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
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.
<table>
<thead>
<tr>
<th>CONTENTS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Series Preface</td>
<td>ix</td>
</tr>
<tr>
<td>One Descriptive Statistics</td>
<td>1</td>
</tr>
<tr>
<td>Two Introduction to Null Hypothesis Testing</td>
<td>28</td>
</tr>
<tr>
<td>Three The Two-Group $t$ Test</td>
<td>48</td>
</tr>
<tr>
<td>Four Correlation and Regression</td>
<td>71</td>
</tr>
<tr>
<td>Five One-Way ANOVA and Multiple Comparisons</td>
<td>97</td>
</tr>
<tr>
<td>Six Power Analysis</td>
<td>122</td>
</tr>
<tr>
<td>Seven Factorial ANOVA</td>
<td>145</td>
</tr>
<tr>
<td>Eight Repeated-Measures ANOVA</td>
<td>172</td>
</tr>
<tr>
<td>Nine Nonparametric Statistics</td>
<td>199</td>
</tr>
<tr>
<td>Appendix A Statistical Tables</td>
<td>226</td>
</tr>
<tr>
<td>Appendix B Answers to Putting it into Practice Exercises</td>
<td>243</td>
</tr>
<tr>
<td>References</td>
<td>275</td>
</tr>
<tr>
<td>Annotated Bibliography</td>
<td>278</td>
</tr>
</tbody>
</table>
CONTENTS

Index 281
Acknowledgments 291
About the Authors 291
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 mater-
ial 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
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
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 \( \bar{X} \) (pronounced “X bar”) or \( M \) when it describes a sample, and \( \mu \) (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., \( \bar{X} \) is a statistic), whereas numbers that summarize an entire population are called parameters (e.g., \( \mu \) 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 sev-
eral 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
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 se-
lected 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](image-url)
assume that if we measure precisely enough, that person will fall into one interval or the other.

**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
be said that we have a uniform distribution. That form of distribution is not likely when dealing with the variables usually measured for people. Often, the distribution of a variable is shaped something like a bell, as in Figure 1.1, and has one mode somewhere in the middle. Values further from the middle are progressively less popular.

**Shapes of Distributions**

Imagine the distribution of 60 students who took a statistics exam. If the class consisted mostly of math majors and English majors the distribution might have two equally high bars, and therefore two modes—one more to the right for the math majors and one more to the left for the English majors, with a dip in between. This distribution would be called *bimodal* (even if the two modes were not exactly equal in frequency), whereas the distribution in Figure 1.1 is called *unimodal*. It is possible for a distribution to have even more than two modes, but we will be dealing only with unimodal distributions. Now imagine that the statistics exam was very easy (if you can). The scores would be bunched up (producing high bars) in the 90s with relatively few low scores trailing off in the negative direction. The mode would be decidedly to one side (the right, or positive, side in this case), and the distribution would appear to have a tail (a series of relatively low bars) on the left. Such a distribution is said to be *negatively skewed*, because the tail is in the negative direction. This kind of distribution often arises when a large portion of the scores are approaching the highest possible score (i.e., there is a ceiling effect).

Positively skewed distributions are probably more common in the social sciences than those with a negative skew. Annual income is a good example. The majority of people in the United States, for instance, are much closer to the lowest possible income (we’ll say it is zero and ignore the possibility of negative income) than to the highest known income. Clearly, there is a floor for income, but no clearly defined ceiling, so the income distribution has a tail that points in the positive direction. The annual incomes for a randomly selected group of people would therefore be very likely to form a positively skewed distribution, as illustrated in Figure 1.2.

**CHOOSING A MEASURE OF CENTRAL TENDENCY**

One of the most important reasons to draw (or have a computer draw) a histogram is to look at the shape of the distribution with which you are dealing. With a very large sample—and especially with a population—the distribution will be fairly smooth and very likely unimodal with an approximate bell shape. However, the shape may be symmetrical or skewed (either positively or negatively). The shape can be important: For example, strong skewing can affect your choice of
In a symmetrical, unimodal distribution the three measures of central tendency we have described—the mean, the median, and the mode—will all be in the same spot, so it doesn’t matter which you choose. However, in a skewed distribution extreme scores have a larger effect on the mean than on the median, so while both of these measures are pulled away from the mode, the mean is pulled further. This is illustrated in Figure 1.2.

It is easy to understand why the skewing does not move the median much. Although the long positive tail includes some very high values, the tail represents only a small percentage of the sample. Moving the median just a little in the hump of the distribution (where the bars are high) can have a large effect on the percentage on each side of the median. Moving the median a little toward the tail can compensate for the small extra percentage that is contained in the tail. Once a score is to the right of the median, moving it much further to the right has no effect on the median, because that wouldn’t change the fact that 50% of the scores are still on each side of the median. The mean, however, is sensitive to the actual values of all the scores, and a few very large scores on one side of the distribution can noticeably pull the mean to that side. That’s why for some purposes the mean can be considered a misleading measure of central tendency, as we will explain next.

Suppose that Figure 1.2 displays the incomes of employees for one particular company. To make the argument that the employees are well paid, the company president would be happy to report that the mean annual income is $35,000. However, you can see that the vast majority of employees earn less than this amount; the mean is being unduly influenced by the incomes of a relatively few executives at the company. The regular workers of the company would prefer to use the median as a description of the average salary. Whereas the majority of the scores in a distribution can be above or below the mean, the median is always near the middle because 50% of the scores are above and 50% below it.
When a news report refers to an average or mean number, it is usually referring to the arithmetic mean, but read closely: The author could be referring to a median or even a mode or other measure in an imprecise way (the measures of central tendency just described are summarized in Rapid Reference 1.2). However, regardless of which measure of central tendency is being used, you should notice that the wider the distribution, the harder it can be to describe it with just one number: The endpoints of the distribution can be very far from the middle, no matter how the middle is defined. Measuring the width of the distribution can be an important complement to locating the middle. This is our next topic.

**MEASURES OF VARIABILITY**

As a sixth-grade English teacher, which class of 20 students would you rather teach, one whose average reading score is 6.3 (a bit above grade level) or 5.8 (a bit below)? Perhaps you like a challenge, but you would probably guess that the 6.3 class would be easier to teach. But what if we tell you that the students in the “5.8” class range from 5.6 to 6.0, whereas the “6.3” class ranges from 5.7 to 6.9? Given these ranges, the more homogeneous (“5.8”) class would likely be the easier to teach.

**The Range and Semi-Interquartile Range**

The simplest way to measure the width of a distribution is to calculate its range. The range is just the highest minus the lowest score, plus one unit if you are dealing with a continuous scale (e.g., the range of the 5.8 class is 6.0 – 5.6 + .1 = .4 + .1 = .5, because the upper limit of 6.0 is really 6.05 and the lower real limit of 5.6 is 5.55). The problem with the range is that it can be dramatically influenced by one extreme score. Add a 7.0 reader to the 5.8 class and the 5.8 class will now have a larger range than the 6.3 class. However, the range of the 5.8 class would then be misleading; it is still a very homogeneous class, with just one very advanced student who needs to be dealt with separately.

One way to modify the range so that it is not affected by extreme scores is to measure the range of the middle 50% of the scores. This modified range is found by subtracting the 25th percentile of the distribution from the 75th percentile. Hence, it is called the interquartile range. If you divide this range by 2, you get a measure called the semi-interquartile range (SIQ), which is roughly the average of the distances from the median to the 25th and 75th percentiles. The SIQ gives you a typical amount by which scores tend to differ from the median (about half are closer and half are further away than the SIQ), and this is one very useful way to describe the variability of a distribution. The SIQ range can be very useful for descriptive purposes, especially when dealing with ordinal data or with a distribution that has
extreme scores on one or both sides of its median. Measures that make use of all of the scores at hand are usually more useful for describing the spread of the scores when you want to extrapolate from your sample to a larger population. We will describe such a measure shortly.

**The Summation Sign**

An obvious way to measure the amount of variability in a distribution is to find the distance of each score from some measure of central tendency, and then average these differences together to find a typical amount of deviation from the middle. If your variability measure will use all of your scores it makes sense to anchor it to a measure of central tendency that does the same—that is, the mean. Expressed in words, we can propose a measure of variability that is equal to the average of all of the scores’ deviations from the mean. At this point, mathematical notation, which so many students find annoying, can be really helpful in defining complex statistics in a compact and unambiguous way. The uppercase Greek letter sigma (\(\Sigma\)) is often used as a way of telling you to add your scores together; it is therefore called, in this context, the *summation sign*. If you follow the summation sign with a letter representing the variable you are measuring (e.g., \(\Sigma X\)), this is a shorthand way of telling you to add all of your scores together. This notation allows us to write a very simple formula for the mean of a set of scores:

\[
\mu = \frac{\sum X_i}{N}
\]  

(1.1)

The subscript \(i\) associated with \(X\) is there to remind you that there is more than just one \(X\); there are a whole series of values to be added up. Statistical purists would like us to put “\(i = 1\)” under the summation sign and \(N\) above it (to remind you to start adding with the first score and not to stop until you have added the \(N\)th score), but we will always use \(\Sigma\) to mean “add them all up,” so that extra notation won’t be necessary. Note that Formula 1.1 is a very convenient way of saying that if you add up all of your scores, and then divide by the number (\(N\)) of scores that you added, the result will equal the mean.

**The Mean Deviation**

Next, we can apply Formula 1.1 to the deviations of scores from the mean rather than to the scores themselves. This can be expressed symbolically as follows:

\[
\frac{\sum (X_i - \mu)}{N}
\]
The problem with the above expression is that it is always equal to zero. This is actually an important property of the mean—that it is a balance point in any distribution, such that the sum of deviations above it equals the sum of deviations below it. However, if we want to know the average distance of scores from the mean we are not concerned with the sign of a deviation, just its magnitude. That idea can be expressed mathematically in the following formula:

$$\text{MD} = \frac{\sum |X_i - \mu|}{N}$$  \hspace{1cm} (1.2)

MD stands for the mean deviation, and the vertical bars around $X_i - \mu$ tell us to take the absolute value of the deviation. Since the deviations are now all positive, they don’t cancel each other out, and we are left with a number that is literally the average of the absolute deviations from the mean. The mean deviation gives us a good description of the variability in a set of scores, and one that makes a good deal of sense. Unfortunately, it is rarely used, mainly because MD is not useful when extrapolating from samples to populations. The reason we are describing MD to you is that the most common measure of variability is just like the MD, only a little different.

**Variance and Standard Deviation**

If you were to square the deviations instead of taking their absolute values, and then average these squared deviations, not only would you get rid of the negative deviations, but the result would be an important measure of variability called the variance; it is symbolized by a lowercase sigma being squared, as in the following formula:

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$  \hspace{1cm} (1.3)

The numerator of this expression, the sum of the squared deviations from the mean, has its own abbreviation; it is known as the sum of squares, or even more briefly as SS. The variance is useful in advanced statistics, but it is not helpful as a descriptive measure of your set of scores, because it is in terms of squared scores. Taking the square root of the variance produces a good descriptive measure of variability that can also be useful for advanced statistics. The resulting measure is called the standard deviation, and it is symbolized by a lowercase sigma (without being squared), as in Formula 1.4.

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$  \hspace{1cm} (1.4)
It is important to realize that taking the square root after averaging the squared deviations does not entirely remove the effect of squaring. Otherwise, the standard deviation would always be the same as the mean deviation. Although MD and \( \sigma \) can be the same for a set of scores (e.g., when there are only two scores), \( \sigma \) is usually larger and can be quite a bit larger if there are a few extreme scores. In fact, the sensitivity of \( \sigma \) to extreme scores can be seen as a drawback. Just as the median can be a better descriptive measure than the mean when there are extreme scores, so too MD (or the SIQ) can be better than \( \sigma \) for descriptive purposes. But as we shall see shortly, \( \sigma \) plays a role in a very common distribution that makes it more useful than MD in advanced statistics. And even though \( \sigma \) is usually larger than MD for the same set of scores, \( \sigma \) is usually in the same ballpark, and therefore a good descriptive measure. The variability measures just described are summarized in Rapid Reference 1.2.

**Rapid Reference 1.2**

**Measures of Central Tendency**

The mode can be found with any scale of measurement; it is the only measure of typicality that can be used with a nominal scale.

The median can be used with ordinal, as well as interval/ratio, scales. It can even be used with scales that have open-ended categories at either end (e.g., \( 10 \) or more). It is not greatly affected by outliers, and it can be a good descriptive statistic for a strongly skewed distribution.

The mean can only be used with interval or ratio scales. It is affected by every score in the distribution, and it can be strongly affected by outliers. It may not be a good descriptive statistic for a skewed distribution, but it plays an important role in advanced statistics.

**Measures of Variability**

The range tells you the largest difference that you have among your scores. It is strongly affected by outliers, and being based on only two scores, it can be very unreliable.

The SIQ range has the same properties as described for the median, and is often used as a companion measure to the median.

The mean deviation, and the two measures that follow, can only be used with interval/ratio scales. It is a good descriptive measure, which is less affected by outliers than the standard deviation, but it is not used in advanced statistics.

The variance is not appropriate for descriptive purposes, but it plays an important role in advanced statistics.

The standard deviation is a good descriptive measure of variability, although it can be affected strongly by outliers. It plays an important role in advanced statistics.
THE NORMAL DISTRIBUTION

The best-known mathematical distribution and the one that is the most often applicable to variables in the social sciences is the one called the normal distribution. The normal distribution (ND), or normal curve as it is often called, has many convenient mathematical properties, but the one that is most relevant to us at this point is that the ND is completely determined by two of its characteristics (called parameters): its mean and its standard deviation. In other words, if two NDs have exactly the same \( \mu \) and \( \sigma \), they will overlap each other perfectly. You can see how the ND depends on \( \mu \) and \( \sigma \) by looking at the mathematical equation for the ND:

\[
f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}
\]  

(1.5)

\( f(x) \) is short for “function of” and it translates into \( y \), the vertical height of the curve at that value for \( x \); \( e \), like \( \pi \), is a constant (\( e = 2.718 \ldots \)). The exponent next to \( e \) has a minus sign, so the smaller the exponent, the higher the curve. The exponent is smallest when it is zero (\( e^0 = 1.0 \)), which occurs when \( X = \mu \), so the curve has its mode when \( X \) is at the mean.

One of the reasons the ND is so important to science (both physical and social) is that many variables in nature have distributions that look a lot like the ND. A common way that the ND arises is when many different independent factors contribute to the value of a variable, and each factor can just as easily contribute positively or negatively in any given case. If 20 factors contribute to a variable, a common result is 10 factors contributing positively and 10 negatively, leading to a middle value. Cases in which all 20 factors pull in the same direction will be rare and, therefore, so will extreme values on the variable. Something like this is probably acting to determine the heights of adult humans.

Let’s look at a likely distribution of the heights for an entire population of adult women (once again, it is simpler to look at one gender at a time; see Figure 1.3). The height distribution looks a lot like the ND, except for one simple fact: The height distribution ends on either side—there is no chance of finding an adult woman less than 2 feet or more than 9 feet tall. The true ND never ends; looking again at Formula 1.5, we see that the height of the curve does not fall to zero until the negative exponent of \( e \) reaches infinity. Moreover, the actual height distribution may not be perfectly symmetrical, and the curve may bend in a way that is slightly different from the ND. Still, it is so much easier to deal with the ND than real population distributions that it is common just to assume that the ND applies (with the same \( \mu \) and \( \sigma \) as the real distribution) and ignore the relatively small discrepancies that inevitably exist between the real distribution and the ND.