sure to note whether or not your hypotheses were supported. *Be careful when interpreting direction* and refer back to the coding scheme to make sure you understand the relationship.

Part IV Multivariate Techniques

Chapter 13 introduces multivariate analytical techniques or statistics that allow us to analyze the relationships between more than two variables at a time. These statistics are extremely useful for probing causal relationships between variables and are commonly reported in the professional research literature. In particular, this chapter introduces regression analysis, which is the basis for many of the most popular and powerful statistical techniques in use today.

These techniques are designed to be used with variables measured at the interval-ratio level of measurement and the mathematics underlying these techniques can become very complicated, so the chapter focuses on the simplest possible applications, relies on SPSS for computation, and stresses interpretation.

13 Partial Correlation and Multiple Regression and Correlation

LEARNING OBJECTIVES

By the end of this chapter, you will be able to:

1. Compute and interpret partial correlation coefficients.
2. Find and interpret the least-squares multiple regression equation with partial slopes.
3. Find and interpret standardized partial slopes or beta-weights (b^*).
4. Calculate and interpret the coefficient of multiple determination (R^2).
5. Explain the limitations of partial and multiple regression analysis.
6. Use SPSS to generate partial correlations and conduct multiple regression analysis.

USING STATISTICS

The statistical techniques presented in this chapter are used to analyze the association between more than two interval-ratio-level variables. Examples of research situations in which these techniques are useful include:

1. A family sociologist studies dual-wage-earner families and wonders whether husbands increase their housework as the number of children increases. Is the relationship the same regardless of the husband's education?
2. A criminologist has data that show a moderate and positive correlation between poverty and crime rate for a sample of 542 U.S. counties and wonders whether the relationship is the same in densely populated urban counties and less populated rural counties? Is it the same in counties with high and low levels of education?
3. A demographer is investigating the relationship between fertility rates and the educational levels of females in 76 nations. Her data show a moderate, negative relationship between the variables: As education increases, fertility rates tend to decrease. Is the relationship the same in Christian and non-Christian nations? Does the relationship retain its strength for nations at all levels of affluence?

As I mentioned at the beginning of Chapter 12, social science research is multivariate by nature and involves the simultaneous analysis of scores of variables. Some of the most powerful and widely used statistical tools for multivariate analysis are introduced in this chapter. We cover techniques that are used to analyze causal relationships between variables and to make predictions, both crucial endeavors in any science.

These techniques are based on Pearson's *r* (see Chapter 12) and are most appropriately used with high-quality, precisely measured interval-ratio variables. As we have noted on many occasions, such data are relatively rare in social science research, and the techniques presented in this chapter are commonly used on variables measured at the ordinal level and with nominal-level variables in the form of dummy variables (see Chapter 12).

We first consider partial correlation analysis, a technique that allows us to examine bivariate relationships while controlling for a third variable. This is an important technique for many reasons but, perhaps most importantly, because it allows us to deal with the possibility that a third variable is affecting a bivariate relationship. As we pointed out in Chapter 12, this is one of the tests that must be passed in order to argue for a causal relationship.

Second, we consider multiple regression and correlation, techniques that allow researchers to assess the effects, separately and in combination, of more than one independent variable on the dependent variable. These techniques also provide important statistical tools for examining the effect of third variables on our bivariate relationships.

Throughout this chapter, we focus on research situations involving three variables and very small samples to illustrate the mathematical routines and basic logic. Extensions to situations involving four or more variables are relatively straightforward, but the computations become very complex. Thus, after working through a simplified example problem, we will use SPSS to execute the computations, as we did with Pearson's *r* (Chapter 12). Likewise, there are a few end-of-chapter problems with small data sets so that you can be sure you understand the mathematical routines, but most problems require the use of SPSS.

Partial Correlation

In Chapter 12, we used Pearson's *r* to measure the strength and direction of bivariate relationships. To provide an example, we looked at the relationship between husband's contribution to housework (the dependent, or *Y*, variable) and number of children (the independent, or *X*, variable) for a sample of 12 families. We found a positive relationship of moderate strength ($r = 0.50$) and concluded that husbands increase their contribution to housework as the number of children increases.

You might wonder, as researchers commonly do, whether this relationship holds true for *all* types of families. For example, might husbands from different religious orientations respond differently? Would politically conservative husbands behave differently than husbands who are liberal? Would more educated husbands respond differently than less educated husbands? We can address issues like these by means of a technique called **partial correlation**, in which we observe how the bivariate relationship changes when a third variable, such as religion, political ideology, or education, is introduced. Third variables are often referred to as ***Z*** variables or **control variables**.

Partial correlation proceeds by first computing Pearson's r for the bivariate relationship (sometimes called the **zero-order correlation**) and then computing the partial (or first-order) correlation coefficient. If the partial correlation coefficient differs from the zero-order correlation coefficient, we conclude that the third variable has an effect on the bivariate relationship. For example, if well-educated husbands respond differently to an additional child than less-educated husbands, the partial correlation coefficient will differ in strength (and perhaps in direction) from the bivariate correlation coefficient.

Before considering matters of computation, we'll consider the possible relationships between the partial and bivariate correlation coefficients and what they might mean.

Types of Relationships

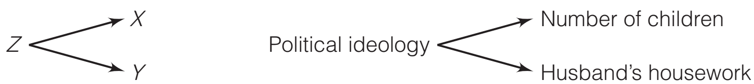
Direct Relationship. One possibility is that the partial correlation coefficient is essentially the same value as the bivariate coefficient. Imagine, for example, that we found a partial correlation coefficient of +0.49 after controlling for husband's education, compared to the zero-order Pearson's r of +0.50. This would mean that the third variable (husband's education) has no effect on the relationship between number of children and husband's hours of housework. In other words, husbands respond in a similar way to additional children regardless of their education. This outcome is consistent with the conclusion that there is a **direct** or **causal relationship** (see Figure 13.1) between X and Y and that the third variable (Z) is irrelevant and should be discarded from further consideration. The next step in the research project might be to run additional tests with other likely control variables (e.g., the researcher might control for the religion or ethnicity of the family).

FIGURE 13.1 A Direct Relationship Between *X* and *Y*

Spurious and Intervening Relationships. A second possible outcome occurs when the partial correlation coefficient is much weaker than the bivariate correlation, perhaps even dropping to 0. This outcome is consistent with two different relationships. The first is called a **spurious relationship**: The control variable (Z) is a cause of both the independent (X) and the dependent (Y) variable (see Figure 13.2). This outcome would mean that X and Y are not actually related. They appear to be related only because both are dependent on a common cause (Z). Once Z is taken into account, the apparent relationship between X and Y disappears.

What would a spurious relationship look like? Imagine that we controlled for the political ideology of parents and found that the partial correlation coefficient was much weaker than the bivariate Pearson's r. This might indicate that the number of children does not actually change the husband's contribution to housework (that is, the relationship between X and Y is not direct). Rather, political ideology might be the mutual cause of both of the other variables: Perhaps more

FIGURE 13.2 A Spurious Relationship Between *X* and *Y*



STATISTICS
IN EVERYDAY LIFE

Spuriousness

Spurious relationships are part of everyday life, not just scientific research. We sometimes leap to incorrect conclusions about "what causes what" as we consider the news or chat with friends, and being aware of spurious relationships can enhance your appreciation of the complexity of social life and the subtlety of causal relationships. As you think about spuriousness, it may be helpful to refer to some simplified examples. Here are three statements, each of which asserts that one variable (*X*) causes another (*Y*). Each relationship is spurious. Can you identify the common cause of both *X* and *Y* for each statement? The answers appear below.

1 People in hospitals are more likely to die. Therefore, hospitals (*X*) cause death (*Y*).
2 The greater the number of firefighters at a fire, the greater the damage. Therefore, firefighters (*X*) cause fire damage (*Y*).
3 Men who listen to a lot of folk music from the 1960s are more likely to suffer from sexual impotence. Therefore, folk music (*X*) causes impotence (*Y*).

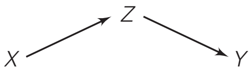
Answers:

1 *Z* = illness and serious injury. The sick and injured are more likely to be in a hospital and more likely to die
2 *Z* = extent of the blaze. More serious fires attract more firefighters and result in more damage
3 *Z* = age. Older men are more likely to listen to this genre of music and more likely to be impotent.

conservative families are more likely to follow traditional gender role patterns (in which husbands contribute less to housework) *and* have more children.

This pattern (partial correlation much weaker than the bivariate correlation) is also consistent with an **intervening** relationship between the variables (see Figure 13.3). In this situation, *X* and *Y* are not linked directly but are connected through the control variable. Again, once *Z* is controlled, the apparent relationship between *X* and *Y* disappears.

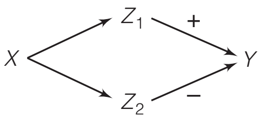
FIGURE 13.3 An Intervening Relationship Between *X* and *Y*



How can we tell the difference between spurious and intervening relationships? This distinction cannot be made on statistical grounds: Spurious and intervening relationships look exactly the same in terms of statistics. The researcher may be able to distinguish between these two relationships in terms of the time order of the variables (i.e., which came first) or on theoretical grounds, but not on statistical grounds.

Interaction. A third possible relationship between variables should be mentioned, even though it *cannot* be detected by partial correlation analysis. This relationship, called interaction, occurs when the relationship between *X* and *Y* changes markedly under the various values of *Z*. For example, if we controlled for social class and found that husbands in middle-class families increased their contribution to housework as the number of children increased while husbands in working-class families did just the reverse, we would conclude that there was interaction among these three variables. In other words, there would be a positive relationship between *X* and *Y* for one category of *Z* and a negative relationship for the other category, as illustrated in Figure 13.4.

FIGURE 13.4 An Interactive Relationship Between *X*, *Y*, and *Z*



STATISTICS
IN EVERYDAY LIFE

Interaction
It may be helpful to think about interaction as an unexpectedly intense reaction when variables are combined. For example, suppose a study of the causes of street crime found moderately strong associations with poverty ("areas with more poverty have higher levels of street crime") and age ("areas with high percentages of teenagers and young adults have higher levels of street crime") but very strong relationships when the variables are combined ("areas with *both* high poverty *and* younger populations have *much* higher levels of crime"). This may be an example of interaction. Each independent variable has its own relationship with crime, but their combined effects have a surprisingly strong impact on the dependent variable.

Computing and Interpreting the Partial Correlation Coefficient

The logic and computation of the partial correlation coefficient requires some new concepts and terminology. We'll introduce the terminology first and then move on to the formula.

Terminology. Partial correlation requires that we deal with more than one bivariate relationship, and we need to differentiate between them with subscripts. Thus, the symbol r_{yx} will refer to the correlation coefficient between variable Y and variable X; r_{yz} will refer to the correlation coefficient between Y and Z; and r_{xz} will refer to the correlation coefficient between X and Z. Recall that correlation coefficients calculated for bivariate relationships are often referred to as zero-order correlations.

Partial correlation coefficients, or first-order partials, are symbolized as $r_{yx.z}$. The variable to the right of the dot in the subscript is the control variable. Thus, $r_{yx.z}$ refers to the partial correlation coefficient that measures the relationship between variables X and Y while controlling for variable Z. The formula for the first-order partial is

FORMULA 13.1

$$r_{yx.z} = \frac{r_{yx} - (r_{yz})(r_{xz})}{\sqrt{1 - r_{yz}^2}\sqrt{1 - r_{xz}^2}}$$

Note that you must first calculate the zero-order coefficients between all possible pairs of variables (variables X and Y, X and Z, and Y and X) before solving this formula.

Computation. To illustrate the computation of a first-order partial, we will return to the relationship between number of children (X) and husband's contribution to housework (Y) for 12 dual-career families. The zero-order r between these two variables ($r_{yx} = 0.50$) indicated a moderate, positive relationship (as number of children increased, husbands tended to contribute more to housework). Suppose the researcher wished to investigate the possible effects of husband's education on the bivariate relationship?. The original data (from Table 12.1) and the scores of the 12 families on the new variable are presented in Table 13.1.

TABLE 13.1 Scores on Three Variables for 12 Dual-Wage-Earner Families

Family	Husband's Housework (*Y*)	Number of Children (*X*)	Husband's Years of Education (*Z*)
A	1	1	12
B	2	1	14
C	3	1	16
D	5	1	16
E	3	2	18
F	1	2	16
G	5	3	12
H	0	3	12
I	6	4	10
J	3	4	12
K	7	5	10
L	4	5	16

TABLE 13.2 Zero-Order Correlations

	Husband's Housework (*Y*)	Number of Children (*X*)	Husband's Years of Education (*Z*)
Husband's Housework (*Y*)	1.00	0.50	−0.30
Number of Children (*X*)		1.00	−0.47
Husband's Years of Education (*Z*)			1.00

The zero-order correlations, as presented in the correlation matrix in Table 13.2, indicate that the husband's contribution to housework is positively related to number of children ($r_{yx} = 0.50$), that better-educated husbands tend to do less housework ($r_{yz} = -0.30$), and that families with better-educated husbands have fewer children ($r_{xz} = -0.47$).

Is the relationship between husband's housework (*Y*) and number of children (*X*) affected by husband's years of education (*Z*)? Substituting the zero-order correlations into Formula 13.1, we would have

$$r_{yx.z} = \frac{r_{yx} - (r_{yz})(r_{xz})}{\sqrt{1 - r_{yz}^2}\sqrt{1 - r_{xz}^2}}$$

$$r_{yx.z} = \frac{(0.50) - (-0.30)(-0.47)}{\sqrt{1 - (-0.30)^2}\sqrt{1 - (-0.47)^2}}$$

$$r_{yx.z} = \frac{(0.50) - (0.14)}{\sqrt{1 - 0.09}\sqrt{1 - 0.22}}$$

$$r_{yx.z} = \frac{0.36}{\sqrt{0.91}\sqrt{0.78}}$$

$$r_{yx.z} = \frac{0.36}{(0.95)(0.88)}$$

$$r_{yx.z} = \frac{0.36}{0.84}$$

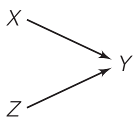
$$r_{yx.z} = 0.43$$

Interpretation. The first-order partial ($r_{yx.z} = 0.43$) measures the strength of the relationship between husband's housework (Y) and number of children (X) while controlling for husband's education (Z). It is lower in value than the zero-order coefficient ($r_{yx} = 0.50$), but the difference in the two values is not great. This result suggests a direct relationship between variables X and Y. That is, when controlling for husband's education, the statistical relationship between husband's housework and number of children is essentially unchanged. Regardless of education, husband's hours of housework increase with the number of children.

Our next step in statistical analysis would probably be to select another control variable. The more the bivariate relationship retains its strength across a series of controls for third variables (Z's), the stronger the evidence for a direct relationship between X and Y.

In closing, I should mention an additional possible outcome, in which the partial correlation coefficient is greater in value than the zero-order coefficient. This outcome would be consistent with a causal model in which the variable taken as independent and the control variable each had a separate effect on the dependent variable and were uncorrelated with each other. This relationship is depicted in Figure 13.5. The absence of an arrow between X and Z indicates that they have no mutual relationship.

FIGURE 13.5 A Possible Causal relationship Among Three Variables



This pattern means that both X and Z should be treated as independent variables, and the next step in the statistical analysis would probably involve multiple correlation and regression, techniques that will be presented in the remainder of this chapter. *(For practice in computing and interpreting partial correlation coefficients, see problems 13.1a and b, 13.4a, and 13.5.)*

ONE STEP AT A TIME — Computing the Partial Correlation Coefficient

Compute Pearson's r for all pairs of variables. Be clear about which variable is independent (X), which is dependent (Y), and which is the control (Z).

Step	***Operation***
	To solve Formula 13.1:
1.	Multiply r_{yz} by r_{xz}.
2.	Subtract the value you found in step 1 from r_{yx}. *This value is the numerator of Formula 13.1.*
3.	Square the value of r_{yz}.
4.	Subtract the quantity you found in step 3 from 1.
5.	Take the square root of the quantity you found in step 4.
6.	Square the value of r_{xz}.
7.	Subtract the quantity you found in step 6 from 1.
8.	Take the square root of the quantity you found in step 7.
9.	Multiply the quantity you found in step 8 by the value you found in step 5. *This value is the denominator of Formula 13.1.*
10.	Divide the quantity you found in step 2 by the quantity you found in step 9. *This is the partial correlation coefficient.*

ONE STEP AT A TIME Interpreting the Partial Correlation Coefficient

Choose the description below that comes closest to matching the relationship between the zero-order correlation (r_{yx}) and the partial correlation ($r_{yx.z}$).

Step	*Operation*
1.	The partial correlation coefficient ($r_{yx.z}$) is roughly the same value as the zero-order correlation (r_{yx}). A good rule of thumb for "roughly the same" is a difference of less than 0.10. This outcome is evidence that that the control variable (Z) has no effect and that the relationship between X and Y is direct.
2.	The partial correlation coefficient is much less (say, more than 0.10 less) than the bivariate correlation. This is evidence that the control variable (Z) changes the relationship between X and Y. The relationship between X and Y is either spurious (Z causes both X and Y) or intervening (X and Y are linked by Z).
3.	Be aware that that X, Y, and Z may have an interactive relationship in which the relationship between X and Y changes for each category of Z. Partial correlation analysis cannot detect interactive relationships.

STATISTICS IN EVERYDAY LIFE

Education and Income

An analysis of a representative sample of Americans age 25 to 65 found a correlation of 0.40 between years of schooling and income, a moderate, positive relationship. Income increases as education increases and education explains about 16% of the variation in income.

Is this relationship affected by age? The partial correlation between education and income, controlling for age, is 0.41, essentially the same as the zero-order correlation. The relationship between education and income is direct: regardless of age, income increases as education increases. You may find this analysis reassuring as you move to the completion of your degree.

Source: Full 2012 General Social Survey. Variables: *educ, rincome06*, and *age*. $N = 975$

Applying Statistics 13.1 Using SPSS to Determine If the Relationship Between Fertility and Women's Education Is Affected by Affluence

In Applying Statistics 12.2, we found a strong negative relationship ($r = -0.78$) between women's education (X) and fertility rates (Y) for the nations in the *Intl-POP* database. Now, we ask if the bivariate relationship is spurious and caused by the affluence (Z) of the nations. Do these variables appear to be related only because women in more affluent nations have more education *and* fewer children?

Start SPSS, load the *Intl-POP* database, and click **Analyze** and **Correlate** but choose **Partial** (not **Bivariate**) from the sub-menu. Enter *TFR* (Total Fertility Rate) and *EducFemales* (the measure of the level of schooling for females we used previously) in the "Variables:" window. In the bottom "Controlling for:" window, enter *GNIcap* (Gross National Income per capita). This variable measures national affluence: The higher the score on this variable, the richer the nation. Click **OK**.

The output box will display a correlation matrix with the control variable (*GNIcap*) listed on the left. Find the partial correlation between education for females and Total Fertility Rate in the matrix. The value of -0.55 is much weaker (much more than our rule of thumb of 0.10)

(*continued*)

Applying Statistics 13.1 (*continued*)

than the bivariate correlation (−0.78) but still in the "moderate to strong" range.

This ambiguous result suggests that the relationship between education and fertility is greatly affected by the affluence of the nation and at least partly spurious. However, regardless of the affluence of the nation, more educated women do tend to have fewer children. Clearly, further analysis is required to clarify this relationship, perhaps using other measures of fertility, education, and affluence.

Multiple Regression: Predicting the Dependent Variable

In Chapter 12, the least-squares regression line was introduced as a way of describing the overall linear relationship between two interval-ratio variables and of predicting scores on Y from scores on X. This line was the best-fitting line to summarize the bivariate relationship and was defined by this formula:

FORMULA 13.2

$$Y = a + bX$$

The least-squares regression line can be modified to include (theoretically) any number of independent variables. This technique is called **multiple regression**. For ease of explication, we will confine our attention to the case involving two independent variables. The least-squares multiple regression equation is

FORMULA 13.3

$$Y = a + b_1X_1 + b_2X_2$$

where b_1 = the partial slope of the linear relationship between the first independent variable (X_1) and Y
b_2 = the partial slope of the linear relationship between the second independent variable (X_2) and Y

Some new notation and some new concepts are introduced in this formula. First, while the dependent variable is still symbolized as Y, the independent variables are differentiated by subscripts. Thus, X_1 identifies the first independent variable, and X_2, the second. The symbol for the slope (b) is also subscripted to identify the independent variable with which it is associated.

Partial Slopes

A major difference between the multiple and bivariate regression equations concerns the slopes (b's). In the case of multiple regression, the b's are called **partial slopes**, and they show the amount of change in Y for a unit change in the independent variable while controlling for the effects of the other independent variables in the equation. The partial slopes are thus analogous to partial correlation coefficients and represent the direct effect of the associated independent variable on Y.

Computing Partial Slopes. The partial slopes for the independent variables are determined by Formulas 13.4 and 13.5:[1]

[1]Partial slopes can be computed from zero-order slopes, but Formulas 13.4 and 13.5 are easier to use.

FORMULA 13.4

$$b_1 = \left(\frac{s_y}{s_1}\right)\left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)$$

FORMULA 13.5

$$b_2 = \left(\frac{s_y}{s_2}\right)\left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right)$$

where b_1 = the partial slope of X_1 on Y
b_2 = the partial slope of X_2 on Y
s_y = the standard deviation of Y
s_1 = the standard deviation of the first independent variable (X_1)
s_2 = the standard deviation of the second independent variable (X_2)
r_{y1} = the bivariate correlation between Y and X_1
r_{y2} = the bivariate correlation between Y and X_2
r_{12} = the bivariate correlation between X_1 and X_2

To illustrate the computation of the partial slopes, we will assess the combined effects of number of children (X_1) and husband's education (X_2) on husband's contribution to housework. All the relevant information can be calculated from Table 13.1 and is noted here:

Husband's Housework	Number of Children	Husband's Education
$\bar{Y} = 3.3$	$\bar{X}_1 = 2.7$	$\bar{X}_2 = 13.7$
$s_y = 2.1$	$s_1 = 1.5$	$s_2 = 2.6$

Zero-order Correlations
$r_{y1} = 0.50$
$r_{y2} = -0.30$
$r_{12} = -0.47$

The partial slope for the first independent variable (X_1) is

$$b_1 = \left(\frac{s_y}{s_1}\right)\left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)$$
$$b_1 = \left(\frac{2.1}{1.5}\right)\left(\frac{0.50 - (-0.30)(-0.47)}{1 - (-0.47)^2}\right)$$
$$b_1 = (1.4)\left(\frac{(0.50) - (0.14)}{1 - 0.22}\right)$$
$$b_1 = (1.4)\left(\frac{0.36}{0.78}\right)$$
$$b_1 = (1.4)(0.46)$$
$$b_1 = 0.64$$

For the second independent variable (X_2), the partial slope is

$$b_2 = \left(\frac{s_y}{s_2}\right)\left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right)$$

$$b_2 = \left(\frac{2.1}{2.6}\right)\left(\frac{-0.30 - (+0.50)(-0.47)}{1 - (-0.47)^2}\right)$$

$$b_2 = (0.81)\left(\frac{-0.30 - (-0.24)}{1 - 0.22}\right)$$

$$b_2 = (0.81)\left(\frac{-0.06}{0.78}\right)$$

$$b_2 = (0.81)(-0.08)$$

$$b_2 = -0.07$$

ONE STEP AT A TIME Computing Partial Slopes

These procedures apply when there are two independent variables and one dependent variable. For more complex situations, use a computerized statistical package such as SPSS to do the calculations.

Step ***Operation***

Compute the partial slope associated with the first independent variable by using Formula 13.4:

1. Divide s_y by s_1.
2. Multiply r_{y2} by r_{12}.
3. Subtract the value you computed in step 2 from r_{y1}.
4. Square the value of r_{12}.
5. Subtract the value you computed in step 4 (r_{12}^2) from 1.
6. Divide the value you computed in step 3 by the value you computed in step 5.
7. Multiply the value you computed in step 6 by the value you computed in step 1. *This value is the partial slope associated with the first independent variable.*

Compute the partial slope associated with the second independent variable by using Formula 13.5:

1. Divide s_y by s_2.
2. Multiply r_{y1} by r_{12}.
3. Subtract the value you computed in step 2 from r_{y2}.
4. Square the value of r_{12}.
5. Subtract the value you computed in step 4 from 1.
6. Divide the value you computed in step 3 by the value you computed in step 5.
7. Multiply the value you computed in step 6 by the value you computed in step 1. *This value is the partial slope associated with the second independent variable.*

Interpreting partial slopes:

The value of a partial slope is the change in the value of Y for a unit increase in the value of the associated independent variable while controlling for the effects of the other independent variable.

ONE STEP AT A TIME **Computing the *Y* Intercept**

Step	***Operation***
Find the Y *intercept using Formula 13.6:*	
1.	Multiply the mean of X_2 by b_2.
2.	Multiply the mean of X_1 by b_1.
3.	Subtract the quantity you found in step 1 from the quantity you found in step 2.
4.	Subtract the quantity you found in step 3 from the mean of *Y*. *The result is the value of* a*, the* Y *intercept.*

Finding the *Y* Intercept. Now that partial slopes have been determined for both independent variables, the *Y* intercept (*a*) can be found. Note that *a* is calculated from the mean of the dependent variable (symbolized as $\bar{Y}$) and the means of the two independent variables ($\bar{X}_1$ and $\bar{X}_2$).

FORMULA 13.6

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

Substituting the proper values for the example problem at hand, we would have

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$
$$a = 3.3 - (0.64)(2.7) - (-0.07)(13.7)$$
$$a = 3.3 - 1.7 - (-1.0)$$
$$a = 2.6$$

Predicting *Y*´ with the Least-Squares Multiple Regression Line. For our example problem, the full least-squares multiple regression equation is

$$Y = a + b_1X_1 + b_2X_2$$

$$Y = 2.6 + (0.64\ )X_1 + (-0.07)X_2$$

As was the case with the bivariate regression line, this formula can be used to predict scores on the dependent variable from scores on the independent variables. For example, what would be our best prediction of husband's housework (Y') for a family of four children ($X_1 = 4$) where the husband had completed 11 years of schooling ($X_2 = 11$)? Substituting these values into the least-squares equation, we would have

$$Y' = 2.6 + (0.64\ )(4) + (-0.07)(11)$$
$$Y' = 2.6 + 2.6 - 0.8$$
$$Y' = 4.4$$

Our prediction would be that this husband would contribute 4.4 hours per week to housework. This prediction is, of course, a kind of "educated guess," which is unlikely to be perfectly accurate. However, we will make fewer errors of prediction using the least-squares line (and, thus, incorporating information from the independent variables) than we would using any other method of prediction (assuming, of course, that there is a linear association between the independent

ONE STEP AT A TIME **Using the Multiple Regression Line to Predict Scores on *Y***

Step	*Operation*
1.	Choose a value for X_1. Multiply this value by the value of b_1.
2.	Choose a value for X_2. Multiply this value by the value of b_2.
3.	Add the values you found in steps 1 and 2 to the value of *a*, the *Y* intercept. *The resulting value is the predicted score on* Y.

and the dependent variables). *(All end-of-chapter problems offer opportunities to find unstandardized slopes and the* Y *intercept and to predict* Y *scores. The first three problems stress hand calculations and all others require the use of SPSS.)*

Multiple Regression: Assessing the Effects of the Independent Variables

The least-squares multiple regression equation (Formula 13.3) is used to isolate the separate effects of the independent variables and to predict scores on the dependent variable. However, in many situations, using this formula to determine the relative importance of the various independent variables will be awkward—especially when the independent variables differ in terms of units of measurement (e.g., number of children vs. years of education). When the units of measurement differ, a comparison of the partial slopes will not necessarily tell us which independent variable has the strongest effect and is thus the most important. Comparing the partial slopes of variables that differ in units of measurement is a little like comparing apples and oranges.

We can make it easier to compare the effects of the independent variables by converting all variables in the equation to a common scale and thereby eliminate variations in the values of the partial slopes that are solely a function of differences in units of measurement. We can, for example, standardize all distributions by changing the scores of all variables to *Z* scores. Each distribution of scores would then have a mean of 0 and a standard deviation of 1 (see Chapter 5), and comparisons between the independent variables would be much more meaningful.

Computing the Standardized Regression Coefficients

To standardize the variables to the normal curve, we could actually convert all scores into the equivalent *Z* scores and then recompute the slopes and the *Y* intercept. This would require a good deal of work; fortunately, a shortcut is available for computing the slopes of the standardized scores directly. These **standardized partial slopes**, called **beta-weights**, are symbolized b^*.

Beta-Weights. The beta-weights show the amount of change in the standardized scores of Y for a one-unit change in the standardized scores of each independent variable while controlling for the effects of all other independent variables.

Formulas and Computation for Beta-Weights. When we have two independent variables, the beta-weight for each is found by Formulas 13.7 and 13.8:

FORMULA 13.7

$$b_1^* = b_1\left(\frac{s_1}{s_y}\right)$$

FORMULA 13.8

$$b_2^* = b_2\left(\frac{s_2}{s_y}\right)$$

We can now compute the beta-weights for our sample problem to see which of the two independent variables has the stronger effect on the dependent variable. For the first independent variable, number of children (X_1):

$$b_1^* = b_1\left(\frac{s_1}{s_y}\right)$$
$$b_1^* = (0.64)\left(\frac{1.5}{2.1}\right)$$
$$b_1^* = (0.64)(0.71)$$
$$b_1^* = 0.45$$

For the second independent variable, husband's education (X_2):

$$b_2^* = b_2\left(\frac{s_2}{s_y}\right)$$
$$b_2^* = (-0.07)\left(\frac{2.6}{2.1}\right)$$
$$b_2^* = (-0.07)(1.24)$$
$$b_2^* = -0.09$$

Comparing the values of the beta-weights, we see that number of children (X_1) has a stronger effect than husband's education (X_2) on husband's housework (Y). Furthermore, the net effect (after controlling for the effect of husband's education) of the first independent variable is positive, while the net effect of the second independent variable is negative.

The Standardized Least-Squares Regression Line. Using standardized scores, the least-squares regression equation can be written as

FORMULA 13.9

$$Z_y = a_z + b_1^* Z_1 + b_2^* Z_2$$

Where: Z indicates that all scores have been standardized to the normal curve

ONE STEP AT A TIME **Computing Beta-Weights (b^*)**

These procedures apply when there are two independent variables and one dependent variable. For more complex situations, use a computerized statistical package such as SPSS to do the calculations.

Step	***Operation***
To compute the beta-weight associated with the first independent variable using Formula 13.7:	
1.	Divide s_1 by s_y.
2.	Multiply the value you found in step 1 by the partial slope of the first independent variable (b_1). *This value is the beta-weight associated with the first independent variable.*
To compute the beta-weight associated with the second independent variable using Formula 13.8:	
1.	Divide s_2 by s_y.
2.	Multiply the value you found in step a by the partial slope of the second independent variable (b_2). *This value is the beta-weight associated with the second independent variable.*

ONE STEP AT A TIME **Interpreting Beta-Weights (b^*)**

A beta-weight (or standardized partial slope) shows the increase in the value of *Y* for a unit increase in the value of the associated independent variable while controlling for the effects of the other independent variable, after all variables have been standardized (or transformed to *Z* scores).

The standardized regression equation can be further simplified by dropping the term for the *Y* intercept, because this term will always be 0 when scores have been standardized. This value is the point where the regression line crosses the *Y* axis and is equal to the mean of *Y* when all independent variables equal 0. This relationship can be seen by substituting 0 for all independent variables in Formula 13.6:

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$
$$a = \bar{Y} - b_1(0) - b_2(0)$$
$$a = \bar{Y}$$

Because the mean of any standardized distribution of scores is 0, the mean of the standardized *Y* scores will be 0 and the *Y* intercept will also be 0 ($a = \bar{Y} = 0$). Thus, Formula 13.9 simplifies to

FORMULA 13.10

$$Z_y = b_1^*Z_1 + b_2^*Z_2$$

The standardized regression equation for our sample problem, with beta-weights noted, would be

$$Z_y = (0.45)Z_1 + (-0.09)Z_2$$

and it is immediately obvious that the first independent variable has a much stronger direct effect on *Y* than does the second independent variable.

As we have seen, multiple regression analysis permits the researcher to summarize the linear relationship among two or more independent variables and a dependent variable. The unstandardized regression equation (Formula 13.2) permits values of Y to be predicted from the independent variables in the original units of the variables. The standardized regression equation (Formula 13.10) allows the researcher to assess easily the relative importance of the various independent variables by comparing the beta-weights. *(All end-of-chapter problems offer opportunities to compute and interpret beta-weights.)*

Multiple Correlation

We use the multiple regression equations to disentangle the separate direct effects of each independent variable on the dependent variable. Using **multiple correlation** techniques, we can also ascertain the combined effects of all independent variables on the dependent variable. We do so by computing the **multiple correlation coefficient (R)** and the **coefficient of multiple determination (R^2)**. The value of the latter statistic represents the proportion of the variance in Y that is explained by all the independent variables combined.

We have seen that number of children (X_1) explains 25% of the variance in Y ($r^2_{y1} = (0.50)^2 = 0.25$, which, when multiplied by 100, is 25%) by itself and that husband's education explains 9% of the variance in Y ($r^2_{y2} = (-0.30)^2 = 0.09$ which, when multiplied by 100, is 9%). The two zero-order correlations cannot simply be added together to ascertain their combined effect on Y, because the two independent variables are also correlated with each other and, therefore, they will "overlap" in their effects on Y and explain some of the same variance. This overlap is eliminated in Formula 13.11:

FORMULA 13.11

$$R^2 = r^2_{y1} + r^2_{y2.1}\left(1 - r^2_{y1}\right)$$

where R^2 = the coefficient of multiple determination
r^2_{y1} = the zero-order correlation between Y and X_1, the quantity squared
$r^2_{y2.1}$ = the partial correlation of Y and X_2, while controlling for X_1, the quantity squared

The first term in this formula (r^2_{y1}) is the coefficient of determination for Y and X_1, or the amount of variation in Y explained by X_1 by itself. To this quantity we add the amount of the variation remaining in Y (given by $1 - r^2_{y1}$) that can be explained by X_2 after the effect of X_1 is controlled ($r^2_{y2.1}$). Basically, Formula 13.11 allows X_1 to explain as much of Y as it can and then adds in the effect of X_2 after X_1 is controlled (thus eliminating the "overlap" in the variance of Y that X_1 and X_2 have in common).

Computing and Interpreting R^2

To observe the combined effects of number of children (X_1) and husband's education (X_2) on husband's housework (Y), we need two quantities. The correlation between X_1 and Y ($r_{y1} = 0.50$) has already been found. Before we can solve

ONE STEP AT A TIME **Computing the Coefficient of Multiple Determination (R^2)**

These procedures apply when there are two independent variables and one dependent variable. For more complex situations, use a computerized statistical package such as SPSS to do the calculations.

Step	***Operation***
	To compute the coefficient of multiple determination by using Formula 13.11:
1.	Find the value of the partial correlation coefficient $r_{y2.1}$.
2.	Square the value you found in step 1.
3.	Square the value of r_{y1}.
4	Subtract the value you found in step 3 from 1.00.
5.	Multiply the value you found in step 4 by the value you found in step 2.
6.	Add the value you found in step 5 to the value you found in step 3. *The result is the coefficient of multiple determination* (R^2).

ONE STEP AT A TIME **Interpreting R^2**

The coefficient of multiple determination (R^2) is the total amount of the variation in Y explained by all independent variables combined.

Formula 13.11, we must first calculate the partial correlation of Y and X_2 while controlling for X_1 ($r_{y2.1}$). This quantity can be found by Formula 13.1 but we will simply report that $r_{y2.1} = -0.08$.

Formula 13.11 can now be solved for our example problem:

$$R^2 = r_{y1}^2 + r_{y2.1}^2(1 - r_{y1}^2)$$
$$R^2 = (0.50)^2 + (-0.08)^2(1 - 0.50^2)$$
$$R^2 = 0.25 + (0.006)(1 - 0.25)$$
$$R^2 = 0.25 + 0.005$$
$$R^2 = 0.255$$

The first independent variable (X_1), number of children, explains 25% of the variance in Y by itself. To this total, the second independent variable (X_2), husband's education, adds only a half a percent, for a total explained variance of 25.5%. In combination, the two independent variables explain a total of 25.5% of the variation in the dependent variable. *(All end-of-chapter problems offer opportunities to find and interpret* R *and* R^2. *The first three problems stress hand calculations and all others require the use of SPSS.)*

Using SPSS to Conduct a Regression Analysis

For reasons that should be obvious at this point in the chapter, researchers use computerized statistical packages like SPSS rather than hand calculations to conduct regression analysis. In this installment of "Using SPSS," we will conduct a regression analysis using the *States* data set. We'll take the rate of traffic fatalities (*TrafDths11*) as the dependent variable. Our independent variables will be the percentage of the state's population older than 65 (*Older*) and population density (*PopDense*), or the number of people per square mile in the state. Will states with older populations have lower or higher rates of traffic deaths? Will more densely populated states be more dangerous for drivers?

To begin:

1. Load the *States* database.
2. From the main menu bar, click "Analyze, Regression," and "Linear."
3. On the "Linear Regression" screen, place *TrafDths11* in the "Dependent" window and *Older* and *PopDense* in the "Independent" window.
4. Click **Statistics** and check "Descriptives." Click "Continue" to return to the "Linear Regression" screen.
5. Click **OK** and results will be printed to the output screen in several separate boxes, only some of which are relevant for our purposes.
6. Find the zero-order correlations in the "**Correlations**" box.
7. Find the values for R ($R = 0.48$) and R^2 ($R^2 = 0.23$) in the "Model Summary" box.
8. The "Coefficients" box displays the values of a and b under **Unstandardized Coefficients**. Look in the column labeled **B**. The value in the first row is a ($a = 0.83$) and the entries in the second and third rows are the partial slopes b ($b = 0.03$ for *Older* and -0.01 for *PopDense*).
9. In the same box, find the beta-weights listed under **Standardized Coefficients** and **Beta** ($b_1^* = 0.17$ and $b_2^* = -0.48$).

We can begin the analysis by noting the zero-order correlations. The rate of traffic fatalities has a weak positive relationship ($r = 0.09$) with age (the higher the percentage of older people, the higher the rate of traffic deaths) and a moderate negative relationship ($r = -0.45$) with population density (states with lower population density have higher rates of traffic deaths).

The multiple regression equation, which could be used to predict scores on Y for various combinations of scores on the independent variables, is:

$$TradDths11 = 0.83 + (0.03)Older + (-0.01)PopDense$$

Together, the two independent variables explain 23% ($R^2 = 0.23$) of the variance in the traffic fatality rates across the states. The beta-weights show that the net effect of age is positive while the effect of population density is negative and that the latter variable has the stronger effect. The traffic fatality rate tends to increase as populations grow older and decrease as populations grow denser. The latter relationship is a reflection of the higher rates of traffic deaths in the less densely settled Western states. The beta-weights are displayed in the standardized multiple regression equation:

$$TradDths11 = (0.17)Older + (-0.48)PopDense$$

Applying Statistics 13.2 Multiple Regression and Correlation

Five recently divorced men have been asked to rate subjectively the success of their adjustment to single life, on a scale ranging from 5 (very successful adjustment) to 1 (very poor adjustment). Is adjustment related to the length of time married or to socioeconomic status as measured by yearly income?

Case	Adjustment (Y)	Years Married (X_1)	Income (X_2)
A	5	5	\$30,000
B	4	7	\$45,000
C	4	10	\$25,000
D	3	2	\$27,000
E	1	15	\$17,000
	$\bar{Y} = 3.4$	$\bar{X}_1 = 7.8$	$\bar{X}_2 = 28,800$
	$s_y = 1.4$	$s_1 = 4.5$	$s_2 = 9173.88$

The zero-order correlations among these three variables are

	Adjustment (Y)	Years Married (X_1)	Income (X_2)
Adjustment (Y)	1.00	−0.62	0.62
Years married (X_1)		1.00	−0.49
Income (X_2)			1.00

These results suggest a strong but opposite relationship between each independent variable and adjustment. Adjustment decreases as years married increases, and it increases as income increases.

To find the multiple regression equation, we must find the partial slopes.

For years married (X_1),

$$b_1 = \left(\frac{s_y}{s_1}\right)\left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)$$
$$b_1 = \left(\frac{1.4}{4.5}\right)\left(\frac{(-0.62) - (0.62)(-0.49)}{1 - (-0.49)^2}\right)$$
$$b_1 = (.31)\left(\frac{(-0.62) - (-0.30)}{1 - (.24)}\right)$$
$$b_1 = (.31)\left(\frac{-0.32}{0.76}\right)$$
$$b_1 = (.31)(-0.42)$$
$$b_1 = -0.13$$

For income (X_2),

$$b_2 = \left(\frac{s_y}{s_2}\right)\left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right)$$
$$b_2 = \left(\frac{1.4}{9,173.88}\right)\left(\frac{(0.62) - (-0.62)(-0.49)}{1 - (-0.49)^2}\right)$$
$$b_2 = (0.00015)\left(\frac{(0.62) - (0.30)}{1 - 0.24}\right)$$
$$b_2 = (0.00015)\left(\frac{0.32}{0.76}\right)$$
$$b_2 = (0.00015)(0.42)$$
$$b_2 = 0.00006$$

The Y intercept would be

$$a = \bar{Y} - b_1\bar{X}_2 - b_2\bar{X}_2$$
$$a = 3.4 - (-0.13)(7.8) - (0.00006)(28,800)$$
$$a = 3.4 - (-1.01) - (1.73)$$
$$a = 2.68$$

The multiple regression equation is

$$Y = a + b_1X_1 + b_2X_2$$
$$Y = 2.68 + (-0.13)X_1 + (0.00006)X_2$$

(*continued*)

What adjustment score could we predict for a male who had been married 30 years ($X_1 = 30$) and had an income of \$50,000 ($X_2 = 50{,}000$)?

$$Y' = 2.68 + (-0.13)(30) + (0.00006)(50{,}000)$$
$$Y' = 2.68 + (-3.9) + (3.0)$$
$$Y' = 1.78$$

To assess which of the two independent variables has the stronger effect on adjustment, the standardized partial slopes must be computed.

For years married (X_1),

$$b_1^* = b_1\left(\frac{s_1}{s_y}\right)$$
$$b_1^* = (-0.13)\left(\frac{4.5}{1.4}\right)$$
$$b_1^* = (-0.13)(3.21)$$
$$b_1^* = -0.42$$

For income (X_2),

$$b_2^* = b_2\left(\frac{s_2}{s_y}\right)$$
$$b_2^* = (0.00006)\left(\frac{9{,}173.88}{1.4}\right)$$
$$b_2^* = (0.00006)(6552.77)$$
$$b_2^* = 0.39$$

The standardized regression equation is

$$Z_y = b_1^* Z_1 + b_2^* Z_2$$
$$Z_y = (-0.42)Z_1 + (0.39)Z_2$$

The independent variables have equal but opposite effects on adjustment.

To assess the combined effects of the two independents on adjustment, the coefficient of multiple determination must be computed:

$$R^2 = r_{y1}^2 + r_{y2.1}^2(1 - r_{y1}^2)$$
$$R^2 = (-0.62)^2 + (0.46)^2(1 - (-0.62)^2)$$
$$R^2 = 0.38 + 0.21(1-0.38)$$
$$R^2 = 0.38 + 0.21(0.62)$$
$$R^2 = 0.38 + 0.13$$
$$R^2 = 0.51$$

The first independent variable, years married, explains 38% of the variation in adjustment by itself. To this quantity, income explains an additional 13% of the variation in adjustment. Taken together, the two independent variables explain a total of 51% of the variation in adjustment.

STATISTICS IN EVERYDAY LIFE

Teen Pregnancies, Revisited

In Chapter 12, we looked at some of the correlates of birth rates for teenagers, using data from the 50 states (see Table 12.5). We found a strong negative bivariate correlation (−0.78) with a measure of education and a strong positive relationship (0.88) with a measure of poverty. What are the combined effects of poverty *and* education on the teen birth rate?

By itself, education explains 61% of the variance in the state teen birth rate. To this, poverty adds 11% for a combined explained variance of 72% ($R^2 = 0.72$).

The beta weights show that poverty has a positive effect (after controlling for education) and that education has a negative effect (after controlling for poverty) and that poverty ($b^* = 0.54$) has a stronger effect than education ($b^* = -0.37$).

Note two things about these results. First, these are unusually high correlations for social science research. It is far more common to deal with correlations in the "weak to moderate" range, partly because our measurements are imprecise.

Second, based on these results, it is tempting to conclude that poor, less educated teenage girls are the most likely to get pregnant. While this conclusion might seem logical, remember that the data refer to *states*, not individuals and these results show only that *states* with higher levels of poverty and lower levels of education tend to have higher birth rates for teens. We would need more information and more research before we could make firm conclusions about the characteristics of pregnant teenagers.

The Limitations of Multiple Regression and Correlation

Multiple regression and correlation are very powerful tools for analyzing the interrelationships among three or more variables. The techniques presented in this chapter permit the researcher to predict scores on one variable from two or more other variables, to distinguish among independent variables in terms of the importance of their direct effects on a dependent variable, and to ascertain the total effect of a set of independent variables on a dependent variable. In terms of the flexibility of the techniques and the volume of information they can supply, multiple regression and correlation represent some of the most powerful statistical techniques available to social science researchers.

Powerful tools are not cheap. They require high-quality data, and measurement at the interval-ratio level can be difficult to attain in social science research. Furthermore, these techniques assume that the interrelationships among the variables follow a particular form. First, they assume that each independent variable has a linear relationship with the dependent variable. How well a given set of variables meets this assumption can be quickly checked with scatterplots.

Second, these techniques assume that there is no interaction among the variables in the equation. If there is interaction among the variables, it will not be possible to estimate or predict the dependent variable accurately by simply adding the effects of the independent variables. There are techniques for handling interaction among the variables in the set, but these techniques are beyond the scope of this text.

Third, multiple regression and correlation assume that the independent variables are uncorrelated with each other. Strictly speaking, this condition means that the zero-order correlation among all pairs of independent variables should be zero; however, practically, we act as if this assumption has been met if the intercorrelations among the independent variables are low.

To the extent that these assumptions are violated, the regression coefficients (especially partial and standardized slopes) and the coefficient of multiple determination (R^2) become less and less trustworthy and the techniques less and less useful. A careful inspection of the bivariate scatterplots will help assess the extent to which the assumptions are reasonable. Finally, we should note that we have covered only the simplest applications of partial correlation and multiple regression and correlation. In terms of logic and interpretation, the extensions to situations involving more variables are relatively straightforward. However, the computations for these situations are extremely complex. If you are faced with a situation involving more than three variables, turn to one of the computerized statistical packages commonly available on college campuses (e.g., SPSS or SAS). These programs require minimal computer literacy and can handle complex calculations in, literally, the blink of an eye. Efficient use of these packages will enable you to avoid drudgery and will free you to do what social scientists everywhere enjoy doing most: pondering the meaning of your results and, by extension, the nature of social life.

STATISTICS
IN EVERYDAY LIFE

Statistics and Baseball

Baseball is a paradise for statistics junkies. Virtually every aspect of what happens on the field is recorded, and there are decades of data to analyze. Can this mountain of information be used for purposes other than settling arguments over trivia among baseball fans?

The answer is a resounding yes, according to Michael Lewis, author of *Moneyball* (which, in 2011, was made into a movie starring Brad Pitt). Lewis relates the story of Billy Beane, who, as general manager of the Oakland A's, was faced with the challenge of fielding a competitive team with one of the smallest budgets in the sport. Beane accomplished this, in part, by hiring statisticians who conducted a series of regression analyses with won/lost records as the dependent variable and every conceivable indicator of performance as independent variables.

Beane and his team of researchers were able to identify a number of measures of player effectiveness that had strong relationships with success (i.e., had high beta-weights) but were not highly valued by other baseball executives. He used these criteria to identify and recruit very effective players for relatively low wages. Under Beane's management, the A's were quite successful (although never winning the World Series), and much of his regression-based system has now become standard practice for baseball and this style of analysis has been adapted for use in other sports.

Source: Lewis, Michael. 2004. *Moneyball.* New York: Norton.

BECOMING A CRITICAL CONSUMER: Reading the Professional Literature

It is unlikely that you will encounter multiple regression in the popular press or in your everyday conversations. On the other hand, multiple regression analysis is widely used in social science research and it is very likely that you will have to deal with articles using this technique—or one of its many variants—in the professional research literature. These articles might appear to be hopelessly complex at first glance and, indeed, they are written for other professionals and assume a high level of statistical sophistication on the part of the reader.

Nevertheless, without denying the challenge, it is quite possible for "ordinary people" to distill the essence of these articles by following a few guidelines and looking for some central elements. The key is to focus on the words, not the numbers. That is, read the text to see what the authors have to say about their results, not the (perhaps impenetrable) array of numbers and symbols. The details of the statistical analysis may be beyond your understanding but you can almost always decipher the words.

Results will typically be presented in summary tables that report unstandardized and standardized coefficients, R^2, and, possibly, other information. An example of this kind of summary table would look like this:

Independent Variables	R^2	Beta-Weights
X_1	0.17	0.47
X_2	0.23	0.32
X_3	0.27	0.16

This table reports that the first independent variable, X_1, has the strongest direct relationship with the dependent variable and explains 17% of the variance in the dependent variable by itself ($R^2 = 0.17$). The second independent variable, X_2, adds 6% to the explained variance ($R^2 = 0.23$ after X_2 is entered into the equation). The third independent variable, X_3, adds 4% to the explained variance of the dependent variable ($R^2 = 0.27$ after X_3 is entered into the equation).

Let's briefly examine an example of an actual research report. Research has established that high levels of violence in the media increase the public's fear of criminal victimization, but sociologist Valerie

(*continued*)

BECOMING A CRITICAL CONSUMER (*continued*)

Callanan wanted to examine several possibilities: Do different types of media have different effects? Does media violence affect different communities in different ways?

Some of her results, based on a large random sample of California residents, are presented in the following table. Several of her independent variables have been deleted to conserve space.

Fear of crime decreases as education increases. People who consumed violence in the media (especially from reality shows) are more fearful, as are victims of crime. All together, these variables explained 17% of the variance in fear of crime.

Multiple Regression Results with Fear of Crime as the Dependent Variable (*N* = 3174)†

Variable	Beta-Weight
Education	−0.17***
Local TV news	0.06***
Crime dramas	0.04
Crime reality shows	0.12**
Crime victim	0.43***
$R^2 =$	0.17

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

†Adapted from Table 2 in the original article.

To learn more, refer to the source given below.

Source: Callanan, Valerie. 2012. “Media Consumption, Perceptions of Crime Risk and Fear of Crime: Examining Race/Ethnic Differences.” *Sociological Perspectives* 55: 93–115.

SUMMARY

1. In partial correlation, we observe what happens to a bivariate relationship when a third variable is controlled. With this technique, we can detect direct and spurious or intervening relationships between X and Y.
2. We can predict scores on a dependent variable from scores on more than one independent variable by using the multiple regression equation with partial slopes. We can disentangle the relative importance of the independent variables with standardized partial slopes or beta-weights.
3. The coefficient of multiple determination (R^2) shows the combined effects of all independent variables on the dependent variable in terms of the proportion of the total variation in Y that is explained.
4. Partial correlation and multiple regression and correlation are some of the most powerful tools available to the researcher and demand high-quality measurement and relationships among the variables that are linear and noninteractive. Further, correlations among the independent variables must be low (preferably 0). Although the price is high, these techniques pay considerable dividends in the volume of precise and detailed information they generate about the interrelationships among the variables.

SUMMARY OF FORMULAS

FORMULA 13.1 Partial correlation coefficient:
$$r_{yx.z} = \frac{r_{yx} - (r_{yz})(r_{xz})}{\sqrt{1 - r_{yz}^2}\sqrt{1 - r_{xz}^2}}$$

FORMULA 13.2 Least-squares regression line (bivariate):
$$Y = a + bX$$

FORMULA 13.3 Least-squares multiple regression line:
$$Y = a + b_1X_1 + b_2X_2$$

FORMULA 13.4 Partial slope for X_1:
$$b_1 = \left(\frac{s_y}{s_1}\right)\left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)$$

FORMULA 13.5 Partial slope for X_2:
$$b_2 = \left(\frac{s_y}{s_2}\right)\left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right)$$

FORMULA 13.6 Y intercept:
$$a = \bar{Y} - b_1\bar{X}_2 - b_2\bar{X}_2$$

FORMULA 13.7	Standardized partial slope (beta-weight) for X_1:	$b_1^* = b_1\left(\frac{s_1}{s_y}\right)$
FORMULA 13.8	Standardized partial slope (beta-weight) for X_2:	$b_2^* = b_2\left(\frac{s_2}{s_y}\right)$
FORMULA 13.9	Standardized least-squares regression line:	$Z_y = a_z + b_1^*Z_1 + b_2^*Z_2$
FORMULA 13.10	Standardized least-squares regression line (simplified):	$Z_y = b_1^*Z_1 + b_2^*Z_2$
FORMULA 13.11	Coefficient of multiple determination:	$R^2 = r_{y1}^2 + r_{y2.1}^2(1 - r_{y1}^2)$

GLOSSARY

Beta-weight (b^*). See **Standardized partial slope.**

Coefficient of multiple determination (R^2). A statistic that equals the proportion of the total variation in the dependent variable that is explained by all independent variables combined.

Control variable. In multivariate statistical analysis, a third variable (Z) whose effect on a bivariate relationship is held constant.

Multiple correlation. A multivariate technique for examining the combined effects of more than one independent variable on a dependent variable.

Multiple correlation coefficient (R). A statistic that indicates the strength of the correlation between a dependent variable and two or more independent variables.

Multiple regression. A multivariate technique that breaks down the separate effects of the independent variables on the dependent variable; used to make predictions of the dependent variable.

Partial correlation. A multivariate technique for examining a bivariate relationship while controlling for other variables.

Partial correlation coefficient. A statistic that shows the relationship between two variables while controlling for other variables; $r_{yx.z}$ is the symbol for the partial correlation coefficient when controlling for one variable.

Partial slope. In a multiple regression equation, the slope of the relationship between a particular independent variable and the dependent variable while controlling for all other independent variables in the equation.

Standardized partial slope (beta-weight). The slope of the relationship between a particular independent variable and the dependent variable when all scores have been normalized.

Zero-order correlation. Correlation coefficient for bivariate relationships.

PROBLEMS

The data sets in the first three problems are unrealistically small to allow you to practice the computational routines presented in this chapter. Most problems use SPSS.

13.1 *SOC* In problem 12.1, data regarding voter turnout in five cities was presented. For the sake of convenience, the data for three of the variables are presented again here along with descriptive statistics and zero-order correlations.

City	Turnout	Unemployment Rate	% Negative Ads
A	55	5	60
B	60	8	63
C	65	9	55
D	68	9	53
E	70	10	48
Mean =	63.6	8.2	55.8
s =	5.5	1.7	5.3

	Unemployment Rate	Negative Ads
Turnout	0.95	−0.87
Unemployment Rate		−0.70

a. Compute the partial correlation coefficient for the relationship between turnout (Y) and unemployment (X) while controlling for the effect of negative advertising (Z). What effect does this control variable have on the bivariate relationship? Is the relationship between turnout and unemployment direct? *(HINT: Use Formula 13.1)*

b. Compute the partial correlation coefficient for the relationship between turnout (Y) and negative advertising (X) while controlling for the effect of unemployment (Z). What effect does this have on the bivariate relationship? Is the relationship between turnout and negative advertising direct? *(HINT: Use Formula 13.1 You will need this partial correlation to compute the multiple correlation coefficient.)*

c. Find the unstandardized multiple regression equation with unemployment (X_1) and negative ads (X_2) as the independent variables. What turnout would be expected in a city in which the unemployment rate was 10% and 75% of the campaign ads were negative? (*HINT: Use Formulas 13.4 and 13.5 to compute the partial slopes and then use Formula 13.6 to find* a*, the* Y *intercept. The regression line is stated in Formula 13.3. Substitute 10 for* X_1 *and 75 for* X_2 *to compute predicted* Y.)

d. Compute beta-weights for each independent variable. Which has the stronger impact on turnout? (*HINT: Use Formulas 13.7 and 13.8 to calculate the beta-weights.*)

e. Compute the coefficient of multiple determination (R^2). How much of the variance in voter turnout is explained by the two independent variables combined? *(HINT: Use Formula 13.11. You calculated* $r^2_{y2.1}$ *in part b of this problem.)*

f. Write a paragraph summarizing your conclusions about the relationships among these three variables.

13.2 **PS** Problem 12.4 presented data on ten precincts. The information is reproduced here.

Precinct	Percentage Democrat	Percentage Minority	Percentage Voter Turnout
A	33	9	36
B	78	15	60
C	62	18	64
D	50	10	56
E	45	12	55
F	56	8	52
G	85	20	25
H	25	0	42
I	13	5	89
J	33	9	88

Take voter turnout as the dependent variable and do the following:

a. Find the multiple regression equations (unstandardized).

b. What turnout would you expect for a precinct in which 0% of the voters were Democrats and 5% were minorities?

c. Compute beta-weights for each independent variable and compare their relative effect on turnout. Which was the more important factor?

d. Compute R^2.

e. Write a paragraph summarizing your findings.

13.3 **SW** Twelve families have been referred to a counselor, and she has rated each of them on a scale that measures family cohesiveness. Also, she has information on family income and number of children currently living at home. Take family cohesion as the dependent variable.

Family	Cohesion Score	Income	Number of Children
A	10	30,000	5
B	10	70,000	4
C	9	35,000	4
D	5	25,000	0
E	1	55,000	3
F	7	40,000	0
G	2	60,000	2
H	5	30,000	3
I	8	50,000	5
J	3	25,000	4
K	2	45,000	3
L	4	50,000	0

a. Find the multiple regression equations (unstandardized).

b. What level of cohesion would be expected in a family with an income of $20,000 and six children?

c. Compute beta-weights for each independent variable and compare their relative effect on cohesion. Which was the more important factor?

d. Compute R^2.

e. Write a paragraph summarizing your findings.

Statistical Analysis Using SPSS

13.4 **SOC** Load the *GSS2012* data set and analyze the relationships between church attendance (the dependent or Y variable), and two independent variables: age and number of children. You analyzed some of these relationships in Problem 12.7. Now, we will use both partial correlation and multiple regression and correlation to further examine these relationships.

- Zero-order correlations.
  - Click **Analyze → Correlate → Bivariate**.
  - Enter *attend*, *age, and childs* in the **Variables:** window.
  - Click **OK**.
- Partial Correlation Analysis: Is the relationship between *attend* and *age* affected by *childs*?
  - Click **Analyze → Correlate → Partial**.
  - Enter *attend* and *age* in the "Variables:" window and *childs* in the "Controlling for:" window.
  - Click **OK**.
- Partial Correlation Analysis: Is the relationship between *attend* and *childs* affected by *age*?
  - Click **Analyze → Correlate → Partial**.
  - Enter *attend* and *childs* in the "Variables:" window and *age* in the "Controlling for:" window.
  - Click **OK**.
- Multiple Regression and Correlation: What are the effects of *childs* and *age* on *attend*?
  - Click **Analyze → Regression → Linear**.
  - Move *attend* to the "Dependent" window and *age* and *childs* to the "Independent" window
  - Click **Statistics** and check "Descriptives". Click **Continue** to return to the "Linear Regression" screen.
  - Click **OK**.

a. Summarize the results for the partial correlation analysis in a paragraph in which you report the value of the zero-order and partial correlations for all relationships. Does the relationship between *attend* and *age* seem to be direct? How about the relationship between *attend* and *childs*?

b. State the unstandardized multiple regression equation. *(HINT: The values for* a *and* b *are in the "Coefficients" box of the output, under the column labeled* B. *The value in the first row is* a *and the values in the second and third rows are the slopes or* b.*)*

$$Y = \_\_\_ + \_\_\_ X_1 + \_\_\_ X_2$$

c. State the standardized multiple regression equation. What is the direction of each relationship? Which independent variable had the stronger effect on *attend*? *(HINT: The beta-weights are in the "Coefficients" box, under "Standardized Coefficients" and "Beta").*

$$Z_y = \_\_\_\_ Z_1 + \_\_\_ Z_2$$

d. Report the value of R^2. What percentage of the variance in *attend* is explained by the two independent variables combined? How does this compare to the amount of the variance explained by each independent variable alone? *(HINTS: R^2 is in the "Model Summary" box. You can compute r^2 from the r's in the "Correlations" window of the output.)*

Bivariate:	*r*	r^2
attend and		
childs		
age		
Multiple:	$R^2 =$	

13.5 SOC Load the *GSS2012* data set and analyze the relationships between number of hours spent on the Internet each week (the dependent or *Y* variable), and two independent variables: age and years of education. We will use both partial correlation and multiple regression and correlation to examine these relationships.

- Partial Correlation Analysis: Is the relationship between *wwwhr* and *age* affected by *educ*?
  - Get the zero-order correlations first:
    - Click **Analyze → Correlate → Bivariate**.
    - Enter *wwwhr, educ,* and *age* in the **Variables:** window.
  - Click **Analyze → Correlate → Partial**
  - Enter *wwwhr* and *age* in the "Variables:" window and *educ* in the "Controlling for:" window.
  - Click **OK**.
- Multiple Regression and Correlation: What are the effects of *educ* and *age* on *wwwhr*?
  - Click **Analyze → Regression → Linear**.
  - Move *wwwhr* to the "Dependent" window and *age* and *educ* to the "Independent" window
  - Click **Statistics** and check "Descriptives." Click **Continue** to return to the "Linear Regression" screen.
  - Click **OK**.

a. Summarize the results for the partial correlation analysis in a paragraph in which you report the value of the zero-order and partial correlation. Does the relationship between *wwwhr* and *age* seem to be direct?

b. State the unstandardized multiple regression equation. (*HINT: The values for* a *and* b *are in the "Coefficients" box of the output, under the column labeled* B. *The value in the first row is* a *and the values in the second and third rows are the slopes or* b.)

$$Y = \_\_\_ + \_\_\_ X_1 + \_\_\_ X_2$$

c. State the standardized multiple regression equation. What is the direction of each relationship?

Which independent variable had the stronger effect on *wwwhr*? (*HINT: The beta-weights are in the "Coefficients" box, under "Standardized Coefficients" and "Beta"*).

$$Z_y = \_\_\_\_ Z_1 + \_\_\_ Z_2$$

d. Report the value of R^2. What percentage of the variance in *wwwhr* is explained by the two independent variables combined? How does this compare to the amount of the variance explained by each independent variable alone? *(HINTS:* R^2 *is in the "Model Summary" box and you can compute* r^2 *from the r's in the "Correlations" window of the output.)*

Bivariate:	*r*	r^2
wwwhr and		
educ		
age		
Multiple:	$R^2 =$	

13.6 SOC/CJ Use the *States* data set to analyze the homicide rate (*Hom12*). The independent variables will be population density (*Popdense*) and unemployment rate (*Unemployment*)

- Click **Analyze → Regression → Linear**.
- Move *Hom12* to the "Dependent" window and *Popdense* and *Unemployment* to the "Independent" window.
- Click **Statistics** and check "Descriptives." Click **Continue** to return to the "Linear Regression" screen.
- Click **OK**.

a. State the unstandardized multiple regression equation. (*HINT: The values for* a *and* b *are in the "Coefficients" box of the output, under the column labeled* B. *The value in the first row is* a *and the values in the second and third rows are the slopes or* b.)

$$Y = \_\_\_ + \_\_\_ X_1 + \_\_\_ X_2$$

b. State the standardized multiple regression equation. What is the direction of each relationship? Which independent variable had the stronger effect on *Hom12*? *(HINT: The beta-weights are in the "Coefficients" box, under "Standardized Coefficients" and "Beta").*

$$Z_y = \_\_\_\_ Z_1 + \_\_\_ Z_2$$

c. Report the value of R^2. What percentage of the variance in *Hom12* is explained by the two independent variables combined? How does this compare to the amount of the variance explained by each independent variable alone? *(HINTS:* R^2 *is in the "Model Summary" box and you can compute* r^2 *from the r's in the "Correlations" window of the output.)*

Bivariate:	*r*	r^2
Hom12 and		
Popdense		
Unemployment		
Multiple:	$R^2 =$	

13.7 SOC/CJ Use the *States* data set to analyze the rate of car theft (*Carthft12*). The independent variables will be population density (*Popdense*) and unemployment rate (*Unemployment*)

- Click **Analyze → Regression → Linear**.
- Move *Carthft12* to the "Dependent" window and *Popdense* and *Unemployment* to the "Independent" window.
- Click **Statistics** and check "Descriptives." Click **Continue** to return to the "Linear Regression" screen.
- Click **OK**.

a. State the unstandardized multiple regression equation. (*HINT: The values for* a *and* b *are in the "Coefficients" box of the output, under the column labeled* B. *The value in the first row is* a *and the values in the second and third rows are the slopes or* b.)

$$Y = \_\_\_ + \_\_\_ X_1 + \_\_\_ X_2$$

b. State the standardized multiple regression equation. What is the direction of each relationship? Which independent variable had the stronger effect on *Carthft12*? *(HINT: The beta-weights are in the "Coefficients" box, under "Standardized Coefficients" and "Beta.")*

$$Z_y = \_\_\_\_ Z_1 + \_\_\_ Z_2$$

c. Report the value of R^2. What percentage of the variance in *Carthft12* is explained by the two independent variables combined? How does this compare to the amount of the variance explained by each independent variable alone? *(HINTS:* R^2 *is in the "Model Summary" box and you can compute* r^2 *from the r's in the "Correlations" window of the output.)*

Bivariate:	*r*	r^2
Carthft12 and		
Popdense		
Unemployment		
Multiple:	$R^2 =$	

13.8 SOC In Problem 12.8, you used the *States* database to analyze the bivariate relationships between region (South/Non-South), education (percentage of the population with a college degree), and infant mortality rates. Now, you will take education and region as the independent variables and look at their combined effects on infant mortality rates.

- Click **Analyze → Regression → Linear**.
- Move *InfantMort* to the "Dependent" window and *College* and *SthDUMMY* to the "Independent" window.
- Click **Statistics** and check "Descriptives." Click **Continue** to return to the "Linear Regression" screen.
- Click **OK**.

a. State the unstandardized multiple regression equation.

$$Y = \_\_\_ + \_\_\_X_1 + \_\_\_X_2$$

b. State the standardized multiple regression equation. What is the direction of each relationship? Which independent variable had the stronger effect on *InfantMort*?

$$Zy = \_\_\_\_ Z_1 + \_\_\_ Z_2$$

c. Report the value of R^2. What percentage of the variance in *InfantMort* is explained by the two independent variables combined? How does this compare to the amount of the variance explained by each independent variable alone? *(HINTS:* R^2 *is in the "Model Summary" box and you can compute* r^2 *from the r's in the "Correlations" window of the output.)*

Bivariate:	r	r^2
InfantMort and		
College		
SthDummy		
Multiple:	$R^2 =$	

13.9 SOC In Problem 12.11, you used the *Intl-Pop* database to analyze the bivariate relationships between cell phone usage, affluence (gross national income per capita), and urbanization. Now, you will take affluence and urbanization as the independent variables and look at their combined effects on cell phone usage.

- Click **Analyze → Regression → Linear**.
- Move *CellPhones* to the "Dependent" window and *GNIcap* and *Urban* to the "Independent" window.
- Click **Statistics** and check "Descriptives." Click **Continue** to return to the "Linear Regression" screen.
- Click **OK**.

a. State the unstandardized multiple regression equation.

$$Y = \_\_\_ + \_\_\_X_1 + \_\_\_X_2$$

b. State the standardized multiple regression equation. What is the direction of each relationship? Which independent variable had the stronger effect on *CellPhones*?

$$Z_y = \_\_\_\_ Z_1 + \_\_\_ Z_2$$

c. Report the value of R^2. What percentage of the variance in *CellPhones* is explained by the two independent variables combined? How does this compare to the amount of the variance explained by each independent variable alone?

Bivariate:	r	r^2
CellPhones and		
GNIcap		
Urban		
Multiple:	$R^2 =$	

13.10 PS Use the *Intl-Pop* database to analyze the correlates of political and civil rights (*Rights*). Take affluence and income inequality as the independent variables and look at their combined effects on *Rights*.

NOTE: "Rights" is scored so that lower scores mean greater freedom.

- Click **Analyze → Regression → Linear**.
- Move *Rights* to the "Dependent" window and *GNIcap* and *IncomeIneq* to the "Independent" window.
- Click **Statistics** and check "Descriptives." Click **Continue** to return to the "Linear Regression" screen.
- Click **OK**.

a. State the unstandardized multiple regression equation.

$$Y = \_\_\_ + \_\_\_X_1 + \_\_\_X_2$$

b. State the standardized multiple regression equation. What is the direction of each relationship? Which independent variable had the stronger effect on *Rights*?

$$Z_y = \_\_\_\_ Z_1 + \_\_\_ Z_2$$

c. Report the value of R^2. What percentage of the variance in *Rights* is explained by the two independent variables combined? How does this compare to the amount of the variance explained by each independent variable alone?

Bivariate:	*r*	r^2
Rights and		
GNIcap		
IncomeIneq		
Multiple:	$R^2 =$	

13.11 SOC/PS In Problem 12.13, you analyzed the relationships between voter turnout (*Voters*), education (*College*), and affluence (*MdHHInc*) for the States database. Now, analyze the combined effects of the independent variables on turnout. With the States database open:

- Click **Analyze** → **Regression** → **Linear**.
- Move *Voters* to the "Dependent" window and *College* and *MdHHInc* to the "Independent" window.
- Click **Statistics** and check "Descriptives." Click **Continue** to return to the "Linear Regression" screen.
- Click **OK**.

a. State the unstandardized multiple regression equation.

$$Y = \_\_\_ + \_\_\_ X_1 + \_\_\_ X_2$$

b. State the standardized multiple regression equation. What is the direction of each relationship? Which independent variable had the stronger effect on *Voters*?

$$Z_y = \_\_\_\_ Z_1 + \_\_\_ Z_2$$

c. Report the value of R^2. What percentage of the variance in *Voters* is explained by the two independent variables combined? How does this compare to the amount of the variance explained by each independent variable alone?

Bivariate:	*r*	r^2
Voters and		
College		
MdHHInc		
Multiple:	$R^2 =$	

YOU ARE THE RESEARCHER

A Multivariate Analysis of TV Viewing and Success

The two following projects continue the investigations begun in Chapter 12. You will use the **Regression** command to analyze the combined effects of your independent variables (including dummy variables if you used any) on the dependent variables from Chapter 12.

Project 1: Who Watches TV?

This project follows up on Project 1 from Chapter 12. The dependent variable is *tvhours*, which, as you recall, measures the number of hours per day the respondent watches television.

Step 1: Choosing Independent Variables

Choose two of the four independent variables you selected in Project 1 from Chapter 12. All other things being equal, you might choose the variables that had the strongest relationship with *tvhours*. Remember that independent variables *cannot* be nominal in level of measurement unless you recode the variable into a dummy variable. You may use any interval-ratio or ordinal-level variables with more than three categories or scores. Once you have selected your variables, list them in the table below and describe exactly what they measure.

SPSS Variable Name	What Exactly Does This Variable Measure?

Step 2: Stating Hypotheses

Restate your hypotheses from Chapter 12 and add some ideas about the relative importance of your two independent variables. Which do you expect to have the stronger effect? Why?

SPSS Variable Name	Hypotheses

Step 3: Running the Regression Procedure

Click **Analyze → Regression → Linear** and place *tvhours* in the "Dependent:" box and your selected independent variables in the "Independent(s):" box. Click **Statistics** and **Descriptives**, and then click **OK** to get your results.

Step 4: Recording Results

Summarize your results by filling in the blanks below with information from the "Coefficients" box. The ***Y*** intercept (*a*) is reported in the top row of the box as a "constant." The slopes (*b*) are reported in the "Unstandardized Coefficients" column under "B," and the beta-weights (b^*) are reported in the "Standardized Coefficients" column under "Beta."

1. $a =$ ________
2. a. Slope for your first independent variable (b_1) = ________.
   b. Slope for your second independent variable (b_2) = ________.
3. State the least squares multiple regression equation (see Formula 13.3).

$$Y = \_\_\_ + (\_\_)X_1 + (\_\_)X_2$$

4. a. Beta-weight of your first independent variable (b_1^*) = ________.
   b. Beta-weight of your second independent variable (b_2^*) = ________.
5. State the standardized least-square regression line

$$Z_y = (\_\_)Z_1 + (\_\_)Z_2$$

6. $R^2 =$ ________

Step 5: Analyzing and Interpreting Results

Write a short summary of your results. Your summary needs to identify your variables and distinguish between independent and dependent variables. You also need to report the strength and direction of relationships as indicated by the slopes and beta-weights and the total explained variance (R^2). How much of the variance does your first independent variable explain? How much more is explained by your second independent variable?

Project 2: What Are the Correlates of Success?

In this second project, we will continue the investigation into the correlates of success, which we began in Chapter 12.

Step 1: Choosing Independent Variables

Use two of the four independent variables you selected in Project 2 from Chapter 12. All other things being equal, you might choose the variables that had the strongest relationship with the variable you selected to be the dependent variable (either *income06* or *rank*). Remember that independent variables *cannot* be nominal in level of measurement unless you recode the variable into a dummy variable. You may use any interval-ratio or ordinal level variables with more than three categories or scores. Once you have selected your variables, list them in the table below and describe exactly what they measure.

SPSS Variable Name	What Exactly Does This Variable Measure?

Step 2: Stating Hypotheses

Restate your hypotheses from Chapter 12 and add some ideas about the relative importance of your two independent variables. Which do you expect to have the stronger effect? Why?

Step 3: Running the Regression Procedure

Click **Analyze → Regression → Linear** and place your dependent variable in the "Dependent:" box and your selected independent variables in the "Independent(s):" box. Click **Statistics** and **Descriptives**, and then click **OK** to get your results.

Step 4: Recording Results

Summarize your results by filling in the blanks below with information from the "Coefficients" box. The ***Y*** intercept (*a*) is reported in the top row of the box as a "constant." The slopes are reported in the "Unstandardized Coefficients" column under "B," and the beta-weights are reported in the "Standardized Coefficients" column under "Beta."

1. a = ________
2. a. Slope for your first independent variable (b_1) = ________.
   b. Slope for your second independent variable (b_2) = ________.
3. State the least squares multiple regression equation (see Formula 13.3).

$$Y = \_\_\_ + (\_\_)X_1 + (\_\_)X_2$$

4. a. Beta-weight of your first independent variable (b_1^*) = ________.
   b. Beta-weight of your second independent variable (b_2^*) = ________.
5. State the standardized least-square regression line

$$Z_y = (\_\_)Z_1 + (\_\_)Z_2$$

6. R^2 = ________

Step 5: Analyzing and Interpreting Results

Write a short summary of your results. Your summary needs to identify your variables and distinguish between independent and dependent variables. You also need to report the strength and direction of relationships as indicated by the slopes and beta-weights and the total explained variance (R^2). How much of the variance does your first independent variable explain? How much more is explained by your second independent variable?

Appendix A Area Under the Normal Curve

Column (a) lists *Z* scores from 0.00 to 4.00. Only positive scores are displayed, but, since the normal curve is symmetrical, the areas for negative scores will be exactly the same as areas for positive scores. Column (b) lists the proportion of the total area between the *Z* score and the mean. Figure A.1 displays areas of this type. Column (c) lists the proportion of the area beyond the *Z* score, and Figure A.2 displays this type of area.

FIGURE A.1 Area Between Mean and *Z*

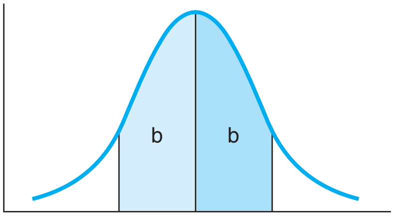
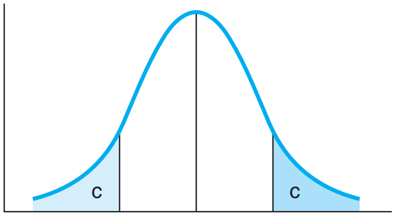


FIGURE A.2 Area Beyond *Z*



(a) *Z*	(b) Area Between Mean and *Z*	(c) Area Beyond *Z*
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
0.04	0.0160	0.4840
0.05	0.0199	0.4801
0.06	0.0239	0.4761
0.07	0.0279	0.4721
0.08	0.0319	0.4681
0.09	0.0359	0.4641
0.10	0.0398	0.4602
0.11	0.0438	0.4562
0.12	0.0478	0.4522
0.13	0.0517	0.4483
0.14	0.0557	0.4443
0.15	0.0596	0.4404
0.16	0.0636	0.4364
0.17	0.0675	0.4325
0.18	0.0714	0.4286
0.19	0.0753	0.4247
0.20	0.0793	0.4207
0.21	0.0832	0.4168
0.22	0.0871	0.4129
0.23	0.0910	0.4090
0.24	0.0948	0.4052
0.25	0.0987	0.4013
0.26	0.1026	0.3974
0.27	0.1064	0.3936
0.28	0.1103	0.3897
0.29	0.1141	0.3859
0.30	0.1179	0.3821
0.31	0.1217	0.3783
0.32	0.1255	0.3745
0.33	0.1293	0.3707
0.34	0.1331	0.3669
0.35	0.1368	0.3632
0.36	0.1406	0.3594
0.37	0.1443	0.3557
0.38	0.1480	0.3520
0.39	0.1517	0.3483
0.40	0.1554	0.3446

(a) Z	(b) Area Between Mean and Z	(c) Area Beyond Z	(a) Z	(b) Area Between Mean and Z	(c) Area Beyond Z
0.41	0.1591	0.3409	0.98	0.3365	0.1635
0.42	0.1628	0.3372	0.99	0.3389	0.1611
0.43	0.1664	0.3336	1.00	0.3413	0.1587
0.44	0.1700	0.3300	1.01	0.3438	0.1562
0.45	0.1736	0.3264	1.02	0.3461	0.1539
0.46	0.1772	0.3228	1.03	0.3485	0.1515
0.47	0.1808	0.3192	1.04	0.3508	0.1492
0.48	0.1844	0.3156	1.05	0.3531	0.1469
0.49	0.1879	0.3121	1.06	0.3554	0.1446
0.50	0.1915	0.3085	1.07	0.3577	0.1423
0.51	0.1950	0.3050	1.08	0.3599	0.1401
0.52	0.1985	0.3015	1.09	0.3621	0.1379
0.53	0.2019	0.2981	1.10	0.3643	0.1357
0.54	0.2054	0.2946	1.11	0.3665	0.1335
0.55	0.2088	0.2912	1.12	0.3686	0.1314
0.56	0.2123	0.2877	1.13	0.3708	0.1292
0.57	0.2157	0.2843	1.14	0.3729	0.1271
0.58	0.2190	0.2810	1.15	0.3749	0.1251
0.59	0.2224	0.2776	1.16	0.3770	0.1230
0.60	0.2257	0.2743	1.17	0.3790	0.1210
0.61	0.2291	0.2709	1.18	0.3810	0.1190
0.62	0.2324	0.2676	1.19	0.3830	0.1170
0.63	0.2357	0.2643	1.20	0.3849	0.1151
0.64	0.2389	0.2611	1.21	0.3869	0.1131
0.65	0.2422	0.2578	1.22	0.3888	0.1112
0.66	0.2454	0.2546	1.23	0.3907	0.1093
0.67	0.2486	0.2514	1.24	0.3925	0.1075
0.68	0.2517	0.2483	1.25	0.3944	0.1056
0.69	0.2549	0.2451	1.26	0.3962	0.1038
0.70	0.2580	0.2420	1.27	0.3980	0.1020
0.71	0.2611	0.2389	1.28	0.3997	0.1003
0.72	0.2642	0.2358	1.29	0.4015	0.0985
0.73	0.2673	0.2327	1.30	0.4032	0.0968
0.74	0.2703	0.2297	1.31	0.4049	0.0951
0.75	0.2734	0.2266	1.32	0.4066	0.0934
0.76	0.2764	0.2236	1.33	0.4082	0.0918
0.77	0.2794	0.2206	1.34	0.4099	0.0901
0.78	0.2823	0.2177	1.35	0.4115	0.0885
0.79	0.2852	0.2148	1.36	0.4131	0.0869
0.80	0.2881	0.2119	1.37	0.4147	0.0853
0.81	0.2910	0.2090	1.38	0.4162	0.0838
0.82	0.2939	0.2061	1.39	0.4177	0.0823
0.83	0.2967	0.2033	1.40	0.4192	0.0808
0.84	0.2995	0.2005	1.41	0.4207	0.0793
0.85	0.3023	0.1977	1.42	0.4222	0.0778
0.86	0.3051	0.1949	1.43	0.4236	0.0764
0.87	0.3078	0.1922	1.44	0.4251	0.0749
0.88	0.3106	0.1894	1.45	0.4265	0.0735
0.89	0.3133	0.1867	1.46	0.4279	0.0721
0.90	0.3159	0.1841	1.47	0.4292	0.0708
0.91	0.3186	0.1814	1.48	0.4306	0.0694
0.92	0.3212	0.1788	1.49	0.4319	0.0681
0.93	0.3238	0.1762	1.50	0.4332	0.0668
0.94	0.3264	0.1736	1.51	0.4345	0.0655
0.95	0.3289	0.1711	1.52	0.4357	0.0643
0.96	0.3315	0.1685	1.53	0.4370	0.0630
0.97	0.3340	0.1660	1.54	0.4382	0.0618

(a) Z	(b) Area Between Mean and Z	(c) Area Beyond Z
1.55	0.4394	0.0606
1.56	0.4406	0.0594
1.57	0.4418	0.0582
1.58	0.4429	0.0571
1.59	0.4441	0.0559
1.60	0.4452	0.0548
1.61	0.4463	0.0537
1.62	0.4474	0.0526
1.63	0.4484	0.0516
1.64	0.4495	0.0505
1.65	0.4505	0.0495
1.66	0.4515	0.0485
1.67	0.4525	0.0475
1.68	0.4535	0.0465
1.69	0.4545	0.0455
1.70	0.4554	0.0446
1.71	0.4564	0.0436
1.72	0.4573	0.0427
1.73	0.4582	0.0418
1.74	0.4591	0.0409
1.75	0.4599	0.0401
1.76	0.4608	0.0392
1.77	0.4616	0.0384
1.78	0.4625	0.0375
1.79	0.4633	0.0367
1.80	0.4641	0.0359
1.81	0.4649	0.0351
1.82	0.4656	0.0344
1.83	0.4664	0.0336
1.84	0.4671	0.0329
1.85	0.4678	0.0322
1.86	0.4686	0.0314
1.87	0.4693	0.0307
1.88	0.4699	0.0301
1.89	0.4706	0.0294
1.90	0.4713	0.0287
1.91	0.4719	0.0281
1.92	0.4726	0.0274
1.93	0.4732	0.0268
1.94	0.4738	0.0262
1.95	0.4744	0.0256
1.96	0.4750	0.0250
1.97	0.4756	0.0244
1.98	0.4761	0.0239
1.99	0.4767	0.0233
2.00	0.4772	0.0228
2.01	0.4778	0.0222
2.02	0.4783	0.0217
2.03	0.4788	0.0212
2.04	0.4793	0.0207
2.05	0.4798	0.0202
2.06	0.4803	0.0197
2.07	0.4808	0.0192
2.08	0.4812	0.0188
2.09	0.4817	0.0183
2.10	0.4821	0.0179
2.11	0.4826	0.0174
2.12	0.4830	0.0170
2.13	0.4834	0.0166
2.14	0.4838	0.0162
2.15	0.4842	0.0158
2.16	0.4846	0.0154
2.17	0.4850	0.0150
2.18	0.4854	0.0146
2.19	0.4857	0.0143
2.20	0.4861	0.0139
2.21	0.4864	0.0136
2.22	0.4868	0.0132
2.23	0.4871	0.0129
2.24	0.4875	0.0125
2.25	0.4878	0.0122
2.26	0.4881	0.0119
2.27	0.4884	0.0116
2.28	0.4887	0.0113
2.29	0.4890	0.0110
2.30	0.4893	0.0107
2.31	0.4896	0.0104
2.32	0.4898	0.0102
2.33	0.4901	0.0099
2.34	0.4904	0.0096
2.35	0.4906	0.0094
2.36	0.4909	0.0091
2.37	0.4911	0.0089
2.38	0.4913	0.0087
2.39	0.4916	0.0084
2.40	0.4918	0.0082
2.41	0.4920	0.0080
2.42	0.4922	0.0078
2.43	0.4925	0.0075
2.44	0.4927	0.0073
2.45	0.4929	0.0071
2.46	0.4931	0.0069
2.47	0.4932	0.0068
2.48	0.4934	0.0066
2.49	0.4936	0.0064
2.50	0.4938	0.0062
2.51	0.4940	0.0060
2.52	0.4941	0.0059
2.53	0.4943	0.0057
2.54	0.4945	0.0055
2.55	0.4946	0.0054
2.56	0.4948	0.0052
2.57	0.4949	0.0051
2.58	0.4951	0.0049
2.59	0.4952	0.0048
2.60	0.4953	0.0047
2.61	0.4955	0.0045
2.62	0.4956	0.0044
2.63	0.4957	0.0043
2.64	0.4959	0.0041
2.65	0.4960	0.0040
2.66	0.4961	0.0039

(a) Z	(b) Area Between Mean and Z	(c) Area Beyond Z
2.67	0.4962	0.0038
2.68	0.4963	0.0037
2.69	0.4964	0.0036
2.70	0.4965	0.0035
2.71	0.4966	0.0034
2.72	0.4967	0.0033
2.73	0.4968	0.0032
2.74	0.4969	0.0031
2.75	0.4970	0.0030
2.76	0.4971	0.0029
2.77	0.4972	0.0028
2.78	0.4973	0.0027
2.79	0.4974	0.0026
2.80	0.4974	0.0026
2.81	0.4975	0.0025
2.82	0.4976	0.0024
2.83	0.4977	0.0023
2.84	0.4977	0.0023
2.85	0.4978	0.0022
2.86	0.4979	0.0021
2.87	0.4979	0.0021
2.88	0.4980	0.0020
2.89	0.4981	0.0019
2.90	0.4981	0.0019
2.91	0.4982	0.0018
2.92	0.4982	0.0018
2.93	0.4983	0.0017
2.94	0.4984	0.0016
2.95	0.4984	0.0016
2.96	0.4985	0.0015
2.97	0.4985	0.0015
2.98	0.4986	0.0014
2.99	0.4986	0.0014
3.00	0.4986	0.0014
3.01	0.4987	0.0013
3.02	0.4987	0.0013
3.03	0.4988	0.0012
3.04	0.4988	0.0012
3.05	0.4989	0.0011
3.06	0.4989	0.0011
3.07	0.4989	0.0011
3.08	0.4990	0.0010
3.09	0.4990	0.0010
3.10	0.4990	0.0010
3.11	0.4991	0.0009
3.12	0.4991	0.0009

(a) Z	(b) Area Between Mean and Z	(c) Area Beyond Z
3.13	0.4991	0.0009
3.14	0.4992	0.0008
3.15	0.4992	0.0008
3.16	0.4992	0.0008
3.17	0.4992	0.0008
3.18	0.4993	0.0007
3.19	0.4993	0.0007
3.20	0.4993	0.0007
3.21	0.4993	0.0007
3.22	0.4994	0.0006
3.23	0.4994	0.0006
3.24	0.4994	0.0006
3.25	0.4994	0.0006
3.26	0.4994	0.0006
3.27	0.4995	0.0005
3.28	0.4995	0.0005
3.29	0.4995	0.0005
3.30	0.4995	0.0005
3.31	0.4995	0.0005
3.32	0.4995	0.0005
3.33	0.4996	0.0004
3.34	0.4996	0.0004
3.35	0.4996	0.0004
3.36	0.4996	0.0004
3.37	0.4996	0.0004
3.38	0.4996	0.0004
3.39	0.4997	0.0003
3.40	0.4997	0.0003
3.41	0.4997	0.0003
3.42	0.4997	0.0003
3.43	0.4997	0.0003
3.44	0.4997	0.0003
3.45	0.4997	0.0003
3.46	0.4997	0.0003
3.47	0.4997	0.0003
3.48	0.4997	0.0003
3.49	0.4998	0.0002
3.50	0.4998	0.0002
3.60	0.4998	0.0002
3.70	0.4999	0.0001
3.80	0.4999	0.0001
3.90	0.4999	<0.0001
4.00	0.4999	<0.0001

Appendix B

Distribution of *t*

Use this table to find the critical region (step 3 of the five-step model) for tests of significance with sample means when sample size (*N*) is small. First, choose an alpha level and a one- or two-tailed test. Then find degrees of freedom to find the *t* score that marks the beginning of the critical region.

Degrees of Freedom (*df*)	Level of Significance for One-Tailed Test					
	0.10	0.05	0.025	0.01	0.005	0.0005
	Level of Significance for Two-Tailed Test					
	0.20	0.10	0.05	0.02	0.01	0.001
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	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
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 Ltd., London (1974), 6th edition (previously published by Oliver & Boyd Ltd., Edinburgh).

Appendix C

Distribution of Chi Square

Use this table to find the critical region (step 3 of the five-step model) for tests of significance with chi square. Choose an alpha level and find the degrees of freedom to find the chi square score that marks the beginning of the critical region.

df	0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.000157	0.000628	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341	16.268
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.465
5	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.517
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566	45.315
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932	46.797
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314	52.620
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642	54.052
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963	55.476
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278	56.893
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588	58.302
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892	59.703

Source: Table IV of Fisher and Yates: *Statistical Tables for Biological, Agricultural and Medical Research*, published by Longman Group Ltd. London (1974), 6th edition (previously published by Oliver & Boyd Ltd., Edinburgh). Reprinted by permission of Addison Wesley Longman Ltd.

Appendix D

Distribution of *F*

Use this table to find the critical region (step 3 of the five-step model) for analysis of variance tests. Choose an alpha level of either 0.05 or 0.01 and then find the degrees of freedom to find the *F* ratio that marks the beginning of the critical region.

$p = 0.05$

$n_1 \rightarrow$ / $n_2 \downarrow$	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

Values of n_1 and n_2 represent the degrees of freedom associated with the between and within estimates of variance, respectively.

Source: From Table V of Fisher & Yates: *Statistical Tables for Biological, Agricultural, and Medical Research*, published by Longman Group, Ltd., London (1974), 6th edition (previously published by Oliver & Boyd, Ltd., Edinburgh).

$p = 0.01$

$n_1 \rightarrow$ / $n_2 \downarrow$	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366
2	98.49	99.01	99.17	99.25	99.30	99.33	99.36	99.42	99.46	99.50
3	34.12	30.81	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.27	9.89	9.47	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65
18	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.94	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

Values of n_1 and n_2 represent the degrees of freedom associated with the between and within estimates of variance, respectively.

Appendix E Using Statistics: Ideas for Research Projects

This appendix presents outlines for four research projects, each of which requires you to use SPSS to analyze the 2012 General Social Survey (GSS), the database that has been used throughout this text. The research projects should be completed at various intervals during the course, and each project permits a great deal of choice on the part of the student.

The first project stresses description and should be done after completing Chapters 2–4. The second involves estimation and should be completed in conjunction with Chapter 6. The third project uses inferential statistics and should be done after completing Part II, and the fourth project combines inferential statistics with measures of association (with an option for multivariate analysis) and should be done after Part III or Part IV.

Project 1: Descriptive Statistics

1. Select five variables from the 2012 General Social Survey *other than* those you have used for end-of-chapter exercises (*NOTE: Your instructor may specify a different number of variables*), and use the **Frequencies** command to get frequency distributions and summary statistics for each variable. Click the **Statistics** button on the **Frequencies** command window and request the mean, median, mode, standard deviation, and range. Make a note of all relevant information when it appears on screen, or print a hard copy. See Appendix G for a list of variables available in the 2012 GSS.

2. For each variable, get bar or line charts to summarize the overall shape of the distribution of the variable.

3. Inspect the frequency distributions and graphs, and choose *appropriate* measures of central tendency and, for ordinal- and interval-ratio-level variables, dispersion. Also, for interval-ratio and ordinal variables with many scores, check for skew both by using the line chart and by comparing the mean and median. Write a sentence or two of description for each variable, being careful to include a description of the overall shape of the distribution (see Chapter 2), the central tendency (Chapter 3), and the dispersion (Chapter 4). For nominal- and ordinal-level variables, be sure to explain any arbitrary numerical codes. For example, on the variable *class* in the 2012 GSS (see Appendix G), a 1 is coded as "lower class," a 2 indicates "working class," and so forth. This is an ordinal-level variable, so you might choose to report the median as a measure of central tendency. If the median score on *class* were 2.45, for example, you

might place that value in context by reporting that "the median is 2.45, about halfway between 'working class' and 'middle class.'"

Here are examples of minimal summary sentences for your report, using *fictitious* data:

- For a nominal-level variable (e.g., gender), report the mode and some detail about the overall distribution. For example: "Most respondents were married (57.5%), but divorced (17.4%) and single (21.3%) individuals were also common."
- For an ordinal-level variable (e.g., occupational prestige), use the median (and perhaps the mean and mode) and the range. For example: "The median number of times that respondents read the newspaper was about once a week (a score of 3) and scores ranged from 1 ('Every day') to 5 ('Never'). The most common score was 2 ('A few times a week') and the average score was 2.13."
- For an interval-ratio-level variable (e.g., age), use the mean (and perhaps the median or mode) and the standard deviation (and perhaps the range). For example: "Average age for this sample was 42.3. Respondents ranged from 18 to 94 years of age with a standard deviation of 15.37."

Project 2: Estimation

In this exercise, you will use the 2012 GSS sample to estimate characteristics of the U.S. population. You will use SPSS to generate the sample statistics and then use either Formula 6.2 or Formula 6.3 to find the confidence interval and state each interval in words.

Estimating Means

1. There are relatively few interval-ratio variables in the 2012 GSS, and for this part of the project you may use ordinal variables that have at least three categories or scores. Choose a total of three variables that fit this description *other than* the variables you used in Chapter 6. (*NOTE: Your instructor may specify a different number of variables*.)
2. Use the **Descriptives** command to get means, standard deviations, and sample size (*N*), and use this information to construct 95% confidence intervals for the mean of each of your variables. Make a note of the mean, standard deviation, and sample size, or print a hard copy. Use Formula 6.2 to compute the confidence intervals.
3. For each variable, write a summary sentence reporting the variable, the interval itself, the confidence level, and the sample size. Write in plain English, as if you were reporting results in a newspaper. Most importantly, you should make it clear that you are estimating characteristics of the population of the entire United States. For example, a summary sentence might look like this: "Based on a random sample of 1231 respondents, I estimate at the 95% level

that U.S. drivers average between 64.46 and 68.22 miles per hour when driving on interstate highways."

Estimating Proportions

1. Choose three nominal- or ordinal-level variables *other than* the variables you used in Chapter 6. (*NOTE: Your instructor may specify a different number of variables*.)
2. Use the **Frequencies** command to get the percentage of the sample in the various categories of each variable. Change the percentages (remember to use the "valid percents" column) to proportions and construct confidence intervals for one category of each variable (e.g., the percent female for *sex*) using Formula 6.3.
3. For each variable, write a summary sentence reporting the variable, the interval, the confidence level, and the sample size. Write in plain English, as if you were reporting results in a newspaper. Remember to make it clear that you are estimating a characteristic of the U.S. population.
4. For any one of the intervals you constructed in either Project A or Project B, identify each of the following concepts and terms, and briefly explain their role in estimation: sample, population, statistic, parameter, EPSEM, representative, confidence level.

Project 3: Significance Testing

In this exercise, you will use the 2012 GSS sample to conduct tests of significance. You will use the five-step model to conduct the tests.

Two-Sample *t* Test (Chapter 8)

1. Choose two different dependent variables from the interval-ratio or ordinal-level variables that have three or more scores. (*NOTE: Your instructor may specify a different number of variables*.) Choose independent variables that might logically be a cause of your dependent variables. Remember that, for a *t* test, independent variables can have *only* two categories, but you can still use independent variables with more than two categories by (a) using the "Grouping Variable" box to specify the exact categories (e.g., select scores of 1 and 5 on *marital* to compare married with never-married respondents), or (b) collapsing the scores of variables with more than two categories by using the **Recode** command. Independent variables can be any level of measurement, and you may use the same independent variable for both tests. (See "You Are the Researcher" in Chapter 9 for instructions on how to recode variables.)
2. Start SPSS, load the 2012 GSS dataset, and click **Analyze**, **Compare Means**, and then **Independent Samples T Test**. Name your dependent variable(s) in the "Test Variable" window and your independent variable in the "Grouping Variable" window. You will also need to specify the scores used to define the groups on the independent variable. See Chapter 8 for examples. Make a note of the test results (group means, obtained *t* score, significance, sample size) or print a hard copy. Repeat the procedure for the second dependent variable.

3. Write up the results of the test. At a minimum, your report should clearly identify the independent and dependent variables, the sample statistics, the value of the test statistic (step 4), the results of the test (step 5), and the alpha level you used.

Analysis of Variance (Chapter 9)

1. Choose two different dependent variables from the interval-ratio or ordinal-level variables that have three or more scores. (*NOTE: Your instructor may specify a different number of variables*.) Choose independent variables that might logically be a cause of your dependent variables and that have between three and five categories. You may use the same independent variables for both tests.

2. If necessary, start SPSS, load the 2012 GSS dataset, and click **Analyze**, **Compare Means**, and then **One-way Anova**. The "One-way Anova" window will appear. Find your dependent variable in the variable list on the left and click the arrow to move the variable name into the **Dependent List**: box. Note that you can request more than one dependent variable at a time. Next, find the name of your independent variable and move it to the **Factor**: box. Click **Options** and then click the box next to **Descriptive** in the **Statistics** box to request means and standard deviations. Click **Continue** and **OK**. Make a note of the test results or print a hard copy. Repeat, if necessary, for your second dependent variable.

3. Write up the results of the test. At a minimum, your report should clearly identify the independent and dependent variables, the sample statistics (category means), the value of the test statistic (step 4), the results of the test (step 5), the degrees of freedom, and the alpha level you used.

Chi Square (Chapter 10)

1. Choose two different dependent variables of any level of measurement that have five or fewer (preferably two to three) scores. (*NOTE: Your instructor may specify a different number of variables*.) For each dependent variable, choose an independent variable that might logically be a cause. Independent variables can be any level of measurement as long as they have five or fewer (preferably two to three) categories. Output will be easier to analyze if you use variables with few categories. You may use the same independent variable for both tests.

2. Click **Analyze**, **Descriptive Statistics**, and then **Crosstabs**. The "Crosstabs" dialog box will appear. Highlight your first dependent variable and move it to the "Rows" box. Next, highlight your independent variable and move it to the "Columns" box. Click the **Statistics** button and click the box next to **chi square**. Click **Continue** and **OK**. Make a note of the results or print a hard copy. Repeat for your second dependent variable.

3. Write up the results of the test. At a minimum, your report should clearly identify the independent and dependent variables, the value of the test statistic (step 4), the results of the test (step 5), the degrees of freedom, and the alpha level you used. It is almost always desirable to report the column percentages as well.

Project 4: Analyzing the Strength and Significance of Relationships

Using Bivariate Tables

1. From the 2012 GSS dataset, select either

   - One dependent variable and three independent variables (possible causes)

   OR

   - One independent variable and three possible dependent variables (possible effects)

   Variables can be from any level of measurement but must have only a few (two to five) categories or scores. (*NOTE: Your instructor may specify a different number of variables*.)

2. Develop research questions or hypotheses about the relationships between variables. Make sure the causal links you suggest are sensible and logical.

3. Use **Crosstabs** to generate bivariate tables. See Chapters 10 and 11 for examples. Place your dependent variable(s) in the rows and independent variable(s) in the columns of the crosstabs tables. In the "Crosstabs" dialog box, click the **Statistics** button and choose **chi square**, **phi** or **Cramer's *V***, and **gamma** for every table you request. Also, click the **Cells** button and get column percentages for every table you request. Make a note of results as they appear on the screen or print hard copies.

4. Write a report that presents and analyzes these relationships. Be clear about which variables are dependent and which are independent. For each combination of variables, report the test of significance and an appropriate measure of association. In addition, for each relationship, report and discuss column percentages, pattern or direction of the relationship, and strength of the relationship.

Using Interval-Ratio Variables

1. From the 2012 GSS, select either

   - One dependent variable and three independent variables (possible causes)

   OR

   - One independent variable and three dependent variables (possible effects)

   Variables should be interval-ratio in level of measurement, but you may use ordinal-level variables as long as they have more than three scores. (*NOTE: Your instructor may specify a different number of variables*.)

2. Develop research questions or hypotheses about the relationships between variables. Make sure the causal links you suggest are sensible and logical.

3. Use **Regression** and **Scatterplot** (click **Graphs**, **Legacy Dialogs**, and then **Scatter**) to analyze the bivariate relationships. Make a note of results (including r, r^2, slopes, beta-weights, and a) as they appear on the screen, or print hard copies.

4. Write a report that presents and analyzes these relationships. Be clear about which variables are dependent and which are independent. For each combination of variables, report the strength and direction of the relationship. Include r, r^2, and the beta-weights in your report.

Optional Multivariate Analysis. Pick one of the bivariate relationships you produced in step 3 and find another logical independent variable. Run **Regression** again with both independent variables and analyze the results. How much improvement is there in the explained variance after the second independent variable is included? Write up the results of this analysis and include them in your summary paper for this project.

Appendix F

An Introduction to SPSS

Computers have affected virtually every aspect of human society, and, as you would expect, their impact on the conduct of social research has been profound. Researchers routinely use computers to organize data and compute statistics—activities that humans often find dull, tedious, and difficult but that a computer can accomplish with accuracy and ease. This division of labor allows social scientists to spend more time on analysis and interpretation—activities that humans typically enjoy but that are beyond the power of computers (so far, at least).

These days, the skills needed to use computers successfully are quite accessible, even for people with little or no experience. This appendix will prepare you to use a statistics program called SPSS—the **S**tatistical **P**ackage for the **S**ocial **S**ciences. If you have used a computer mouse to "point and click," you are ready to use SPSS. Even if you are completely unfamiliar with computers, you will find this program accessible. After you finish this appendix, you will be ready to do the exercises found at the end of most chapters of this text.

A word of caution before we begin: This appendix is intended only as an *introduction* to SPSS. It will give you an overview of the program and enough information so that you can complete the assignments in the text. It is unlikely, however, that this appendix will answer all your questions or provide solutions to all the problems you might encounter. So this is a good place to tell you that SPSS has an extensive and easy-to-use **Help** facility that will provide assistance as you request it. You should familiarize yourself with this feature and use it as needed. To get help, simply click on the **Help** command on the toolbar across the top of the SPSS screen.

SPSS is a **statistical package** (or **statpak**): a set of computer programs that work with data and compute statistics as requested by the user (you). Once you have entered the data for a particular group of observations, you can easily and quickly produce an abundance of statistical information without doing any computations or writing any computer programs yourself.

Why bother to learn this technology? The truth is that the labor saving capacity of computers is sometimes exaggerated, and there are research situations in which they are unnecessary. If you are working with a small number of observations or need only a few uncomplicated statistics, then statistical packages may not be helpful. However, as the number of cases increases and as your requirements for statistics become more sophisticated, computers and statpaks will become more and more useful.

An example should make this point clearer. Suppose you have gathered a sample of 150 respondents and the only thing you want to know about these people is their average age. To compute an average, as you know, you add the scores and divide by the number of cases. How long do you think it would take you to add 150 two-digit numbers (ages) with a hand calculator? If you entered the scores at the rate of one per second—60 scores per minute—it would take about 3 or 4 minutes to enter the ages and get the average. Even if you worked slowly and carefully and added the scores a second and third time to check your math, you could probably complete all calculations in less than 15 or 20 minutes. If this were all the information you needed, computers and statpaks would not save you any time.

Such a simple research project is not realistic, however. Typically, social science researchers deal with dozens or even hundreds of variables and hundreds or thousands of cases. While you could add 150 numbers in perhaps 3 or 4 minutes, how long would it take to add the scores for 1500 cases? What are the chances of adding 1500 numbers without making significant errors of arithmetic? The more complex the research situation, the more valuable and useful statpaks become. SPSS can produce statistical information in a few keystrokes or clicks of the mouse that might take you minutes, hours, or even days to produce with a hand calculator.

Clearly, this is technology worth mastering by any social researcher. With SPSS, you can avoid the drudgery of mere computation, spend more time on analysis and interpretation, and conduct research projects with very large data sets. Mastery of this technology might be very handy indeed in your senior-level courses, in a wide variety of jobs, or in graduate school.

Getting Started—Databases and Computer Files

Before statistics can be calculated, SPSS must have some data to process. A **data base** is an organized collection of related information, such as the responses to a survey. For purposes of computer analysis, a database is organized into a **file**—a collection of information that is stored under the same name in the memory of the computer, on a disk or flash drive, or on some other medium. Words as well as numbers can be saved in files. If you've ever used a word-processing program to type a letter or term paper, you probably saved your work in a file so that you could update or make corrections at a later time. Data can be stored in files indefinitely. Because it can take months to conduct a thorough data analysis, the ability to save a database is another advantage of using computers.

For the SPSS exercises in this text, we will rely primarily on the *GSS2012*, a shortened version of the General Social Survey (GSS) for 2012. The complete General Social Survey includes hundreds of items of information (age, gender, opinion about such social issues as capital punishment, and so forth) for thousands of respondents. Some of you will be using a student version of SPSS, which is limited in the number of cases and variables it can process. To accommodate those limits, I have reduced the database to 49 items of information and fewer than 1500 respondents.

The GSS is an especially valuable resource for (at least) two reasons. First, the respondents to the survey are chosen so that the sample is representative of the entire U.S. population. A representative sample reproduces, in miniature form, the characteristics of the population from which it was taken (see Chapter 6). So when you analyze the *GSS2012*, you are in effect analyzing U.S. society as of 2012. The data are real, and the relationships you will analyze reflect some of the most important and sensitive issues in American life.

Second, the GSS has been administered regularly since 1972 and provides a continuous record of how American society has evolved and changed over the last four decades on issues ranging from abortion to euthanasia. The GSS has been the basis for hundreds of research projects by professional social researchers.

Although we will rely primarily on the *GSS2012*, we will also use three other databases. The first two listed below are used throughout the text but the third is used only to construct graphs in Chapter 2.

- The *States* database includes a variety of information for the 50 states.
- The *Intl-POP* database presents mostly population data (e.g., birth rates and growth rates) for a group of about 100 nations.
- The *CrimeTrends84-10* database includes five measures of criminal activity in the 50 states.

Appendix G presents codebooks for all four databases. A codebook lists all the variables in the database, explains what each variable measures, and, where necessary, lists the meaning of the scores for each variable. To illustrate, please examine the codebook for the *GSS2012*. The codebook lists all variables alphabetically by their variable names, along with the survey question used to generate the variable.

Note that the variable names in the left column are abbreviations of the related survey item. In some cases (e.g., *age*), the variable names are easy to figure out. In other cases, the variable names (like *abany* and *fefam*) are not so obvious. Previous versions of SPSS required that variable names be no more than eight characters and this sometimes necessitated extreme abbreviation. Although SPSS no longer restricts variable names to eight characters (as reflected in the other databases), we retain these short names for ease of comparability with older versions of the GSS.

Appendix G also shows the wording of the item that generated the variable. For example, *abany* consists of responses to a question about legal abortion and the variable name is formed from the question: “Should an *ab*ortion be possible for *any* reason?” The *fefam* variable consists of responses to the statement “It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family.”

Where necessary, Appendix G lists the codes (or scores) of a variable along with their meanings. Thus, a score of 1 on *abany* means that the respondent approves of a legal abortion “for any reason” and a score of 0 on *attend* means that the respondent never attends religious services.

Notice that some of the possible responses to *abany*, *fefam*, and other variables in Appendix G are labeled IAP (“Not Applicable”), NA, or DK. The first of these responses means that the item in question was not given to a respondent. The full GSS is very long, and, to keep the time frame for completing the survey reasonable, not all respondents are asked every question. NA stands for “No Answer” and means that the respondent was asked the question but refused to

FIGURE F.1 The Data Analysis Process

Database	→	Statpak	→	Output	→	Analysis
(raw information)		(computer programs)		(statistics and graphs)		(interpretation)

answer. DK stands for "Don't Know," which means that the respondent did not have the requested information. All three of these scores are "missing values," and, as "noninformation," they should be eliminated from statistical analysis. Missing values are common on surveys, and, as long as they are not too numerous, they are not a particular problem.

It is important that you understand the difference between a statpak (SPSS) and a database (*GSS2012, States,* and *Intl-POP*) and what we are ultimately after here. A database consists of information. A statpak processes the information in the database and produces statistics. Our goal is to apply the statpak to the database to produce output (for example, statistics and graphs) that we can analyze and use to answer questions. The process might be diagrammed as in Figure F.1.

Statpaks like SPSS are general research tools that can be used to analyze databases of all sorts; they are not limited just to the databases supplied with this text. In the same way, our databases could be analyzed with statpaks other than SPSS, including SAS and Stata, that may be available on your campus.

Starting SPSS and Loading a Database

If you are using the complete, professional version of SPSS, you will probably be working in a computer lab, and you can begin running the program immediately. If you are using the student version of the program on your personal computer, the first thing you need to do is install the software. Follow the instructions that came with the program and return to this appendix when the installation is complete.

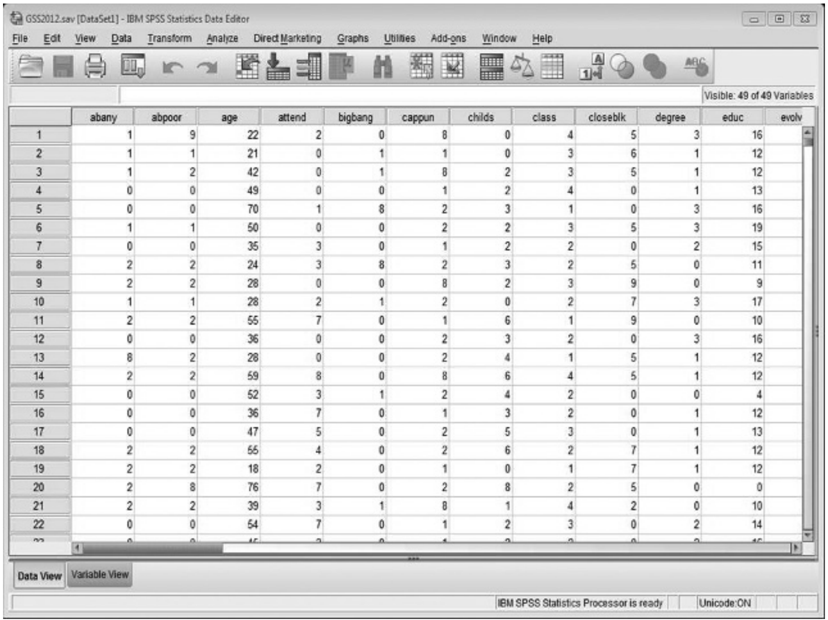
To start SPSS, find the icon (or picture) on the screen of your monitor that has an "IBM SPSS Statistics" label attached to it. Move the cursor on the monitor screen over this icon and then double-click the left button on the mouse. This will start SPSS.

After a few seconds, the SPSS screen will appear. Depending on which version of SPSS you are using, you may see an introductory screen that asks "What would you like to do?" or that lists "New and Recent Files" on the left. In either case, close this initial screen by clicking the "x" in the upper right corner of the screen.

The next screen you will see is the "SPSS Data Editor" screen. Note that there is a list of commands across the very top of the screen. These commands begin with **File** at the far left and end with **Help** at the far right. This is the main menu bar for SPSS. When you click any of these words, a menu of commands and choices will drop down. Basically, you tell SPSS what to do by clicking on your desired choices from these menus. Sometimes submenus will appear, and you will need to specify your choices further.

The first thing we need to do is open a dataset. Click the **File** command on the left of the menu bar and a submenu will drop down. Click **Open** and then

Data from the submenu. You need to find the *GSS2012* data file, which may already be stored on the computer, probably under the name *GSS2012.sav* or something similar. If not, you can download the dataset from the website for this text. Check with your instructor to make sure you know how to access the dataset. Once you have located the data, click on the file name and SPSS will load it. After the dataset is loaded, your computer screen will look like this:



SPSS provides the user with a variety of options for displaying information about the data file and output on the screen. I recommend that you tell the program to display lists of variables by name (e.g., *age, abany*) rather than by labels (e.g., AGE OF RESPONDENT, or ABORTION IF WOMAN WANTS FOR ANY REASON). Lists displayed this way will be easier to read and compare to Appendix G. To do this, click **Edit** on the main menu bar and then click **Options** from the drop-down submenu. A dialog box labeled "Options" will appear with a series of tabs along the top. The "General" options should be displayed; if not, click on this tab. On the "General" screen, find the box labeled "Variable Lists" and, if they are not already selected, click **Display names** and **alphabetical** and then click **OK**. If you make changes, a message may appear on the screen that tells you that changes will take effect the next time a data file is opened.

The procedures covered in this section are summarized in Table F.1.

TABLE F.1 Summary of Commands

To start SPSS	Click the SPSS icon.
To open a data file or database	Click **File**, **Open**, and **Data**. Find the file name of the data set you want to open and double-click it.
To set the display options for lists of variables	Click **Edit** and **Options**. On the "General" tab, make sure "Display names" and "Alphabetical" are selected and then click **OK**.

Working with Databases

Note that the "SPSS Data Editor" window is organized into a two-dimensional grid, with columns running up and down (vertically) and rows running across (horizontally). Each column is a variable or item of information from the database. The names of the variables are listed at the tops of the columns. Remember that you can find the meanings of these variable names in the codebooks in Appendix G.

Another way to decipher the meanings of variable names is to click **Utilities** on the menu bar and then click **Variables**. The "Variables" window will open. This window has two parts. On the left is a list of all variables in the database, arranged in alphabetical order and with the first variable highlighted. On the right is the "Variable Information" window, with information about the highlighted variable. The first variable in the *GSS2012* database is listed as *abany*. The "Variable Information" window displays a fragment of the question that was actually asked during the survey ("ABORTION IF WOMAN WANTS FOR ANY REASON") and shows the possible scores on this variable (a score of 1 = yes and a score of 2 = no), along with some other information.

The same information can be displayed for any variable in the dataset. For example, with the *GSS2012* database loaded, find the variable *marital*. You can do this by using the arrow keys on your keyboard or the slider bar on the right of the variable list window. You can also move through the list by typing the first letter of the variable name you are interested in. For example, type "m" and you will be moved to the first variable name in the list that begins with that letter. Now you can see that *marital* measures marital status and that a score of "1" indicates that the respondent was married, and so forth. What do *relig* and *marhomo* measure? Close this window by clicking the **Close** button at the bottom of the window.

Examine the "Data Editor" window displaying the *GSS2012* database a little more. Each row of the window (reading across, from left to right) contains the scores of a particular respondent on all the variables in the database. Note that the upper-left cell is highlighted (outlined in a darker border than the other cells). This cell contains the score of respondent 1 on the first variable. The second row contains the scores of respondent 2, and so forth. You can move around in this window with the arrow keys on your keyboard.

The SPSS commands presented in this section are summarized in Table F.2. We are now prepared to perform some statistical operations with the 2012 GSS database.

TABLE F.2 Summary of Commands

To move around the "Data Editor" window	Use any of these options: 1. Click the cell you want to highlight. 2. Use the arrow keys on your keyboard. 3. Move the slider buttons. 4. Click the arrows on the right-hand or bottom margins.
To get information about a variable	1. See Appendix G. 2. Click **Utilities** and then **Variables**. Scroll through the list until you find the variable in which you are interested. Information about the variable will appear in the window on the right.

Putting SPSS to Work: Producing Statistics

At this point, the database on the screen is just a mass of numbers with little meaning for you. That's okay because you will not actually have to read any information from this screen. All of the statistical operations you will conduct will begin by clicking the **Analyze** command from the menu bar, selecting a procedure and statistics, and then naming the variable or variables you would like to process.

To illustrate, let's have SPSS produce a frequency distribution for the variable *sex*. Frequency distributions are tables that display the number of times each score of a variable occurred (see Chapter 2). So when we complete this procedure, we will know the number of males and females in the 2012 GSS sample.

With the *GSS2012* data set loaded, click **Analyze** on the menu bar. From the menu that drops down, click **Descriptive Statistics** and then **Frequencies**. The "Frequencies" window appears, with the variables listed in alphabetical order in the box on the left. The first variable (*abany*) will be highlighted. Use the slider button or the arrow keys on the right-hand margin of this box to scroll through the variable list until you highlight the variable *sex*, or type "s" to move to the approximate location.

Once the variable name has been highlighted, click the arrow button in the middle of the screen to move the variable name to the box on the right-hand side of the screen. SPSS will produce frequency distributions for all variables listed in this box, but for now we will confine our attention to *sex*. Click the **OK** button in the "Frequencies" window and, in seconds, a frequency distribution will be produced.

SPSS sends all tables and statistics to the Output window, or SPSS viewer. This window is now "closest" to you, and the Data Editor window is "behind" the Output window. If you want to return to the Data Editor, click on any part of it if it is visible; it will move to the "front" and the Output window will be "behind" it. To display the Data Editor window if it is not visible, minimize the Output window by clicking the "_" box in the upper-right corner.

Frequencies. The output from SPSS is reproduced as Table F.3. What can we tell from this table? The score labels (male and female) are printed at the left, with the number of cases (Frequency) in each category of the variable one column to the right. As you can see, there are 666 males and 791 females in the sample. The next two columns give information about percentages, and the last column to the right displays cumulative percentages. We will defer a discussion of this last column until a later exercise.

TABLE F.3 An Example of SPSS Output

RESPONDENTS SEX

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MALE	666	45.7	45.7	45.7
	FEMALE	791	54.3	54.3	100.0
	Total	1457	100.0	100.0	

One of the percentage columns is labeled “Percent” and the other is labeled “Valid Percent.” The difference between these two columns lies in the handling of missing values. The “Percent” column is based on all cases, including people who did not respond to the item (NA) and people who said they did not have the requested information (DK). The “Valid Percent” column excludes all missing scores. Because we will almost always want to ignore missing scores, we will pay attention only to the “Valid Percent” column. Note that for *sex*, there are no missing scores, and the two columns are identical.

Printing and Saving Output

Once you have gone to the trouble of producing statistics, a table, or a graph, you will probably want to keep a permanent record. There are two ways to do this. First, you can print a hard copy of the contents of the Output window to take with you. To do this, click on **File** and then click **Print**. Alternatively, find the icon of a printer (third from the left) in the row of icons just below the menu bar and click on it.

The other way to create a permanent record is to save the output to the computer’s memory, to a flash drive, or to some other device. To do this, click **Save** from the **File** menu. The **Save** dialog box opens. Give the output a name (some abbreviation such as “frequency-sex” might do) and, if necessary, specify the location you are using. Click **OK**, and the table will be permanently saved.

Ending Your SPSS Session

Once you have saved or printed your work, you may end your SPSS session. Click on **File** from the menu bar and then click **Exit**. If you haven’t already done so, you will be asked whether you want to save the contents of the output window. You may save the frequency distribution at this point if you wish. Otherwise, click **No**. The program will close, and you will be returned to the screen from which you began.

Appendix G

Codebooks for All Datasets

Codebook for the General Social Survey, 2012

The General Social Survey (GSS) is a public opinion poll that has been regularly conducted by the National Opinion Research Council since 1972. A version of the 2012 GSS is available at the website for this text and is used in many of the SPSS demonstrations and in end-of-chapter exercises. Our version of the 2012 GSS includes 49 variables for a randomly selected subsample of about 1500 of the original respondents. This codebook lists each item in the dataset. The variable names are those used in the data files. The questions have been reproduced exactly as they were asked (with a few exceptions to conserve space), and the numbers beside each response are the scores recorded in the data file.

The dataset includes variables that measure demographic or background characteristics of the respondents, including gender, age, race, religion, and several indicators of socioeconomic status. Also included are items that measure opinion on current and controversial topics such as abortion, capital punishment, and homosexuality.

Most variables in the dataset have codes for "missing data." These codes are italicized in the listings below for easy identification. The codes refer to various situations in which the respondent does not or cannot answer the question and are excluded from all statistical operations. The codes are: *IAP* (the respondent was not asked the question), *DK* ("Don't know," that is, the respondent didn't have the requested information), and *NA* ("No answer," that is, the respondent refused to answer).

Please tell me if you think it should be possible for a woman to get a legal abortion if . . .

abany — She wants it for any reason.

1. Yes
2. No

0. IAP, 8. DK, 9. NA

abpoor — The family is very low income and can't afford any more children. (Same scoring as *abany*)

age — Age of respondent

18–89. Actual years

99. *NA*

attend — How often do you attend religious services?

0. Never
1. Less than once per year
2. Once or twice a year
3. Several times per year
4. About once a month
5. 2–3 times a month
6. Nearly every week
7. Every week
8. More than once a week
9. *DK or NA*

bigbang — The universe began with a huge explosion. Is that true or false?

1. True
2. False

0. IAP, 8. DK, 9. NA

cappun — Do you favor or oppose the death penalty for persons convicted of murder?

1. Favor
2. Oppose

0. IAP, 8. DK, 9. NA

childs — How many children have you ever had?

1–7. Actual number
8. Eight or more
9. DK, NA

class — Subjective class identification

1. Lower class
2. Working class
3. Middle class
4. Upper class

0. IAP, 8. DK, 9. NA

closeblk — In general, how close do you feel to blacks?

1. Not at all close
5. Neither one nor the other
9. Very close

0. IAP, 98. DK, 99. NA

degree — Respondent's highest degree

0. Less than high school
1. High school
2. Assoc./junior college
3. Bachelor's
4. Graduate

7. IAP, 8. DK, 9. NA

educ — Highest year of school completed

0–20. Actual number of years
97. IAP, 98. DK, 99. NA

evolved — Human beings, as we know them today, developed from earlier species of animals.

1. True
2. False

0. IAP, 8. DK, 9. NA

fear — Is there any area right around here—that is, within a mile—where you would be afraid to walk alone at night?
1. Yes
2. No
0. IAP, 8. DK, 9. NA

fechld — A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.
1. Strongly disagree
2. Disagree
3. Agree
4. Strongly agree
0. IAP, 8. DK, 9. NA

fefam — It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family.
1. Strongly agree
2. Agree
3. Disagree
4. Strongly disagree
0. IAP, 8. DK, 9. NA

getahead — Some people say that people get ahead by their own hard work; others say that lucky breaks or help from other people are more important. Which do you think is most important?
1. Hard work more important
2. Hard work, luck equally important
3. Luck more important
0. IAP, 8. DK, 9. NA

gunlaw — Would you favor or oppose a law which would require a person to obtain a police permit before he or she could buy a gun?
1. Favor
2. Oppose
0. IAP, 8. DK, 9. NA

happy — Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?
1. Very happy
2. Pretty happy
3. Not too happy
0. IAP, 8. DK, 9. NA

hrs1 — How many hours did you work last week?
1–89. Actual hours
−1. IAP, 98. DK, 99. NA

income06 — Respondent's total family income from all sources

1. Less than 1000
2. 1000 to 2999
3. 3000 to 3999
4. 4000 to 4999
5. 5000 to 5999
6. 6000 to 6999
7. 7000 to 7999
8. 8000 to 9999
9. 10,000 to 12,499
10. 12,500 to 14,999
11. 15,000 to 17,499
12. 17,500 to 19,999
13. 20,000 to 22,499
14. 22,500 to 24,999
15. 25,000 to 29,999
16. 30,000 to 34,999
17. 35,000 to 39,999
18. 40,000 to 49,999
19. 50,000 to 59,999
20. 60,000 to 74,999
21. 75,000 to 89,999
22. 90,000 to 109,999
23. 110,000 to 129,999
24. 130,000 to 149,999
25. 150,000 or more

98. DK, 99. NA

letdie1 — When a person has a disease that cannot be cured, do you think doctors should be allowed by law to end the patient's life by some painless means if the patient and his family request it?

1. Yes
2. No

0. IAP, 8. DK, 9. NA

letin1 — Do you think the number of immigrants to America nowadays should be

1. Increased a lot
2. Increased a little
3. Remain the same as it is
4. Reduced a little
5. Reduced a lot

0. IAP, 8. DK, 9. NA

marhomo — Homosexual couples should have the right to marry one another.

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

0. IAP, 8. DK, 9. NA

marital — Are you currently married, widowed, divorced, separated, or have you never been married?

1. Married
2. Widowed
3. Divorced
4. Separated
5. Never married

9. NA

news — How often do you read the newspaper?
1. Every day
2. A few times a week
3. Once a week
4. Less than once a week
5. Never
0. IAP, 8. DK, 9. NA

obey — How important is it for a child to learn obedience to prepare him or her for life?
1. Most important
2. 2nd most important
3. 3rd most important
4. 4th most important
5. Least important
0. IAP, 8. DK, 9. NA

paeduc — Father's highest year of school completed
0–20. Actual number of years
97. IAP, 98. DK, 99. NA

parsol — Compared to your parents when they were the age you are now, do you think your own standard of living now is much better, somewhat better, about the same, somewhat worse, or much worse than theirs was?
1. Much better
2. Somewhat better
3. About the same
4. Somewhat worse
5. Much worse
0. IAP, 8. DK, 9. NA

partnrs5 — How many sex partners have you had over the past five years?
0. No partners
1. 1 partner
2. 2 partners
3. 3 partners
4. 4 partners
5. 5–10 partners
6. 11–20 partners
7. 21–100 partners
8. More than 100 partners
9. 1 or more, don't know the number, 95. Several, 98. DK, 99. NA, −1. IAP

pillok — Do you strongly agree, agree, disagree, or strongly disagree that methods of birth control should be available to teenagers between the ages of 14 and 16 if their parents do not approve?
1. Strongly agree
2. Agree
3. Disagree
4. Strongly disagree
0. IAP, 8. DK, 9. NA

polviews — I'm going to show you a seven-point scale on which the political views that people might hold are arranged from extremely liberal to extremely conservative. Where would you place yourself on this scale?

1. Extremely liberal
2. Liberal
3. Slightly liberal
4. Moderate
5. Slightly conservative
6. Conservative
7. Extremely conservative

0. IAP, 8. DK, 9. NA

postlife — Do you believe there is life after death?

1. Yes
2. No

0. IAP, 8. DK, 9. NA

pray — About how often do you pray?

1. Several times a day
2. Once a day
3. Several times a week
4. Once a week
5. Less than once a week
6. Never

0. IAP, 8. DK, 9. NA

premarsx — There's been a lot of discussion about the way morals and attitudes about sex are changing in this country. If a man and a woman have sex relations before marriage, do you think it is always wrong, almost always wrong, wrong only sometimes, or not wrong at all?

1. Always wrong
2. Almost always wrong
3. Wrong only sometimes
4. Not wrong at all

0. IAP, 8. DK, 9. NA

pres08 — In 2008, did you vote for Obama (the Democratic candidate) or McCain (the Republican candidate)? (Includes only those who said they voted in this election.)

1. Obama
2. McCain

3. Other, 6. No presidential vote, 0. IAP, 8. DK, 9. NA

racecen1 — Race or ethnicity of respondent

1. White
2. Black
3. American Indian or Alaska Native
4. Asian American or Pacific Islander
5. Hispanic

0. IAP, 98. DK

rank — In our society there are groups which tend to be toward the top and those that are toward the bottom. Here we have a scale that runs from top to bottom. Where would you put yourself on this scale?

1. Top	2.
3.	4.
5.	6.
7.	8.
9.	10. Bottom

0. IAP, 98. DK, 99. NA

region — Region of interview

1. New England (ME, VT, NH, MA, CT, RI)
2. Mid-Atlantic (NY, NJ, PA)
3. East North Central (WI, IL, IN, MI, OH)
4. West North Central (MN, IA, MO, ND, SD, NE, KS)
5. South Atlantic (DE, MD, WV, VA, NC, SC, GA, FL, DC)
6. East South Central (KY, TN, AL, MS)
7. West South Central (AK, OK, LA, TX)
8. Mountain (MT, ID, WY, NV, UT, CO, AR, NM)
9. Pacific (WA, OR, CA, AL, HI)

relig — What is your religious preference? Is it Protestant, Catholic, Jewish, some other religion, or no religion?

1. Protestant
2. Catholic
3. Jewish
4. None
5. Other

8. DK, 9. NA

satjob — On the whole, how satisfied are you with the work you do? (*Asked of those currently working or keeping house.*)

1. Very satisfied
2. Moderately satisfied
3. A little dissatisfied
4. Very dissatisfied

0. IAP, 8. DK, 9. NA

sex — Respondent's gender

1. Male
2. Female

sexfreq — About how many times did you have sex during the last 12 months?

0. Not at all
1. Once or twice
2. About once a month
3. 2 or 3 times a month
4. About once a week
5. 2 or 3 times a week
6. More than 3 times a week

−1. IAP, 8. DK, 9. NA

spanking — Do you strongly agree, agree, disagree, or strongly disagree that it is sometimes necessary to discipline a child with a good, hard spanking?

1. Strongly agree
2. Agree
3. Disagree
4. Strongly disagree

0. IAP, 8. DK, 9. NA

spkmslm — There are always some people whose ideas are considered bad or dangerous by other people. Consider a Muslim clergyman who preached hatred of the United States . . . should he be allowed to speak or not?

1. Yes, allowed
2. Not allowed

0. IAP, 8. DK, 9. NA

thnkself — How important is it for a child to learn to think for himself or herself to prepare him or her for life?

1. Most important
2. 2nd most important
3. 3rd most important
4. 4th most important
5. Least important

0. IAP, 8. DK, 9. NA

trust — Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?

1. Can be trusted
2. Can't be trusted

0. IAP, 8. DK, 9. NA

tvhours — On the average day, about how many hours do you personally watch television?

00–24. Actual hours

−1. IAP, 98. DK, 99. NA

wwwhr — Not counting e-mail, about how many hours per week do you use the Web?

0–120. Actual hours

−1. IAP, 998. DK, 999. NA

xmarsex — What is your opinion about a married person having sexual relations with someone other than the marriage partner?

1. Always wrong
2. Almost always wrong
3. Wrong only sometimes
4. Not wrong at all

0. IAP, 8. DK, 9. NA

Codebook for *States* Database

This database includes variables for the 50 states. The data come from a variety of sources and are the latest available information.

AbortRate	Number of legal abortions per 1000 women, 2008
Asian	Percentage of the population Asian, 2010
Asslt12	Number of assaults per 100,000 population, 2012
BirthRate	Number of births per 1000 population, 2012
Black	Percentage of the population black, 2010
Burg12	Number of burglaries per 100,000 population, 2012
Carthft12	Number of car thefts per 100,000 population, 2012
College	Percentage of the population with a bachelor's degree or more
CommTime	Average travel time to work, 2009
CopSpend	Spending on police per capita, 2007
DeathRate	Number of deaths per 1000 population, 2009
DivRate	Divorce rate per 1000 population, 2011
Edspend	Spending on education K–12 per pupil, 2011
FamPoor00	Percentage of families below the poverty line in 2000
FamPoor09	Percentage of families below the poverty line in 2009
ForBorn	Percentage of the population foreign-born, 2009
Hispanic	Percentage of the population Hispanic, 2009
Hom12	Number of homicides per 100,000 population, 2012
Hom95	Number of homicides per 100,000 population, 1995
HS	Percentage of the population with a high school degree or more, 2009
InfantMort	Deaths of infants age 0–1 per 1000 live births, 2010
Internet	Percentage of households using the Internet, 2010
Larc12	Number of larcenies per 100,000 population, 2012
MarrRate	Number of marriages per 1000 population, 2011
MdHHInc	Median household income, 2009
Older	Percentage of the population 65 years of age or older, 2010
PopDense	Number of people per square mile, 2010
PopGrow	Population growth rate, 2000–2010
PopRank	Rank of population size, 2010
Rape12	Number of rapes per 100,000 population, 2012
Robb12	Number of robberies per 100,000 population, 2012

State	Name of state
SthDUMMY	South (0)/Non-South (1)
TeenBirthRate	Births per 1000 females age 15–19, 2011
TrafDths11	Traffic fatalities per 100 million miles driven, 2011
TrafDths90	Traffic fatalities per 100 million miles driven, 1990
Unemplymnt	Unemployment rate, 2012
Voters	Percentage of the population voting for president, 2012
White	Percentage of the population white, 2010
WmnRep	Percentage of women in the state legislature, 2013
Younger	Percentage of the population 18 and younger, 2010

Codebook for International Population (*Intl-POP*) Database

This database includes variables for a sample of 99 nations from various levels of development and geographic regions. Most of the information comes from the Population Reference Bureau's *2013 World Population Data Sheet* but a variety of other sources were also used.

BirthRate	Number of births per 1000 population, 2013
CellPhones	Subscribers per 100 population, 2012
Corruption	Perceived public sector corruption, 2013. Scores are based on the judgements of a panel of experts. The lower the score, the higher the corruption.
DthRate	Number of deaths per 1000 population, 2013
EducFemales	Ratio of number of females enrolled in a post-secondary educational program to number of females in the appropriate age group (e.g., 18–22). The value can exceed 100 if the total enrollment of all age groups is greater than the number in the specific age group
EducMales	Ratio of number of males enrolled in a post-secondary educational program to number of males in the appropriate age group (e.g., 18–22). The value can exceed 100 if the total enrollment of all age groups is greater than the number in the specific age group
FemaleLabor	Percentage of females age 15–24 in paid labor force
GNIcap	Gross national income per capita, 2012
IncomeIneq	Percent of income to richest 20% of population

IncLevel	Income level (World Bank classification), 2012 1. Low income 2. Lower middle income 3. Upper middle income 4. High income
INFM	Infant mortality rate: Death rate for ages 0–1, 2013
LEMen	Life expectancy at birth for men, 2013
LEWomen	Life expectancy at birth for women, 2013
LifeExp	Overall life expectancy at birth, 2013
LOD	Level of development: Collapsed income level 1. Low income 2. Middle income 3. High income
MaleLabor	Percentage of males age 15–24 in paid labor force
Migrate	Net migration rate per 1000 population, 2013
Nation	Name of nation
PctPopOld	Percentage of population older than 65, 2013
PctPopYng	Percentage of population younger than 15, 2013
Popgro	Projected 2050 population as a multiple of 2012 population. For example, a score of 2 on this variable means that the population is projected to double by 2050.
Region	Region of the world 1. East Asia and the Pacific 2. Eastern Europe and Central Asia 3. Latin America and the Caribbean 4. Middle East and North Africa 5. South Asia 6. Sub-Saharan Africa 7. Western Europe 8. North America 9. Oceania
Religiosity	Percentage of the population who agree that “religion is an important part of everyday life.”
Rights	Civil and political rights 1. Free 2. 3. 4. Partly free 5. 6. 7. Not free

TFR	Total fertility rate: Average number of children a woman has in her lifetime, 2013
Urban	Percentage of the population living in urban areas, 2013

Codebook for State Crime Trends (*CrimeTrends84-12*) Database

This database is used solely for graphing exercises in Chapter 2. All information is from the *Uniform Crime Reports* of the Federal Bureau of Investigation, 1984–2012.

AggAssault	Number of aggravated assaults per 100,000 population
Burglary	Number of burglaries per 100,000 population
CarTheft	Number of auto thefts per 100,000 population
Homicide	Number of homicides per 100,000 population
Robbery	Number of robberies per 100,000 population
Year	Years from 1984 to 2012

Answers to Odd-Numbered End-of-Chapter Problems

In addition to answers, this section suggests problem-solving strategies and provides examples of how to phrase or interpret conclusions and answers. You should try to solve and interpret the problems on your own before consulting this section.

In solving these problems, I let my calculator or computer do most of the work. I worked with whatever level of precision these devices permitted and didn't round off until the end or until I had to record an intermediate sum. I always rounded off to two places of accuracy (that is, to two places beyond the decimal point, or to hundredths). If you follow these same conventions, your answers will match mine. However, you should realize that small, generally trivial discrepancies between your answer and mine still might occur. If the difference doesn't seem trivial, you should double-check to make sure you haven't made an error or solve the problem again using a greater degree of precision.

Finally, please allow me a brief disclaimer about mathematical or typographical errors in this section. Let me assure you, first of all, that I know how important this section is for most students and that I worked hard to be certain that these answers are correct. Human fallibility being what it is, however, I know that I cannot make absolute guarantees. Should you find any errors, please let me know so that I can make corrections in the future.

Chapter 1

1.5
- **a.** Nominal
- **b.** Ordinal (The categories can be ranked in terms of degree of honesty with "Returned the wallet with money" the "most honest.")
- **c.** Ordinal
- **d.** Interval-ratio ("Years" has equal intervals and a true zero point.)
- **e.** Interval-ratio
- **f.** Interval-ratio
- **g.** Nominal (The various patterns are different from each other but cannot be ranked from high to low.)
- **h.** Interval-ratio
- **i.** Ordinal
- **j.** Number of accidents: interval-ratio; Severity of accident: ordinal

1.7

	Variable	Level of Measurement	Application
a.	Opinion	Ordinal	Inferential
b.	Letter grade	Ordinal	Descriptive (two variables)
	% correct	Interval-ratio	
c.	Party	Nominal	Inferential
	Gender	Nominal	
	% in favor	Interval-ratio	
d.	Homicides	Interval-ratio	Descriptive (two variables)
	Before/After Change	Nominal	
e.	Satisfaction	Ordinal	Descriptive (one variable)

1.9

	Variable	Level of Measurement	Type
a.	Racial Self ID	Nominal	Independent
	Desirability of categories	Ordinal	Dependent
b.	Number of relationships	Interval-ratio	Dependent
	Number of times having intercourse	Interval-ratio	Dependent
	GPA	Interval-ratio	Independent
c.	% minority voters	Interval-ratio	Independent
	Turnout	Interval-ratio	Dependent
	% minority officials	Interval-ratio	Independent
d.	Per capita income	Interval-ratio	Independent
	School enrollment (Total)	Interval-ratio	Dependent
	School enrollment (Boys)	Interval-ratio	Dependent
	School enrollment (Girls)	Interval-ratio	Dependent
e.	Support for gun control	Ordinal	Dependent
	Years of schooling	Interval-ratio	Independent
	Gender	Nominal	Independent
	Region	Nominal	Independent
	Party preference	Nominal	Independent

Chapter 2

2.1 **a.** Complex A: $(5/20) \times 100 = (0.25) \times 100 = 25.00\%$
Complex B: $(10/20) \times 100 = 50.00\%$
b. Complex A: $4{:}5 = 0.80$
Complex B: $6{:}10 = 0.60$
c. Complex A: $(0/20) = 0.00$
Complex B: $(1/20) = 0.05$
d. $(6/(4 + 6)) \times 100 = (6/10) \times 100 = 60.0\%$
e. Complex A: $8{:}5 = 1.60$
Complex B: $2{:}10 = 0.20$

2.3 Bank robbery rate $= (47/211{,}732) \times 100{,}000 = (0.0002219) \times 100{,}000 = 22.20$
Murder rate $= (13/211{,}732) \times 100{,}000 = (0.00006139) \times 100{,}000 = 6.14$
Auto theft rate $= (23/211{,}732) \times 100{,}000 = (0.00010862) \times 100{,}000 = 10.86$

2.5 For gender:

Sex	Frequency
Male	9
Female	6
Total	15

For age, set $k = 10$. $R = 77 - 23$, or 54, so we can round off interval size to $i = 5$. The first interval will be 20–24, to include the low score of 23, and the highest interval will be 75–79.

Age	Frequency
20–24	1
25–29	2
30–34	3
35–39	2
40–44	1
45–49	3
50–54	1
55–59	1
60–64	0
65–69	0
70–74	0
75–79	1
Total	15

2.9 Set $k = 10$. $R = 92 - 5$, or 87, so set i at 10.

Score	Frequency	Midpoint	Percentage	Cumulative *f*	Cumulative %
0–9	3	4.5	12.0%	3	12.0%
10–19	7	14.5	28.0%	10	40.0%
20–29	6	24.5	24.0%	16	64.0%

(*continued*)

(*continued*)

Score	Frequency	Midpoint	Percentage	Cumulative *f*	Cumulative %
30–39	0	34.5	0.0%	16	64.0%
40–49	2	44.5	8.0%	18	72.0%
50–59	2	54.5	8.0%	20	80.0%
60–69	3	64.5	12.0%	23	92.0%
70–79	0	74.5	0.0%	23	92.0%
80–89	0	84.5	0.0%	23	92.0%
90–99	2	94.5	8.0%	25	100.0%
Total	25		100.0		

2.11 Answers should describe the rise in all crime rates except burglary, up to a peak in the early 1990s, followed by a decline. Burglary rates declined throughout the period, but the rate of decline has leveled off since 2000.

2.13 One possible description of *relig* might be "Protestant was the most common religious denomination by far, followed by Catholic and 'None.' The smallest categories were 'Other' and Jews." A description of hours worked should note the peak in the middle of the graph at 40 hours.

Chapter 3

3.1 "Region of birth" and "religion" are nominal-level variables, "support for legalization" and "opinion of food" are ordinal, and "expenses" and "number of movies" are interval-ratio. The mode, the most common score, is the only measure of central tendency available for nominal-level variables. For the two ordinal-level variables, *don't forget to arrange the scores from high to low* before locating the median. There are 10 freshmen (N is even), so the median for freshmen will be the score halfway between the scores of the two middle cases. There are 11 seniors (N is odd), so the median for seniors will be the score of the middle case. To find the mean for the interval-ratio variables, add the scores and divide by the number of cases.

Variable	Measure	Freshmen	Seniors
Region of birth	Mode	North	North
Legalization	Median	3	5
Expenses	Mean	58.50	72.55
Movies	Mean	5.80	5.18
Food	Median	6	4
Religion	Mode	Protestant	Protestant, None

3.3

Variable	Level of Measurement	Measure of Central Tendency
Gender	Nominal	Mode = Male
Class	Ordinal	Median = Middle
Years	Interval-ratio	Mean = 26.15
Education	Ordinal	Median = High school
Marital status	Nominal	Mode = Married
No. of children	Interval-ratio	Mean = 2.39

3.5

Variable	Level of Measurement	Measure of Central Tendency
Marital status	Nominal	Mode = Married
Race/Ethnic group	Nominal	Mode = White
Age	Interval-ratio	Mean = 27.53
Attitude	Ordinal	Median = 7

3.7

Variable	Level of Measurement	Measure of Central Tendency
Gender	Nominal	Mode = Male
Support	Ordinal	Median = 1
Education	Ordinal	Median = 1
Age	Interval-ratio	Mean = 40.93

3.9 Attitude scales almost always generate ordinal-level data, so the appropriate measure of central tendency would be the median. The median is 9 for the students and 2 for the neighbors. Incidentally, the means are 7.80 for the students and 4.00 for the neighbors.

3.11 Mean = 310.73, median = 271. The higher value for the mean indicates a positive skew. Note the very high score of the United States, Australia, and a few other nations.

3.13 To find the median, you must first rank the scores from high to low. Both groups have 25 cases (*N* is odd), so the median is the score of the 13th case. For freshmen the median is 35, and for seniors the median is 30. The mean score for freshmen is 31.72, and for seniors the mean is 28.60.

3.15 The owners are using the mean while the players are citing the median. This distribution has a positive skew (the mean is greater than the median), as is typical for income data. Note the difference in wording in the two reports: The owners cite the "average" (mean) and the players cite the "typical player" (the score of the middle case or the median).

3.17

Variable	Level of Measurement	Measure of Central Tendency
closeblk	Ordinal	Median = 5
hrs1	Interval-ratio	Mean = 39.75
partnrs5	Ordinal	Median = 1
region	Nominal	Mode = 5 (South Atlantic)

Note: *partnrs5* has an unusual coding scheme. Scores 0–4 are actual numbers (a score of "1" means one partner) but higher values represent broad categories. A score of 6, for example, means "between 11 and 20 partners." The variable is interval-ratio from scores 0 to 4 and then ordinal for the remaining scores. Given this "hybrid" scoring scheme, it is best to be conservative and treat the variable as ordinal.

Chapter 4

4.1 The high score is 50 and the low score is 10, so the range is 50 − 10, or 40. The standard deviation is 12.28.

4.3

	Freshmen		Seniors	
Variable	*R*	*s*	*R*	*s*
Expenses	32	9.21	45	14.43
Movies	14	5.06	14	4.04
Food	10	3.35	8	2.50

4.5

	Canadian Provinces		U.S. States	
	2000	2011	1999	2012
Mean	48,161.54	74,723.85	42,235.15	52,621.92
Median	47,300.00	68,710.00	40,058.00	50,015.00
Range	23,400	41,630	22,548	32,818
Std Dev.	6,753.64	12,209.67	6,635.91	8,880.79

For the Canadian provinces, the mean and median increase across the time period. For both years, the distributions show a positive skew (the mean is greater than the median). The standard deviation and the range increase dramatically, indicating that there is much greater variability in 2011. These changes are largely due to the much higher income in the Northwest Territories. The patterns are similar for the U.S. states, although not as dramatic.

4.7

	Labor Force Participation Rate		Percentage HS Graduates		Mean Income	
State	Male	Female	Male	Female	Male	Female
Mean	72.03	60.75	85.94	87.31	54,271.30	50,760.40
s	3.67	4.14	2.95	2.95	4245.89	4076.04

Males and females are very similar in terms of percentage of high school graduates, but females are less involved in the labor force and, on average, earn about $3500 less per year than males. The females in these ten states are more variable in their labor force participation but are similar to males in dispersion on the other two variables.

4.9

Mean	5.39
Median	4.84
R	9.08
Q	4.13
s	2.35

4.11

	2000	2010
Mean	98.65	160.85
s	31.67	45.90

4.13

	1973	1975
Mean	12.21	19.06
s	12.20	9.63

4.15 Department B has the highest average (mean) but Department A is the most consistent (has the lowest standard deviation). Which is more important, number of arrests or steadiness of performance? Many governors would probably lean towards Department B but a case could be made for Department A. Note that, by these statistics, Department D is the most inconsistent in its performance.

4.17

	Freshmen	Seniors
Mean	31.72	28.60
s	12.53	13.35

The mean score declined, so, on the average, there was less prejudice among seniors. The standard deviation increased, which means that seniors were more variable than freshmen. This reflects the fact that seniors had more low scores *and* more high scores.

4.19

	FamPoor00	FamPoor09
Mean	8.82	9.87
s	2.80	2.75
R	12.50	11.80

Chapter 5

5.1

X_i	*Z* Score	% Area Above	% Area Below
5	−1.67	95.25	4.75
6	−1.33	90.82	9.18
7	−1.00	84.13	15.87
8	−0.67	74.86	25.14
9	−0.33	62.93	37.07
11	0.33	37.07	62.93
12	0.67	25.14	74.86
14	1.33	9.18	90.82
15	1.67	4.75	95.25
16	2.00	2.28	97.72
18	2.67	0.38	99.62

5.3

	Z Scores	Area
a.	0.10 & 1.10	32.45%
b.	0.60 & 1.10	13.86%
c.	0.60	27.43%
d.	0.90	18.41%
e.	0.60 & −0.40	38.11%
f.	0.10 & −0.40	19.52%
g.	0.10	53.98%
h.	0.30	61.79%
i.	0.60	72.57%
j.	1.10	86.43%

5.5

X_i	*Z* Score	Number of Students Above	Number of Students Below
60	−2.00	195	5
57	−2.50	199	1
55	−2.83	200	0
67	−0.83	159	41
70	−0.33	126	74
72	0.00	100	100
78	1.00	32	168
82	1.67	10	190
90	3.00	0	200
95	3.83	0	200

Note: Number of students has been rounded off to the nearest whole number.

5.7

	Z Scores	Area
a.	−2.20	1.39%
b.	1.80	96.41%
c.	−0.20 & 1.80	54.34%
d.	0.80 & 2.80	20.93%
e.	−1.20	88.49%
f.	0.80	21.19%

5.9

	Z Scores	Area
a.	−1.00 & 1.50	0.7745
b.	0.25 & 1.50	0.3345
c.	1.50	0.0668
d.	0.25 & −22.25	0.5865
e.	−1.00 & −22.25	0.1465
f.	−1.00	0.1587

5.11 Yes. The raw score of 110 translates into a *Z* score of 2.88. 99.80% of the area lies below this score, so this individual was in the top 1% on this test.

5.13 For the first event, the probability is 0.0919; for the second, the probability is 0.0655. The first event is more likely.

5.15 a.

	Z score—Freshman	Z score—Senior	Student did better as a
A	0.57	1.25	Senior
B	−0.29	0.50	Senior
C	−1.14	−2.50	Freshman (−1.14 is closer to the mean than 22.50)
D	2.86	2.25	Freshman
E	1.29	1.50	Senior

b. Freshman Year

Score	Z Score	Probability
Less than 52	−0.14	0.4443
Less than 57	0.57	0.7157
Between 40 and 50	−1.86 & −0.43	0.3022
More than 51	−0.29	0.6141
More than 62	1.29	0.0985

Senior Year

Score	Z Score	Probability
Less than 88	−1.00	0.1587
Less than 98	1.50	0.9332
Between 80 and 100	−3.00 & 2.00 −0.00	0.9758
More than 97	1.25	0.1056
More than 85	−1.75	0.9599

Chapter 6

6.1 **a.** 5.2 ± 0.11
b. 100 ± 0.71
c. 20 ± 0.40
d. 1020 ± 5.41
e. 7.3 ± 0.23
f. 33 ± 0.80

6.3

Confidence Level	Alpha	Area Beyond Z	Z score
95%	0.05	0.0250	±1.96
94%	0.06	0.0300	±1.88 or ±1.89
92%	0.08	0.0400	±1.75 or ±1.76
97%	0.03	0.0150	±2.17
98%	0.02	0.0100	±2.33 or ±2.32

6.5 **a.** 2.30 ± 0.04
b. 2.10 ± 0.01, 0.78 ± .07
c. 6.00 ± 0.37

6.7 **a.** 478.23 ± 1.97. The students spent between \$476.26 and \$480.20 on books.
b. 1.5 ± 0.04. The students visited the clinic between 1.46 and 1.54 times on the average.
c. 2.8 ± 0.13
d. 3.5 ± 0.19

6.9 0.14 ± 0.07. The estimate is that between 7% and 21% of the population consists of unmarried couples living together.

6.11 **a.** P_s = 823/1496 = 0.55. The confidence interval is 0.55 ± 0.03. Between 52% and 58% of the population agrees with the statement.
b. P_s = 650/1496 = 0.43. The confidence interval is 0.43 ± 0.03.
c. P_s = 375/1496 = 0.25. The confidence interval is 0.25 ± 0.03.
d. P_s = 1023/1496 = 0.68. The confidence interval is 0.68 ± 0.03.
e. P_s = 800/1496 = 0.53. The confidence interval is 0.53 ± 0.03.

6.13

Alpha (α)	Confidence Level	Z score	Confidence Interval
0.10	90%	±1.65	100 ± 0.74
0.05	95%	±1.96	100 ± 0.88
0.01	99%	±2.58	100 ± 1.15
0.001	99.9%	±3.32	100 ± 1.49
0.0001	99.99%	±3.90	100 ± 1.76

6.15 The confidence interval is 0.51 ± 0.05 and the estimate would be that between 46% and 56% of the population prefer candidate A. The population parameter (P_u) is equally likely to be anywhere in the interval (that is, it's just as likely to be 46% as it is to be 56%), so a winner cannot be predicted.

6.17 The sample proportion is 0.23 (35/150) and the confidence interval is 0.23 ± 0.08. At the 95% confidence level, the estimate would be that between 240 (15%) and 496 (31%) of the 1600 freshmen would be extremely interested. The estimated numbers are found by multiplying *N* (1600) by the upper (0.31) and lower (0.15) limits of the interval.

6.19 **a.** The confidence interval for *age* is 48.21 ± 0.92 and the confidence interval for *hrs1* is 39.75 ± 1.02.
b. The confidence interval for the proportion of Catholics in the population is 0.23 ± 0.03 and the confidence interval for the proportion of supporters of the death penalty in the population is 0.66 ± 0.03.

Chapter 7

7.1 **a.**

Z(Critical)
+1.65 or −1.65
±1.65
±1.89
+2.33 or −2.33
±2.33

b.

df	*t*(Critical)
30	±1.697
23	±2.500
120	±2.617
30	+2.457 or −2.457
60	+1.671 or −1.671

c. **1** *Z*(obtained) = −3.76
2 *t*(obtained) = −2.21
3 *Z*(obtained) = −5.74
4 *Z*(obtained) = 0.66
5 *Z*(obtained) = −0.77

7.3 **a.** *Z*(obtained) = −41.00
b. *Z*(obtained) = −29.09

7.5 *Z*(obtained) = 6.04

7.7 **a.** *Z*(obtained) = −13.66
b. *Z*(obtained) = 25.50

7.9 *t*(obtained) = 4.50

7.11 *Z*(obtained) = 3.06

7.13 *Z*(obtained) = −1.48

7.15 **a.** *Z*(obtained) = −1.49
b. *Z*(obtained) = 2.19
c. *Z*(obtained) = −8.55
d. *Z*(obtained) = −18.07
e. *Z*(obtained) = 2.09
f. *Z*(obtained) = −53.33

7.17 *t*(obtained) = −1.14

7.19 **a.** *Z*(obtained) = −1.28
b. *Z*(obtained) = 3.57
c. *Z*(obtained) = 34.96
d. *Z*(obtained) = −19.09
e. *Z*(obtained) = 35.00
f. *Z*(obtained) = −140.00

Chapter 8

8.1 **a.** $\sigma_{\bar{X}-\bar{X}} = 1.39$, *Z*(obtained) = −2.52
b. $\sigma_{\bar{X}-\bar{X}} = 1.61$, *Z*(obtained) = 2.48

8.3 **a.** $\sigma_{\bar{X}-\bar{X}} = 10.57$, *Z*(obtained) = 1.70
b. $\sigma_{\bar{X}-\bar{X}} = 11.28$, *Z*(obtained) = −2.48

8.5 **a.** $\sigma_{\bar{X}-\bar{X}} = 0.08$, *Z*(obtained) = −11.25
b. $\sigma_{\bar{X}-\bar{X}} = 0.12$, *Z*(obtained) = −3.33
$\sigma_{\bar{X}-\bar{X}} = 0.15$, *Z*(obtained) = 20.00

8.7 These are small samples (combined *N*'s of less than 100), so be sure to use Formulas 8.5 and 8.6 in step 4.
a. $\sigma_{\bar{X}-\bar{X}} = 0.12$, *t*(obtained) = −1.33
b. $\sigma_{\bar{X}-\bar{X}} = 0.13$, *t*(obtained) = 14.85

8.9 **a.** (Brazil) $\sigma_{\bar{X}-\bar{X}} = 0.11$, *Z*(obtained) = 0.91
b. (Ukraine) $\sigma_{\bar{X}-\bar{X}} = 0.15$, *Z*(obtained) = 2.00
c. (Vietnam) $\sigma_{\bar{X}-\bar{X}} = 0.10$, *Z*(obtained) = 2.00
d. (South Africa) $\sigma_{\bar{X}-\bar{X}} = 0.09$, *Z*(obtained) = −4.44
e. (Egypt) $\sigma_{\bar{X}-\bar{X}} = 0.10$, *Z*(obtained) = −3.00

8.11 $P_u = 0.45$, $\sigma_{p-p} = 0.06$, *Z*(obtained) = 0.67

8.13 **a.** $P_u = 0.46$, $\sigma_{p-p} = 0.06$, *Z*(obtained) = 2.17
b. $P_u = 0.80$, $\sigma_{p-p} = 0.07$, *Z*(obtained) = −1.57
c. $P_u = 0.73$, $\sigma_{p-p} = 0.08$, *Z*(obtained) = 0.63

8.15 **a.** $P_u = 0.34$, $\sigma_{p-p} = 0.04$, *Z*(obtained) = 1.50
b. $P_u = 0.20$, $\sigma_{p-p} = 0.05$, *Z*(obtained) = −2.20
c. $P_u = 0.47$, $\sigma_{p-p} = 0.03$, *Z*(obtained) = −2.67
d. $\sigma_{\bar{X}-\bar{X}} = 0.43$, *Z*(obtained) = 1.86
e. $\sigma_{\bar{X}-\bar{X}} = 0.14$, *Z*(obtained) = −5.71
f. $\sigma_{\bar{X}-\bar{X}} = 0.08$, *Z*(obtained) = −5.50

8.17

Variable	Categories	*N*	Mean	*s*	*t* score	*df*	Sig.
educ	Protestant	760	13.41	3.110	1.868	1090	0.062
	Catholic	332	13.02	3.376			
rank	Protestant	715	4.79	1.890	−0.641	1032	0.522
	Catholic	319	4.87	1.757			

There is no significant difference between Protestants and Catholics on either variable.

Chapter 9

9.1

	Grand Mean	SST	SSB	SSW	*F* Ratio
a.	12.17	231.67	173.17	58.5	13.32
b.	6.87	455.73	78.53	377.20	1.25
c.	31.65	8362.55	5053.35	3309.2	8.14

9.3

	Grand Mean	SST	SSB	SSW	*F* Ratio
a.	4.39	86.28	45.78	40.50	8.48
b.	16.44	332.44	65.44	267.00	1.84

For problem 9.3a, with alpha = 0.05 and *df* = 2, 15, the critical *F* ratio would be 3.68. We would reject the null hypothesis and conclude that decision making *does* vary significantly by type of relationship. By inspection of the group means, it seems that the "cohabitational" category accounts for most of the differences.

9.5

Grand Mean	SST	SSB	SSW	*F* Ratio
9.28	213.61	2.11	211.50	0.08

9.7

Grand Mean	SST	SSB	SSW	*F* Ratio
5.40	427.32	124.06	303.26	6.00

9.9

	Grand Mean	SST	SSB	SSW	*F* Ratio
Mexico	3.78	300.98	154.08	146.90	12.59
Canada	6.88	156.38	20.08	136.30	1.77
U.S.	5.13	286.38	135.28	151.10	10.74

At the 0.05 level of significance and *df* = 3, 36, the critical *F* ratio is 2.92. There is a significant difference in support for suicide by class in Mexico and the United States but not in Canada. The category means for Mexico suggest that the upper class accounts for most of the differences. For the United States, there is more variation across the category means, and the working class seems to account for most of the differences. Going beyond the ANOVA test and comparing the grand means, we see that support is highest in Canada and lowest in Mexico.

9.11 **a.** *educ.*

	Group Means	*df*	Sig.	*F* Ratio
Protestant	13.41	4, 1446	0.000	10.602
Catholic	13.02			
Jewish	15.67			
None	14.29			
Other	14.68			

b. *rank*

	Group Means	*df*	Sig.	*F* Ratio
Protestant	4.79	4, 1370	0.367	1.076
Catholic	4.87			
Jewish	4.06			
None	4.72			
Other	5.00			

Chapter 10

10.1 **a.** 1.11
b. 0.00
c. 1.52
d. 1.46

10.3 **a.** A computing table is highly recommended as a way of organizing the computations for chi square:

Computational Table for Problem 10.3

(1)	(2)	(3)	(4)	(5)
f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
6	5	1	1	0.20
7	8	−1	1	0.13
4	5	−1	1	0.20
9	8	1	1	0.13
26	26	0		$\chi^2 = 0.66$

There is one degree of freedom in a 2 × 2 table. With alpha set at 0.05, the critical value for the chi square is 3.841. The obtained chi square is 0.66, so we fail to reject the null hypothesis of independence between the variables. There is no statistically significant relationship between race and services received.

b.

	Race	
Received Services?	Black	White
Yes	60.00%	43.75%
No	40.00%	56.25%
Totals	100.00%	100.00%

The relationship is not significant, but the pattern is that blacks are more likely to receive services than whites.

10.5 **a.**

Computational Table for Problem 10.5a

(1)	(2)	(3)	(4)	(5)
f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
21	17.5	3.5	12.25	0.70
29	32.5	−3.5	12.25	0.38
14	17.5	−3.5	12.25	0.70
36	32.5	3.5	12.25	0.38
100	100.00	0.0		$\chi^2 = 2.16$

With $df = 1$ and alpha = 0.05, the critical value for the chi square is 3.841. The obtained chi square of 2.16 does not fall within this area, so the null hypothesis cannot be rejected. There is no statistically significant relationship between unionization and salary.

Column Percentages

	Status	
Salary	Union	Nonunion
High	60.00%	44.62%
Low	40.00%	55.38%
Totals	100.00%	100.00%

The relationship is not significant but unionized fire departments tend to have higher salary levels.

10.7 The obtained chi square is 5.13, which is significant ($df = 1$, alpha = 0.05). The column percentages show that more affluent communities have higher-quality schools.

10.9 The obtained chi square is 6.67, which is significant ($df = 2$, alpha = 0.05). The column percentages show that shorter marriages have higher satisfaction.

10.11 The obtained chi square is 12.58, which is significant ($df = 4$, alpha = 0.05). The column percentages show that proportionally more of the students living "off campus with roommates" are in the high-GPA category.

10.13 The obtained chi square is 19.34, which is significant ($df = 3$, alpha = 0.05). Legalization was not favored by a majority of any region, but the column percentages show that the West was most in favor.

10.15 **a.**

	Gender	
Preference	Male	Female
Romney	45.21%	45.17%
Obama	54.79%	54.83%
Totals	100.00%	100.00%

$\chi^2 = 0.00$, $df = 1$, $p > 0.05$

b.

	Race/Ethnicity		
Preference	White	Black	Latino
Romney	53.71%	5.00%	40.00%
Obama	46.28%	95.00%	60.00%
Totals	100.00%	100.00%	100.00%

$\chi^2 = 82.21$, $df = 1$, $p < 0.05$

c.

	Education			
Preference	Less than HS	HS Graduate	College Graduate	Postgrad. Degree
Romney	46.15%	60.00%	35.12%	21.28%
Obama	53.85%	40.00%	64.88%	78.72%
Totals	100.00%	100.00%	100.00%	100.00%

$\chi^2 = 51.20$, $df = 3$, $p < 0.05$

d.

	Religion				
Preference	Protestant	Catholic	Jewish	None	Other
Romney	40.24%	66.67%	33.33%	31.82%	45.45%
Obama	59.76%	33.33%	66.67%	68.18%	54.55%
Totals	100.00%	100.00%	100.00%	100.00%	100.00%

$\chi^2 = 42.83$, $df = 4$, $p < 0.05$

10.17

	Religion				
bigbang	LT HS	HS	Jr. Coll	BA	Grad
True	50.0%	49.1%	68.0%	58.0%	79.3%
False	50.0%	50.9%	32.0%	42.0%	20.7%
Totals	100.00%	100.00%	100.00%	100.00%	100.00%

$\chi^2 = 10.83$, $df = 4$, $p < 0.05$

	Religion				
evolved	LT HS	HS	Jr. Coll	BA	Grad
True	50.0%	53.1%	60.9%	52.4%	75.0%
False	50.0%	46.9%	39.1%	47.8%	25.0%
Totals	100.00%	100.00%	100.00%	100.00%	100.00%

$\chi^2 = 7.07$, $df = 4$, $p > 0.05$

Chapter 11

11.1 a.

	Gender	
Support?	Males	Females
For	44.44%	35.00%
Against	55.56%	65.00%
Totals	100.00%	100.00%

Phi = 0.10; Lambda = 0.00; Max. diff. = 9.44 This relationship is weak to moderate.

b.

	Discipline	
Support?	Liberal Arts	Science and Business
For	30.00%	48.15%
Against	70.00%	51.85%
Totals	100.00%	100.00%

Phi = 0.18; Lambda = 0.00; Max. diff. = 18.15 This relationship is moderate to strong.

c.

	Status	
Support?	Tenured	Nontenured
For	45.45%	28.57%
Against	54.55%	71.43%
Totals	100.00%	100.00%

Phi = 0.16; Lambda = 0.00; Max. diff. = 16.88 This relationship is moderate to strong.

11.3 Phi = 0.59; Lambda = 0.54; Max. diff. = 60.00 This is a strong relationship.

11.5 **a.** Cramer's $V = 0.33$; Lambda = 0.12
b. Cramer's $V = 0.26$; Lambda = 0.18
c. Cramer's $V = 0.24$; Lambda = 0.16

11.7 Gamma = −0.61; Max. diff. = 33.56. This is a strong, negative relationship. As authoritarianism increases, efficiency decreases. *NOTE:* These variables are associated, but which causes which? The problem assumes that authoritarianism is the independent variable ("mean bosses cause lazy workers") but could the causal relationship go in the other direction ("lazy workers make bosses mean")? Remember that correlation and causation are two different things.

11.9 $G = -0.17$. This is a weak, negative relationship. The higher the level of alcohol use among parents, the fewer the symptoms of depression.

11.11 $r_s = 0.33$

11.13

sex and	Phi or *V*	Lambda	
fear	Phi = 0.25	$\lambda = 0.07$	Phi and the changes in the column percentages show that this is a moderate to strong relationship. Ignore lambda.
cappun	Phi = 0.08	$\lambda = 0.00$	This is a moderate to weak relationship. Ignore lambda.
satjob	$V = 0.04$	$\lambda = 0.00$	This is a weak relationship. Ignore lambda.

Chapter 12

12.1 *(HINT: When finding the slope, remember that "Turnout" is the dependent, or Y, variable.)*

	For Turnout (*Y*) and		
	Unemployment	Education	Neg. Campaigning
a	39.00	−94.73	114.01
b	3.00	12.67	−0.90
Reg. equation	$Y = 39.0 + 3.0X$	$Y = -94.73 + 12.67X$	$Y = 114.01 + (-0.90)X$
r	0.95	0.98	−0.87
r^2	0.90	0.96	0.76

12.3 *(HINT: When finding the slope, remember that "Number of visitors" is the dependent, or Y, variable.)*
Slope (b) = −0.37
Y intercept (a) = 13.42
$r = -0.31$
$r^2 = 0.10$

12.5 The correlation between the variables ($r = 0.40$) is positive and moderate in strength. This tends to support the coach's argument.

12.7

	attend		*childs*		*rank*		*tvhours*	
	r	r^2	*r*	r^2	*r*	r^2	*r*	r^2
age	0.16*	0.03	0.37*	0.14	−0.10*	0.01	0.15*	0.02

12.9

	Hom12		*CarThft12*	
	r	r^2	r	r^2
PopGrow	−0.04	0.00	0.39	0.15
PopDense	0.04	0.00	−0.02	0.00
Unemployment	0.32	0.10	0.36	0.13

12.11

	GNIcap		*Urban*	
	r	r^2	r	r^2
CellPhones	0.55	0.30	0.58	0.34

Chapter 13

13.1 **a.** For turnout (Y) and unemployment (X) while controlling for negative advertising (Z), $r_{yx.z}$ = 0.95. The relationship between X and Y is not affected by the control variable Z.

b. For turnout (Y) and negative advertising (X) while controlling for unemployment (Z), $r_{yx.z}$ = −0.89. The bivariate relationship is not affected by the control variable.

c. Turnout (Y) = 70.25 + (2.09) unemployment (X_1) + (−0.43) negative advertising (X_2). For unemployment (X_1) = 10 and negative advertising (X_2) = 75, turnout (Y) = 58.90.

d. For unemployment (X_1): = 0.66. For negative advertising (X_2): = −0.41. Unemployment has a stronger effect on turnout than negative advertising. Note that the independent variables' effects on turnout are in opposite directions.

e. $R^2 = 0.98$

13.3 **a.** Cohesion (Y) = 4.31 + (0.65) children (X_1) + (−0.00001) income (X_2).

b. For children (X_1) = 6 and income (X_2) = 20,000, cohesion (Y) = 8.41.

c. For children (X_1): = 0.37. For income (X_2): = −0.06. Number of children has a stronger effect than income. Note that the independent variables' effects on cohesion are in opposite directions.

d. $R^2 = 0.14$

13.5 **a.** The zero-order correlation between *wwwhr* and *age* is −0.20. The partial correlation coefficient when controlling for *educ* is −0.22. The relationship seems to be direct: Internet use decreases with age regardless of level of education.

b. *wwwhr* (Y) = 7.16 + (−0.19) *age* (X_1) + (0.82) *educ* (X_2).

c. *wwwhr* (Z_y) = (−0.22) *age* (Z_1) + (0.16) *educ* (Z_2).

d. $R^2 = 0.06$

13.7 **a.** *Carthft12* (Y) = 59.78 + (−0.07) *PopDense* (X_1) + (21.61) *Unemployment* (X_2).

b. *Carthft12* (Z_y) = (−0.21) *PopDense* (Z_1) + (0.45) *Unemployment* (Z_2).

c.

Bivariate:	r	r^2
Carthft12 and		
PopDense	− 0.02	0.00
Unemployment	0.36	0.13
Multiple:	$R^2 = 0.16$	

13.9 **a.** *CellPhones* (Y) = 45.69 + (0.00) *GNIcap* (X_1) + (0.72) *Urban* (X_2).

b. *CellPhones* (Z_y) = (0.26) *GNIcap* (Z_1) + (0.42) *Urban* (Z_2).

c.

Bivariate:	r	r^2
CellPhones and		
GNIcap	0.55	0.30
Urban	0.60	0.36
Multiple:	$R^2 = 0.39$	

Glossary

Each entry includes a brief definition and the chapter number where the term was introduced.

Alpha (α). The probability of error, or the probability that a confidence interval does not contain the population value. Alpha levels are usually set at 0.10, 0.05, 0.01, or 0.001. Chapter 6

Alpha level (α). The proportion of area under the sampling distribution that contains unlikely sample outcomes, given that the null hypothesis is true. Also, the probability of Type I error. Chapter 7

Analysis of variance. A test of significance appropriate for situations in which we are concerned with the differences between more than two sample means. Chapter 9

ANOVA. See **Analysis of variance**. Chapter 9

Association. The relationship between two (or more) variables. Two variables are said to be associated if the distribution of one variable changes for the various categories or scores of the other variable. Chapter 11

Bar chart. A graph used for nominal and ordinal variables with only a few categories. Categories are represented by bars of equal width, the height of each corresponding to the number (or percentage) of cases in the category. Chapter 2

Beta-weight (b^*). See **Standardized partial slope**. Chapter 13

Bias. A criterion used to select sample statistics as estimators. A statistic is unbiased if the mean of its sampling distribution is equal to the population value of interest. Chapter 6

Bivariate table. A table that displays the joint frequency distributions of two variables. Chapter 10

Boxplot. A graph that presents information about the central tendency and dispersion of a variable. Chapter 4

Cells. The cross-classification categories of the variables in a bivariate table. Chapter 10

Central Limit Theorem. A theorem that specifies the mean, standard deviation, and shape of the sampling distribution, given that the sample is large. Chapter 6

χ^2(critical). The score on the sampling distribution of all possible sample chi squares that marks the beginning of the critical region. Chapter 10

χ^2(obtained). The test statistic computed from sample results. Chapter 10

Chi square (χ^2) test. A nonparametric test of hypothesis for variables that have been organized into a bivariate table. Chapter 10

Class intervals. The categories used in the frequency distributions for interval-ratio variables. Chapter 2

Coefficient of determination (r^2). The proportion of all variation in *Y* that is explained by *X*. Found by squaring the value of Pearson's *r*. Chapter 12

Coefficient of multiple determination (R^2). A statistic that equals the proportion of the total variation in the dependent variable that is explained by all independent variables combined. Chapter 13

Column. The vertical dimension of a bivariate table. By convention, each column represents a score on the independent variable. Chapter 10

Column percentages. Percentages computed within each column of a bivariate table. Chapter 10

Conditional distribution of *Y*. The distribution of scores on the dependent variable for a specific score or category of the independent variable when the variables have been organized into table format. Chapter 11

Conditional means of *Y*. The mean of all scores on *Y* for each value of *X*. Chapter 12

Confidence interval. An estimate of a population value in which a range of values is specified. Chapter 6

Confidence level. A frequently used alternate way of expressing alpha, the probability that an interval estimate will not contain the population value. Confidence levels of 90%, 95%, 99%, 99.9%, and 99.99% correspond to alphas of 0.10, 0.05, 0.01, 0.001, and 0.0001, respectively. Chapter 6

Control variable. In multivariate statistical analysis, a third variable (Z) whose effect on a bivariate relationship is held constant. Chapter 13

Correlation matrix. A table that shows the correlation coefficients between all possible pairs of variables. Chapter 12

Cramer's *V*. A chi square–based measure of association for nominally measured variables that have been organized into a bivariate table with any number of rows and columns. Chapter 11

Critical region (region of rejection). The area under the sampling distribution that, in advance of the test itself, is defined as including unlikely sample outcomes, given that the null hypothesis is true. Chapter 7

Cumulative frequency. An optional column in a frequency distribution that displays the number of cases within an interval and all preceding intervals. Chapter 2

Cumulative percentage. An optional column in a frequency distribution that displays the percentage of cases within an interval and all preceding intervals. Chapter 2

Data. Information expressed as numbers. Chapter 1

Data reduction. Summarizing many scores with a few statistics. Chapter 1

Dependent variable. A variable that is identified as an effect or outcome. The dependent variable is thought to be caused by the independent variable. Chapters 1 and 11

Descriptive statistics. The branch of statistics concerned with (1) summarizing the distribution of a single variable or (2) measuring the relationship between two or more variables. Chapter 1

Deviation. The distance between the score and the mean. Chapter 4

Direct relationship. A multivariate relationship in which the control variable has no effect on the bivariate relationship. Chapter 13

Dispersion. The amount of variety, or heterogeneity, in a distribution of scores. Chapter 4

Dummy variable. A nominal-level variable dichotomized so that it can be used in regression analysis. A dummy variable has two scores, one coded as 0 and the other as 1. Chapter 12

Efficiency. The extent to which the sample outcomes are clustered around the mean of the sampling distribution. Chapter 6

EPSEM. The **E**qual **P**robability of **SE**lection **M**ethod for selecting samples. Every element or case in the population must have an equal probability of selection for the sample. Chapter 6

Expected frequency (f_e). The cell frequencies that would be expected in a bivariate table if the variables were independent. Chapter 10

Explained variation. The proportion of all variation in Y that is attributed to the effect of X. Equal to $\sum(Y' - \bar{Y})^2$. Chapter 12

***F* ratio.** The test statistic computed in step 4 of the ANOVA test. Chapter 9

Five-step model. A step-by-step guideline for conducting tests of hypotheses. A framework that organizes decisions and computations for all tests of significance. Chapter 7

Frequency distribution. A table that displays the number of cases in each category of a variable. Chapter 2

Frequency polygon. See **Line chart**. Chapter 2

Gamma (*G*). A measure of association for "collapsed" ordinal variables in bivariate table format. Chapter 11

Histogram. A graph used for interval-ratio variables. Class intervals are represented by contiguous bars of equal width (equal to the class limits), the height of each corresponding to the number (or percentage) of cases in the interval. Chapter 2

Hypothesis. A specific statement, derived from a theory, about the relationship between variables. Chapter 1

Hypothesis testing. Statistical tests that estimate the probability of sample outcomes if assumptions about the population (the null hypothesis) are true. Chapter 7

Independence. The null hypothesis in the chi square test. Two variables are independent if, for all cases, the classification of a case on one variable has no effect on the probability that the case will be classified in any particular category of the second variable. Chapter 10

Independent random samples. Random samples gathered in such a way that the selection of a particular case for one sample has no effect on the probability that any other particular case will be selected for the other samples. Chapter 8

Independent variable. A variable that is identified as a cause. The independent variable is thought to cause the dependent variable. Chapters 1 and 11

Inferential statistics. The branch of statistics concerned with making generalizations from samples to populations. Chapter 1

Interquartile range (*Q*). The distance from the third quartile to the first quartile. Chapter 4

Intervening relationship. A multivariate relationship wherein a bivariate relationship becomes substantially weaker after a third variable is controlled for. The independent and dependent variables are linked primarily through the control variable. Chapter 13

Lambda (λ). A proportional reduction in error (PRE) measure for nominally measured variables that have been organized into a bivariate table. Chapter 11

Level of measurement. The mathematical characteristic of a variable and the major criterion for selecting statistical techniques. Variables can be measured at any of three levels, each permitting certain mathematical operations and statistical techniques. The characteristics of the three levels are summarized in Table 1.5. Chapter 1

Line chart. A graph used for interval-ratio variables. Class intervals are represented by dots placed over the midpoints, the height of each corresponding to the number (or percentage) of cases in the interval. All dots are connected by straight lines. Chapter 2

Linear relationship. A relationship between two variables in which the observation points (dots) in the scatterplot can be approximated with a straight line. Chapter 12

Marginals. The row and column subtotals in a bivariate table. Chapter 10

Maximum difference. A way to assess the strength of an association between variables that have been organized into a bivariate table. The maximum difference is the largest difference between column percentages for any row of the table. Chapter 11

Mean. The arithmetic average of the scores. $\overline{X}$ represents the mean of a sample, and μ, the mean of a population. Chapter 3

Mean square estimate. An estimate of the variance calculated by dividing the sum of squares within (SSW) or the sum of squares between (SSB) by the proper degrees of freedom. Chapter 9

Measures of association. Statistics that summarize the strength and direction of the relationship between variables. Chapters 1 and 11

Measures of central tendency. Statistics that summarize a distribution of scores by reporting the most typical or representative value of the distribution. Chapter 3

Measures of dispersion. Statistics that indicate the amount of variety, or heterogeneity, in a distribution of scores. Chapter 4

Median (Md). The point in a distribution of scores above and below which exactly half of the cases fall. Chapter 3

Midpoint. The point exactly halfway between the upper and lower limits of a class interval. Chapter 2

Mode. The most common value in a distribution or the largest category of a variable. Chapter 3

μ. The mean of a population. Chapter 6

μ_p. The mean of a sampling distribution of sample proportions. Chapter 6

$\mu_{\overline{X}}$. The mean of a sampling distribution of sample means. Chapter 6

Multiple correlation. A multivariate technique for examining the combined effects of more than one independent variable on a dependent variable. Chapter 13

Multiple correlation coefficient (R). A statistic that indicates the strength of the correlation between a dependent variable and two or more independent variables. Chapter 13

Multiple regression. A multivariate technique that breaks down the separate effects of the independent variables on the dependent variable; used to make predictions of the dependent variable. Chapter 13

N_d. In a bivariate table, the number of pairs of cases ranked in different order on two variables. Chapter 11

N_s. In a bivariate table, the number of pairs of cases ranked in the same order on two variables. Chapter 11

Negative association. A bivariate relationship where the variables vary in opposite directions. As one variable increases, the other decreases, and high scores on one variable are associated with low scores on the other. Chapter 11

Nonparametric test. A "distribution-free" test. These tests do not assume a normal sampling distribution. Chapter 10

Nonprobability sample. Any sample that does not meet the EPSEM criterion. These samples have a variety of uses in social science research but cannot be used to generalize to a population. Chapter 6

Normal curve. A theoretical distribution of scores that is symmetrical, unimodal, and bell-shaped. The standard normal curve always has a mean of 0 and a standard deviation of 1. Chapter 5

Normal curve table. Appendix A; a detailed description of the area between a Z score and the mean of any standardized normal distribution. Chapter 5

Null hypothesis (H_0). A statement of "no difference." In the context of single-sample tests of significance, the null hypothesis states that the population from which the sample was drawn has a certain characteristic or value. Chapter 7

Observed frequency (f_o). The cell frequencies actually observed in a bivariate table. Chapter 10

One-tailed test. A type of hypothesis test used when (1) the direction of the difference can be predicted or (2) concern focuses on outcomes in only one tail of the sampling distribution. Chapter 7

One-way analysis of variance. Applications of ANOVA in which the effect of a single independent variable on a dependent variable is observed. Chapter 9

Outliers. Scores that are very high or very low compared to most scores. Chapter 3.

P_s. Any sample proportion. Chapter 6

P_u. Any population proportion. Chapter 6

Parameter. A characteristic of a population. Chapter 6

Partial correlation. A multivariate technique for examining a bivariate relationship while controlling for other variables. Chapter 13

Partial correlation coefficient. A statistic that shows the relationship between two variables while controlling for other variables; $r_{yx.z}$ is the symbol for the partial correlation coefficient when controlling for one variable. Chapter 13

Partial slope. In a multiple regression equation, the slope of the relationship between a particular independent variable and the dependent variable while controlling for all other independent variables in the equation. Chapter 13

Pearson's *r*. A measure of association for variables that have been measured at the interval-ratio level. Chapter 12

Percentage. The number of cases in a category of a variable divided by the number of cases in all categories of the variable, the entire quantity multiplied by 100. Chapter 2

Percent change. A statistic that expresses the magnitude of change in a variable from time 1 to time 2. Chapter 2

Phi. A chi square–based measure of association for nominally measured variables that have been organized into a bivariate table with two rows and two columns (a 2×2 table). Chapter 11

Pie chart. A graph used for nominal and ordinal variables with only a few categories. A circle (the pie) is divided into segments proportional in size to the percentage of cases in each category of the variable. Chapter 2

Pooled estimate. An estimate of the standard deviation of the sampling distribution of the difference in sample means based on the standard deviations of both samples. Chapter 8

Population. The total collection of all cases in which the researcher is interested. Chapter 1

Positive association. A bivariate relationship where the variables vary in the same direction. As one variable increases, the other also increases, and high scores on one variable are associated with high scores on the other. Chapter 11

Proportion. The number of cases in one category of a variable divided by the number of cases in all categories of the variable. Chapter 2

Proportional reduction in error (PRE). A logic used in the design of certain measures of association. A PRE measure indicates how much knowledge of the independent variable improves our predictions of the dependent variable. Chapter 11

Quantitative research. Research projects that collect data or information in the form of numbers. Chapter 1

Range (*R*). The highest score minus the lowest score. Chapter 4

Rate. The number of actual occurrences of some phenomenon or trait divided by the number of possible occurrences per some unit of time. Rates are usually multiplied by some power of 10. Chapter 2

Ratio. The number of cases in one category divided by the number of cases in some other category. Chapter 2

Regression line. The single best-fitting straight line that summarizes the relationship between two variables. Regression lines are fitted to the data points by the least-squares criterion, whereby the line touches all conditional means of *Y* or comes as close to doing so as possible. Chapter 12

Representative sample. A sample that reproduces the characteristics of the population from which it was drawn. Chapter 6

Research. Any process of gathering information systematically and carefully to answer questions or test theories. Statistics are useful for research projects that collect numerical information or data. Chapter 1

Research hypothesis (H_1). A statement that contradicts the null hypothesis. In the context of single-sample tests of significance, the research hypothesis says that the population from which the sample was drawn does not have a certain characteristic or value. Chapter 7

Row. The horizontal dimension of a bivariate table, conventionally representing a score on the dependent variable. Chapter 10

Sample. A carefully chosen subset of a population. In inferential statistics, information is gathered from a sample and then generalized to a population. Chapter 1

Sampling distribution. The distribution of a statistic for all possible sample outcomes of a certain size. Under conditions specified in two theorems, the sampling distribution will be normal in shape, with a mean equal to the population value and a standard deviation equal to the population standard deviation divided by the square root of *N*. Chapter 6

Scatterplot. A graph that depicts the relationship between two variables. Chapter 12

Σ (uppercase Greek letter sigma). "The summation of." Chapter 3

σ_{p-p}. Symbol for the standard deviation of the sampling distribution of the differences in sample proportions. Chapter 8

$\sigma_{\bar{X}-\bar{X}}$. Symbol for the standard deviation of the sampling distribution of the differences in sample means. Chapter 8

Significance testing. See **Hypothesis testing**. Chapter 7

Simple random sample. A method for choosing cases from a population by which every case and every combination of cases has an equal chance of being included. Chapter 6

Skew. The extent to which a distribution of scores has a few scores that are extremely high (positive skew) or extremely low (negative skew). Chapter 3

Slope (*b*). The amount of change in one variable per unit change in the other; *b* is the symbol for the slope of a regression line. Chapter 12

Spearman's rho (r_s). A measure of association for "continuous" ordinal variables. Chapter 11

Spurious relationship. A multivariate relationship in which a bivariate relationship becomes substantially weaker after a third variable is controlled for. The independent and dependent variables are not causally linked. Rather, both are caused by the control variable. Chapter 13

Standard deviation. The statistic computed by summing the squared deviations of the scores around the mean, dividing by *N*, and, finally, taking the square root of the result. The most important and useful descriptive measure of dispersion; *s* represents the standard deviation of a sample; σ represents the standard deviation of a population. Chapter 4

Standard error of the mean. The standard deviation of a sampling distribution of sample means. Chapter 6

Standardized partial slope (beta-weight). The slope of the relationship between a particular independent variable and the dependent variable when all scores have been normalized. Chapter 13

Stated class limits. The class intervals of a frequency distribution when stated as discrete categories. Chapter 2

Statistics. A set of mathematical techniques for organizing and analyzing data. Chapter 1

Student's *t* distribution. A distribution used to find the critical region for tests of sample means when *s* is unknown and sample size is small. Chapter 7

Sum of squares between (SSB). The sum of the squared deviations of the sample means from the overall mean, weighted by sample size. Chapter 9

Sum of squares within (SSW). The sum of the squared deviations of scores from the category means. Chapter 9

***t*(critical).** The *t* score that marks the beginning of the critical region of a *t* distribution. Chapter 7

***t*(obtained).** The test statistic computed in step 4 of the five-step model. The sample outcome expressed as a *t* score. Chapter 7

Test statistic. The value computed in step 4 of the five-step model that converts the sample outcome into either a *t* score or a *Z* score. Chapter 7

Theory. A generalized explanation of the relationship between two or more variables. Chapter 1

Total sum of squares (SST). The sum of the squared deviations of the scores from the overall mean. Chapter 9

Total variation. The spread of the *Y* scores around the mean of *Y*. Equal to $\sum(Y - \bar{Y})^2$. Chapter 12

Two-tailed test. A type of hypothesis test used when (1) the direction of the difference cannot be predicted or (2) concern focuses on outcomes in both tails of the sampling distribution. Chapter 7

Type I error (alpha error). The probability of rejecting a null hypothesis that is, in fact, true. Chapter 7

Type II error (beta error). The probability of failing to reject a null hypothesis that is, in fact, false. Chapter 7

Unexplained variation. The proportion of the total variation in *Y* that is not accounted for by *X*. Equal to $\sum(Y - Y')^2$. Chapter 12

Variable. Any trait that can change values from case to case. Chapter 1

Variance. The sum of the squared deviations of the scores around the mean, divided by *N*. A measure of dispersion used primarily in inferential statistics and also in correlation and regression techniques; s^2 represents the variance of a sample; σ^2 represents the variance of a population. Chapter 4

***X*.** Symbol used for any independent variable. Chapter 11

X_i. Any score in a distribution. Chapter 3

***Y*.** Symbol used for any dependent variable. Chapter 11

***Y* intercept (*a*).** The point where the regression line crosses the *Y* axis. Chapter 12

Y'. Symbol for predicted score on *Y*. Chapter 12

***Z*.** Symbol for any control variable. Chapter 13

***Z* scores.** Standard scores; the way scores are expressed after they have been standardized to the theoretical normal curve. Chapter 5

***Z*(critical).** The *Z* score that marks the beginning of the critical region on a *Z* distribution. Chapter 7

***Z*(obtained).** The test statistic computed in step 4 of the five-step model. The sample outcomes expressed as a *Z* score. Chapter 7

Zero-order correlation. Correlation coefficient for bivariate relationships. Chapter 13

Index