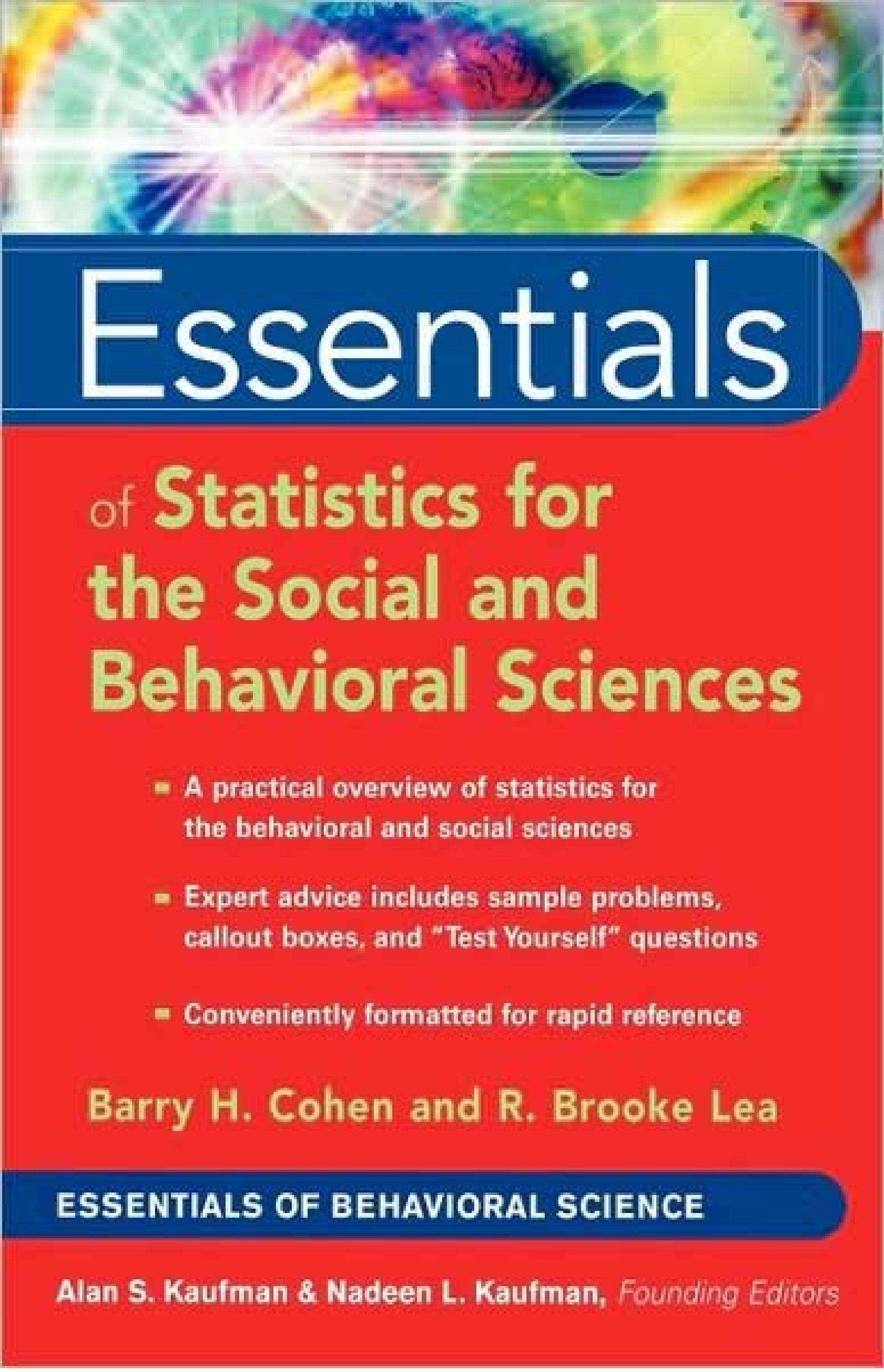
Essentials
of Statistics for the Social and Behavioral Sciences
A practical overview of statistics for the behavioral and social sciences
Expert advice includes sample problems, callout boxes, and "Test Yourself" questions
Conveniently formatted for rapid reference
Barry H. Cohen and R. Brooke Lea
ESSENTIALS OF BEHAVIORAL SCIENCE
Alan S. Kaufman & Nadeen L. Kaufman, Founding Editors

Essentials

of Statistics for the Social and Behavioral Sciences

Barry H. Cohen
R. Brooke Lea

John Wiley & Sons, Inc.

Essentials of Statistics for the Social and Behavioral Sciences

Essentials of Behavioral Science Series

Founding Editors, Alan S. Kaufman and Nadeen L. Kaufman

Essentials of Statistics for the Social and Behavioral Sciences
by Barry H. Cohen and R. Brooke Lea

Essentials of Psychological Testing
by Susana Urbina

Essentials of Research Design and Methodology
by Geoffrey R. Marczyk and David DeMatteo

Essentials

of Statistics for the Social and Behavioral Sciences

Barry H. Cohen
R. Brooke Lea

John Wiley & Sons, Inc.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Cohen, Barry H., 1949–
Essentials of statistics for the social and behavioral science / Barry H. Cohen, R. Brooke Lea.
p. cm. — (Essentials of behavioral sciences series)
Includes bibliographical references and index.
ISBN 0-471-22031-0 (pbk. : alk. paper)
1. Social sciences—Statistical methods. I. Lea, R. Brooke. II. Title. III. Series.

HA29.C65 2003
519.5—dc21 2003049669

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To my dear Aunts: Harriet Anthony and Diana Franzblau

BHC

To Emily and Jackson, the two parameters that keep me normal

RBL

We would like to sincerely thank Irving B. Weiner, Ph.D., ABPP for his assistance as a consulting editor on this project.

Dr. Weiner completed his doctoral studies at the University of Michigan in 1959 and went on to write and edit over 20 books, as well as countless chapters and journal articles. A Diplomate of the American Board of Professional Psychology in both Clinical and Forensic Psychology, he currently serves as Clinical Professor of Psychiatry and Behavioral Medicine at the University of South Florida. Dr. Weiner serves as Chairman of the Wiley Behavioral Sciences Advisory Board and is Editor-in-Chief of the 12-volume Handbook of Psychology, which published in December 2002.

CONTENTS

SERIES PREFACE

In the *Essentials of Behavioral Science* series, our goal is to provide readers with books that will deliver key practical information in an efficient, accessible style. The series features books on a variety of topics, such as statistics, psychological testing, and research design and methodology, to name just a few. For the experienced professional, books in the series offer a concise yet thorough review of a specific area of expertise, including numerous tips for best practices. Students can turn to series books for a clear and concise overview of the important topics in which they must become proficient to practice skillfully, efficiently, and ethically in their chosen fields.

Wherever feasible, visual cues highlighting key points are utilized alongside systematic, step-by-step guidelines. Chapters are focused and succinct. Topics are organized for an easy understanding of the essential material related to a particular topic. Theory and research are continually woven into the fabric of each book, but always to enhance the practical application of the material, rather than to sidetrack or overwhelm readers. With this series, we aim to challenge and assist readers in the behavioral sciences to aspire to the highest level of competency by arming them with the tools they need for knowledgeable, informed practice.

Essentials of Statistics for the Social and Behavioral Sciences concentrates on drawing connections among seemingly disparate statistical procedures and providing intuitive explanations for how the basic formulas work. The authors weave statistical concepts together and thus make the different procedures seem less arbitrary and isolated. The statistical procedures covered here are those considered *essential* to researchers in the field. Only univariate statistics are presented; topics in multivariate statistics (including multiple regression) deserve a separate volume of their own. Further, this book assumes that the reader has a working knowledge of basic statistics or has ready access to an introductory text. Therefore, this book will not bog down the reader down with computational details. Thus, this book should be ideal as a supplementary text for students struggling to understand the material in an advanced (or sophisticated) undergraduate statistics course, or an intermediate course at the master's level. *Essentials of Statistics* is also ideal for researchers in the social and behavioral sciences who have forgotten some of their statistical training and need to brush up on statistics in order to evaluate data, converse knowledgeably with a statistical consultant, or prepare for licensing exams.

Chapter 1 covers the most often used methods of descriptive statistics, and the next four chapters cover the basics of null hypothesis testing and interval estimation for the one-, two-, and multigroup cases, as well as the case of two continuous variables. Chapter 6 is devoted to the increasingly essential topics of power analysis and effect size estimation for the cases covered in Chapters 2 through 5. Chapters 7 and 8 deal with the complex forms of analysis of variance common in experimental social science research. As appropriate, these chapters include material relevant to the larger topic of research design. Finally, Chapter 9 includes some of the most popular methods in nonparametric statistics. Regrettably, many useful topics had to be omitted for lack of space, but the references and annotated bibliography point the reader toward more comprehensive and more advanced texts to fill any gaps. Indeed, we hope that this book will help the reader understand those more advanced sources. Additional material to help readers of this book understand the statistical topics covered in this book, as well as some related and more advanced topics, are posted on the web and can be accessed by following links from *www.psych.nyu.edu/people/faculty/cohen/statstext.html.*

Alan S. Kaufman, PhD, and Nadeen L. Kaufman, EdD, Founding Editors
Yale University School of Medicine

Essentials of Statistics for the Social and Behavioral Sciences

One

DESCRIPTIVE STATISTICS

Social and behavioral scientists need statistics more than most other scientists, especially the kind of statistics included in this book. For the sake of contrast, consider the subject matter of physics. The nice thing about protons and electrons, for instance, is that all protons have the same mass; electrons are a lot lighter, but they also are all identical to each other in mass. This is not to imply that physics is easier than any of the social or behavioral sciences, but the fact that animals and especially humans vary so much from each other along every conceivable dimension creates a particular need to summarize all this variability in order to make sense of it.

The purpose of descriptive statistics is to use just a few numbers to capture the meaning of a much larger collection of observations on many different cases. These cases could be people, animals, or even cities or colleges; or the same cases on many different occasions; or some combination of the two. Often, computing descriptive statistics is just your first step in a process that uses more advanced statistical methods to make estimates about cases that you will never have the opportunity to measure directly. This chapter will cover only descriptive statistics. The remaining chapters will be devoted to more advanced methods called *inferential* statistics.

SAMPLES AND POPULATIONS

Sometimes you have all of the observations in which you are interested, but this is rare. For instance, a school psychologist may have scores on some standardized test for every sixth grader in Springfield County and her only concern is studying and comparing students within the County. These test scores would be thought of as her *population.* More often, you have just a subset of the observations in which you are interested. For instance, a market researcher randomly selects and calls 100 people in Springfield County and asks all of them about their use of the Internet. The 100 observations obtained (Springfield residents are very cooperative) do not include all of the individuals in which the researcher is interested. The

DON'T FORGET

When Will I Use the Statistics in This Chapter?

You have measured the same variable many times, perhaps on many different people, or many different rats, or many different cities (e.g., the total budget for each city), and so on, and now you want to summarize all of those numbers in a compact and descriptive way. If you want to extrapolate from those numbers to cases you have not measured yet, you will need the tools that we will begin to describe in Chapter 2.

100 observations would be thought of as a *sample* of a larger population.

If as a researcher you are interested in the Internet habits of people in Springfield County, your population consists of all the people in that county. If you are really interested in the Internet habits of people in the United States, then that is your population. In the latter case your sample may not be a good representation of the population. But for the purposes of descriptive statistics, populations and samples are dealt with in similar ways. The distinction between *sample* and *population* will become important in the next chapter, when we introduce the topic of *inferential statistics.* For now, we will treat any collection of numbers that you have as a population.

The most obvious descriptive statistic is one that summarizes all of the observations with a single number—one that is the most typical or that best locates the middle of all the numbers. Such a statistic is called a measure of *central tendency.* The best-known measure of central tendency is the arithmetic mean: the statistic you get if you add up all the scores in your sample (or population) and divide by the number of different scores you added. When people use the term *mean* you can be quite sure that they are referring to the arithmetic mean. There are other statistics that are called means; these include the geometric and the harmonic mean (the latter will be discussed in Chapter 5). However, whenever we use the term *mean* by itself we will be referring to the arithmetic mean. Although the mean is calculated the same way for a sample as a population, it is symbolized as $\overline{X}$ (pronounced "X bar") or *M* when it describes a sample, and μ (the lowercase Greek letter mu; pronounced "myoo") when it describes a population. In general, numbers that summarize the scores in a sample are called statistics (e.g., $\overline{X}$ is a statistic), whereas numbers that summarize an entire population are called parameters (e.g., μ is a parameter).

SCALES OF MEASUREMENT

When we calculate the mean for a set of numbers we are assuming that these numbers represent a precise scale of measurement. For instance, the average of 61 inches and 63 inches is 62 inches, and we know that 62 is exactly in the middle of 61 and 63 because an inch is always the same size (the inch that's between 61 and 62 is precisely the same size as the inch between 62 and 63). In this case we can say that our measurement scale has the *interval* property. This property is necessary to justify and give meaning to calculating means and many other statistics on the measurements that we have. However, in the social sciences we often use numbers to measure a variable in a way that is not as precise as measuring in inches. For instance, a researcher may ask a student to express his or her agreement with some political statement (e.g., I think U.S. senators should be limited to two 6-year terms) on a scale that consists of the following choices: 1 = strongly disagree; 2 = somewhat disagree; 3 = neutral; 4 = somewhat agree; 5 = strongly agree. [This kind of scale is called a *Likert scale,* after its inventor, Rensis Likert (1932).]

Ordinal Scales

You might say that a person who strongly agrees and one who is neutral, when averaged together, are equivalent to someone who somewhat agrees, because the mean of 1 and 3 is 2. But this assumes that "somewhat agree" is just as close to "strongly agree" as it is to neutral—that is, that the intervals on the scale are all equal. All we can really be sure of in this case is the order of the responses—that as the responses progress from 1 to 5 there is more agreement with the statement. A scale like the one described is therefore classified as an *ordinal scale.* The more points such a scale has (e.g., a 1 to 10 rating scale for attractiveness), the more likely social scientists are to treat the scale as though it were not just an ordinal scale, but an interval scale, and therefore calculate statistics such as the mean on the numbers that are reported by participants in the study. In fact, it is even common to treat the numbers from a 5-point Likert scale in that way, even though statisticians argue against it. This is one of many areas in which you will see that common practice among social scientists does not agree with the recommendations of many statisticians (and measurement experts) as reported in textbooks and journal articles.

Another way that an ordinal scale arises is through ranking. A researcher observing 12 children in a playground might order them in terms of aggressiveness, so that the most aggressive child receives a rank of 1 and the least aggressive gets a 12. One cannot say that the children ranked 1 and 2 differ by the same amount as the children ranked 11 and 12; all you know is that the child ranked 5, for instance, has been judged more aggressive than the one ranked 6. Sometimes measurements that come from an interval scale (e.g., time in seconds to solve a puzzle) are converted to ranks, because of extreme scores and other problems (e.g., most participants solve the puzzle in about 10 seconds, but a few take several minutes). There is a whole set of procedures for dealing with ranked data, some of which are described in Chapter 9. Some statisticians would argue that these rank-order statistics should be applied to Likert-scale data, but this is rarely done for reasons that will be clearer after reading that chapter.

Nominal Scales

Some of the distinctions that social scientists need to make are just qualitative—they do not have a quantitative aspect, so the categories that are used to distinguish people have no order, let alone equal intervals. For instance, psychiatrists diagnose people with symptoms of mental illness and assign them to a category. The collection of all these categories can be thought of as a *categorical* or *nominal scale* (the latter name indicates that the categories have names rather than numbers) for mental illness. Even when the categories are given numbers (e.g., the *Diagnostic and Statistical Manual of Mental Disorders* used by psychologists and psychiatrists has a number for each diagnosis), these numbers are not meant to be used mathematically (e.g., it doesn't make sense to add the numbers together) and do not even imply any ordering of the categories (e.g., according to the *Diagnostic and Statistical Manual of Mental Disorders,* fourth edition [*DSM-IV*], Obsessive-Compulsive Disorder is 300.3, and Depressive Disorder is 311; but the diagnostic category for someone suffering from Obsessive-Compulsive Disorder *and* Depressive Disorder is not 611.3, nor is it 305.65, the sum and mean of the categories, respectively).

Although you cannot calculate statistics such as the mean when dealing with categorical data, you can compare frequencies and percentages in a useful way. For instance, the percentages of patients that fall into each *DSM-IV* diagnosis can be compared from one country to another to see if symptoms are interpreted differently in different cultures, or perhaps to see if people in some countries are more susceptible to some forms of mental illness than the people of other countries. Statistical methods for dealing with data from both categorical and ordinal scales will be described in Chapter 9.

Ratio Scales

The three scales of measurement described so far are the nominal (categories that have no quantitative order), the ordinal (the values of the scale have an order, but the intervals may not be equal), and the interval scale (a change of one unit on the scale represents the same amount of change anywhere along the scale). One scale we have not yet mentioned is the *ratio scale.* This is an interval scale that has a true zero point (i.e., zero on the scale represents a total absence of the variable being

Rapid Reference 1.1

Measurement Scales

Nominal: Observations are assigned to categories that differ qualitatively but have no quantitative order (e.g., depressed, phobic, obsessive, etc.).

Ordinal: The values have an order that can be represented by numbers, but the numbers cannot be used mathematically, because the intervals may not be equal (e.g., assigning ranks according to the ability of gymnasts on a team).

Interval: One unit on this scale is the same size anywhere along the scale, so values can be treated mathematically (e.g., averaged), but zero on the scale does not indicate a total absence of the variable being measured (e.g., IQ scores).

Ratio: This scale has the interval property plus the zero point is not arbitrary; it represents a true absence of the variable being measured. For instance, weight in pounds has this property, so that if object A is measured as twice as many pounds as object B, then object A has twice as much weight. (You cannot say that someone with an IQ of 120 is twice as smart as someone with an IQ of 60.)

measured). For instance, neither the Celsius nor Fahrenheit scales for measuring temperature qualify as ratio scales, because both have arbitrary zero points. The Kelvin temperature scale is a ratio scale because on that scale zero is absolute zero, the point at which all molecular motion, and therefore all heat, ceases. The statistical methods described in this book do not distinguish between the interval and ratio scales, so it is common to drop the distinction and refer to interval/ratio data. A summary of the different measurement scales is given in Rapid Reference 1.1.

DISPLAYING YOUR DATA

When describing data there are many options for interval/ratio data, such as the mean, but relatively few options for nominal or ordinal data. However, regardless of the scale you are dealing with, the most basic way to look at your data is in terms of frequencies.

Bar Charts

If you have nominal data, a simple *bar chart* is a good place to start. Along a horizontal axis you write out the different categories in any order that is convenient. The height of the bar above each category should be proportional to the number of your cases that fall into that category. If 20 of the patients you studied were phobic and 10 were depressed, the vertical bar rising above "phobic" would be twice as high as the bar above "depressed." Of course, the chart can be rotated to make the bars horizontal, or a pie chart or some other display can be used instead, but the bar chart is probably the most common form of display for nominal data in the social sciences.

Because the ordering of the categories in a bar chart of nominal data is arbitrary, it doesn't quite make sense to talk of the central tendency of the data. However, if you want to talk about the most typical value, it makes some sense to identify the category that is the most popular (i.e., the one with the highest bar). The category with the highest frequency of occurrence is called the *mode.* For instance, among patients at a psychiatric hospital the modal diagnosis is usually schizophrenia (unless this category is broken into subtypes).

The bar chart is also a good way to display data from an ordinal scale, but because the values now have an order, we can talk meaningfully about central tendency. You can still determine the mode—the value with the highest bar (i.e., frequency)—but the mode need not be near the middle of your bar chart (although it usually will be). However, with an ordinal scale you can add up frequencies and percentages in a way that doesn't make sense with a nominal scale. First, let us look at the convenience of dealing with percentages.

Percentile Ranks and the Median

Suppose 44 people in your sample "strongly agree" with a particular statement; this is more impressive in a sample of 142 participants than in a sample of 245 participants (note: in keeping with recent custom in the field of psychology, we will usually use the term *participant* to avoid the connotations of the older term *subject*). The easiest way to see that is to note that in the first case the 44 participants are 31% of the total sample; in the second case, they are only 18%. The percentages make sense without knowing the sample size. Percentages are useful with a nominal scale (e.g., 45% of the patients were schizophrenic), but with an ordinal scale there is the added advantage that the percentages can be summed. For example, suppose that 100 people respond to a single question on a Likert scale with the following percentages: 5% strongly disagree; 9% somewhat disagree; 36% are neutral; 40% agree; and 10% strongly agree. We can then say that 14% (5 + 9) of the people are on the disagree side, or that 14% are below neutral (it's arbitrary, but we are assigning higher values in the agree direction).

We can assign a *percentile rank* (PR) to a value on the scale such that the PR equals the percentage of the sample (or population) that is at or below that value. The PR is 5 for strongly disagree, 14 for somewhat disagree, 50 for neutral, 90 for agree, and 100 for strongly agree (it is always 100, of course, for the highest value represented in your set of scores). A particularly useful value in any set of scores is called the *median.* The median is defined as the middle score, such that half the scores are higher, and half are lower. In other words, the median is the value whose PR is 50. In this example the median is "neutral." The median is a useful measure of central tendency that can be determined with an ordinal, but not a nominal, scale. According to this definition, the median in the preceding example would be somewhere between "neutral" and "somewhat agree." If "neutral" is 3 and "somewhat" agree is 4 on the scale, then some researchers would say that the median is 3.5. But unless you are dealing with an interval scale you cannot use the numbers of your scale so precisely. If all your scores are different, it is easy to see which score is the middle score. If there are only a few different scores (e.g., 1 to 5) but many responses, there will be many scores that are tied, making it less clear which score is in the middle.

Histograms

A slight modification of the bar chart is traditionally used when dealing with interval/ratio data. On a bar chart for nominal or ordinal data there should be some space between any two adjacent bars, but for interval/ratio data it is usually appropriate for each bar to touch the bars on either side of it. When the bars touch, the chart is called a *histogram.* To understand when it makes sense for the bars to touch, you need to know a little about *continuous* and *discrete* scales, and therefore something about discrete and continuous variables. A variable is discrete when it can only take certain values, with none between. Appropriately, it is measured on a discrete scale (whole numbers—no fractions allowed). For example, family size is a discrete variable because a family can consist of three or four or five members, but it cannot consist of 3.76 members.

Height is a continuous variable because for any two people (no matter how close in height) it is theoretically possible to find someone between them in height. So height should be measured on a continuous scale (e.g., number of inches to as many decimal places as necessary). Of course, no scale is perfectly continuous (infinitely precise), but measuring height in tiny fractions of inches can be considered continuous for our purposes. Note that some continuous variables cannot at present be measured on a continuous scale. A variable like charisma may vary continuously, but it can only be measured with a rather crude, discrete scale (e.g., virtually no charisma, a little charisma, moderate charisma, etc.). Data from a continuous scale are particularly appropriate for a histogram.

Consider what a histogram might look like for the heights of 100 randomly selected men (for simplicity, we will look at one gender at a time). If the men range from 62 to 76 inches, the simplest scheme would be to have a total of 15 bars, the first ranging from 61.5 to 62.5 inches, the second from 62.5 to 63.5 inches, and so on until the 15th bar, which goes from 75.5 to 76.5 inches. Looking at Figure 1.1, notice how the bars are higher near the middle, as is the case for many variables (the mode in this case is 69 inches). Now suppose that these men range in weight from 131 to 218 pounds. One bar per pound would require 88 bars (218 – 131 + 1), and many of the bars (especially near either end) would be empty. The solution is to group together values into class intervals. For the weight example, 10-pound intervals starting with 130–139 and ending with 210–219 for a total of nine intervals would be reasonable. A total of eighteen 5-pound intervals (130–134 to 215–219) would give more detail and would also be reasonable. The common guidelines are to use between 10 and 20 intervals, and when possible to start or end the intervals with zeroes or fives (e.g., 160–164 or 161–165).

Note that if you look at what are called the *apparent limits* of two adjacent class intervals, they don't appear to touch—for example, 130–134 and 135–139. However, measurements are being rounded off to the nearest unit, so the *real limits* of the intervals just mentioned are 129.5–134.5 and 134.5–139.5, which obviously do touch. We don't worry about anyone who is exactly 134.5 pounds; we just

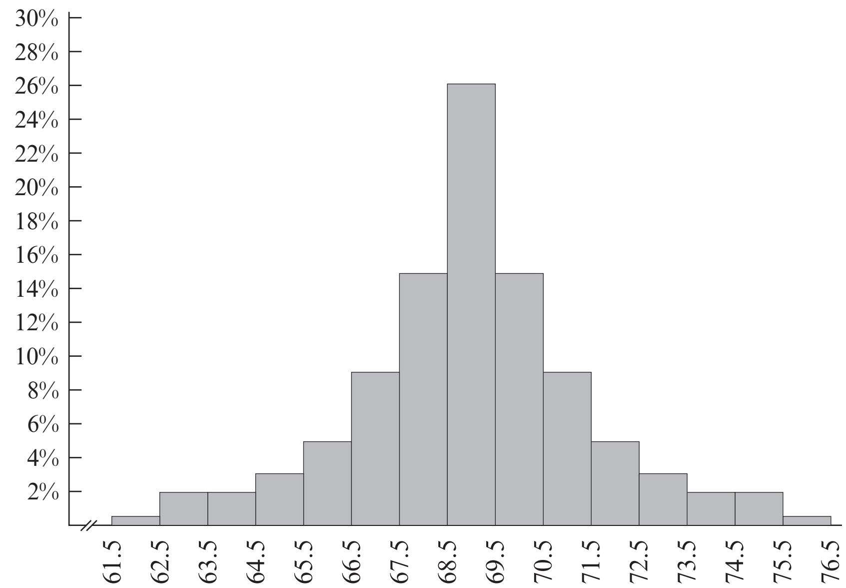

Figure 1.1 A histogram of the heights (in inches) of 100 randomly selected men

assume that if we measure precisely enough, that person will fall into one interval or the other.

DON'T FORGET

If you are dealing with nominal (i.e., categorical) or ordinal data, a bar chart is appropriate (the bars do not touch). If you are dealing with interval or ratio data, a histogram is appropriate; the bars extend to the lower and upper real limits of the interval represented (even if it is a single unit), and therefore adjacent bars do touch.

Percentiles

Percentages can be added, just as with the ordinal scale, to create percentile ranks. For instance, looking at Figure 1.1, we can add the percentages of the first five bars (1 + 2 + 2 + 3 + 5) to find that the PR for 66 inches is 13% (actually 13% is the PR for 66.5 inches, because you have to go to the upper real limit of the interval to ensure that you have surpassed everyone in that interval). Conversely, one can define a *percentile* as a score that has a particular PR. For example, the 22nd percentile is 67 (actually 67.5), because the PR of 67 is 22. The percentiles of greatest interest are the deciles (10%, 20%, etc.), and the quartiles (25%, 50%, 75%).

Unfortunately, these particular percentiles are not likely to fall right in the middle of a bar or right between two bars. For instance, for the data in Figure 1.1, the 1st quartile (25%) is somewhere between 67.5 (PR = 22) and 68.5 (PR = 37). It is common to interpolate linearly between these two points. Because 25 is one fifth of the way from 22 to 37, we say that the 25th percentile is about one fifth of the way from 67.5 to 68.5 or about 67.7. The formula for linear interpolation is given in most introductory statistics texts. Probably the most important percentile of all is the 50th; as we mentioned before, this percentile is called the median. For Figure 1.1, the median is 69.0—that is, half the men have heights below 69.0 inches, and half are taller than 69.0 inches. The mode is the interval represented by 69 inches—that is, 68.5 to 69.5 inches.

Distributions

Figure 1.1 shows you that height is a variable; if it were a constant, all people would have the same height (the number of chambers in the human heart is a constant—everybody has four). Figure 1.1 shows how the values for height are distributed in the sample of 100 men that were measured. A set of values from a variable together with the relative frequency associated with each value is called a *distribution*. Except for the last chapter of the book, all of the statistical methods we will present involve distributions. If all of the heights from 62 to 76 inches were equally represented, all of the bars would be at the same height, and it would