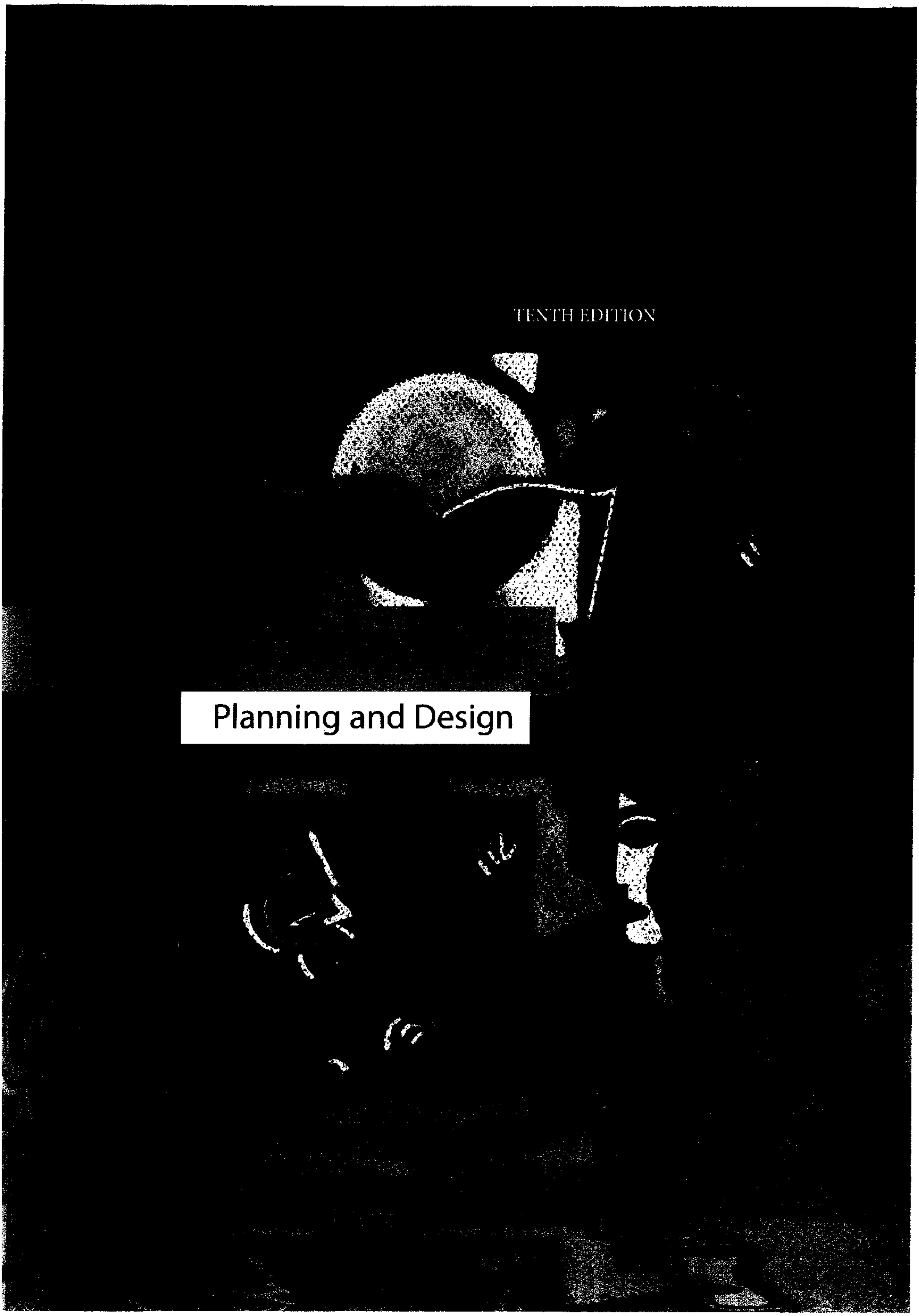
TENTH EDITION

Planning and Design



ISBN-13: 978-0-13-269324-0
ISBN-10: 0-13-269324-0

EAN

9 780132 693240 90000

www.pearsonhighered.com

PEARSON

ALWAYS LEARNING

11 Strategies for Analyzing Quantitative Data

Numbers are meaningless unless we analyze and interpret them in order to reveal the truth that lies beneath them. With statistics, we can summarize large numerical data sets, make predictions about future trends, and determine when different experimental treatments have led to significantly different outcomes. Thus, statistical procedures are among the most powerful tools in the researcher's toolbox.

In quantitative research, we try to make better sense of the world through the numbers we obtain. Sometimes these numbers represent aspects of the observable, physical world, such as the pull of gravity on a concrete object, the temperature of a gas, or the number of people engaging in a particular activity. We may also use numbers to represent nonphysical phenomena, such as how much students learn in the classroom, what beliefs people have about controversial topics, or how much influence various news media are perceived to have. We can then summarize and interpret the numbers by using statistics. In general, we can think of **statistics** as a group of computational procedures that enable us to find patterns and meaning in numerical data.

To some beginning researchers, the field of statistics can appear to be a never-never land in which advanced mathematicians conjure up elusive, hard-to-grasp numerical entities. But in reality, statistics are invaluable and often indispensable tools in research. They provide a means through which numerical data can be made more meaningful, so that the researcher can see their nature and better understand their interrelationships. The first and last question of statistics is precisely the same question that every researcher needs to ask: What do the data mean? In other words, What message do they communicate?

Exploring and Organizing a Data Set

Before employing any statistical procedure—before making a single computation—look closely at your data and explore various ways of organizing them. Using an open mind and your imagination, look for patterns in the numbers. Nothing takes the place of looking carefully, inquiringly, critically—perhaps even naively—at the data.

We take a simple example to illustrate the point. Following are the scores on a reading achievement test for 11 children: Ruth, 96; Robert, 60; Chuck, 68; Margaret, 88; Tom, 56; Mary, 92; Ralph, 64; Bill, 72; Alice, 80; Adam, 76; Kathy, 84. What do you see? Jot down a few observations before you read further.

Now let's try various arrangements of the scores to see what patterns they might reveal. Some of the information may be irrelevant to our research problem. No matter. Careful researchers discover everything possible about their data, whether the information is immediately useful or not.

We begin by forming an alphabetical list of the children's names and their test scores:

Adam	76	Mary	92
Alice	80	Ralph	64
Bill	72	Robert	60
Chuck	68	Ruth	96
Kathy	84	Tom	56
Margaret	88		

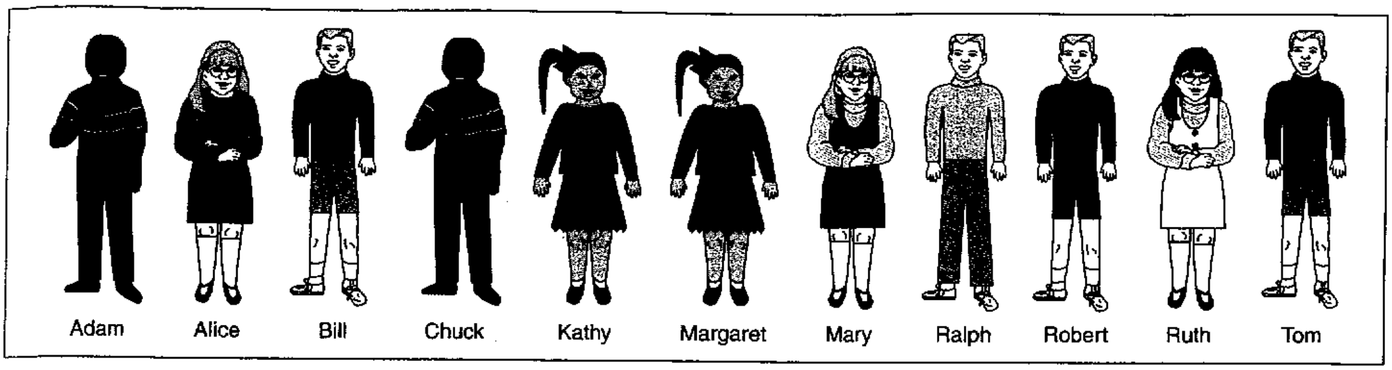


FIGURE 11.1
The 11 children in our reading achievement test sample

When we display the children's test scores in this manner, the scores are no more meaningful, but we have at least isolated individuals and scores so that we can inspect them more easily. Does this arrangement show us anything? Yes. It shows that the highest score was earned by a girl and that the lowest score was earned by a boy. Silly, you say, and meaningless. Perhaps. But it's an observable fact, and it might come in handy later on.

Let's keep the arrangement but view it in another way. In Figure 11.1, we see these 11 boys and girls lined up in a row, still arranged in alphabetical order according to first names. Look! Now we can discern a symmetrical pattern that wasn't previously apparent. No matter whether we start from the left or from the right, we have *one* boy, then *one* girl, then *two* boys, *three* girls, *two* boys, *one* girl, and *one* boy. Putting adjacent children of the same sex together, the arrangement is this:

♂ ♀ ♂♂ ♀♀♀ ♂♂ ♀ ♂

or:

1 1 2 3 2 1 1

Now let's arrange the data differently, separating girls from boys:

Girls		*Boys*	
Alice	80	Adam	76
Kathy	84	Bill	72
Margaret	88	Chuck	68
Mary	92	Ralph	64
Ruth	96	Robert	60
		Tom	56

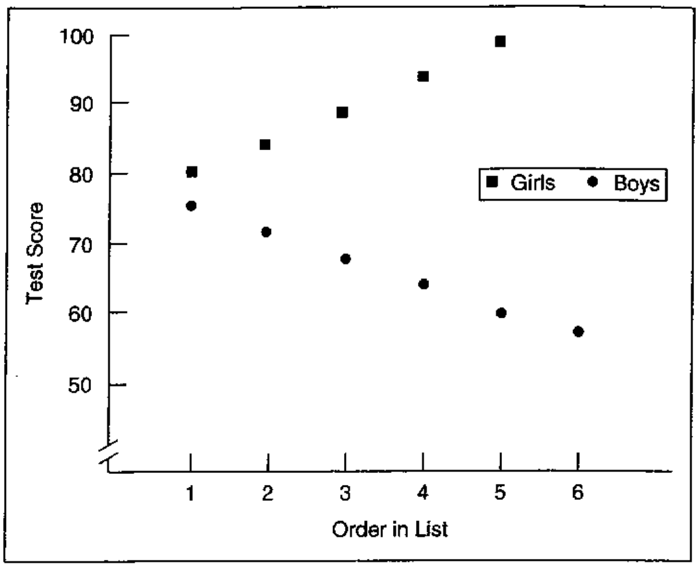
Represented graphically in Figure 11.2, the trends are quite dramatic: The girls' scores increase as we proceed through the alphabet, and the boys' scores decrease.

Not only are there opposite trends, but now we are aware of a very obvious fact that may, up to this point, have escaped our attention: The scores are equidistant from one another. Each score is 4 points either above or below the preceding one.

Whatever we have observed may have no relevance whatsoever for our project, but because it represents *dynamics within the data,* it is important that we see it. That is the point: The researcher should be aware of the dynamics—the phenomena—that are active within the data, whether those phenomena are important to the purpose of the research or not. The astute researcher overlooks nothing.

FIGURE 11.2

A visual representation of the reading achievement test scores



The preceding exercise was, of course, an artificial one. We would be hard pressed to find much meaning in diverging trends for girls versus boys that appear simply through an alphabetical arrangement of first names. Yet for the researcher working in an area of science, observations of a similar kind may reveal important new knowledge. Take the case of a paleontologist and an astronomer who examined growth marks on the spiral-shaped shells of a particular marine mollusk, the chambered nautilus (Kahn & Pompea, 1978). They noticed that each chamber in a shell had an average of 30 growth lines and deduced that (a) the growth lines had appeared at the rate of 1 per day and (b) one chamber had been laid down every lunar month, specifically every 29.53 days. They also concluded that, if their interpretation of the data was correct, it might be possible to determine from fossil shells the length of the ancient lunar months. Because the distance of the moon from Earth can be calculated from the length of the lunar month, the scientists examined nautilus fossils—some of them 420 million years old—and noticed a gradual decrease in the number of growth lines in each chamber as the fossils came from further and further back in prehistoric time. This finding indicated that the moon was once closer to Earth and revolved around it more rapidly than it does now—an observation consistent with generally accepted scientific theory.

In the two examples just presented, we find a fundamental principle about data exploration: *How the researcher prepares the data for inspection or interpretation will affect the meaning that those data reveal. Therefore, every researcher should be able to provide a clear, logical rationale for the procedure used to arrange and organize the data.* We had no rationale whatsoever for arranging the data according to the children's first names. Had we used their last names—which would have been equally illogical—we would still have seen that the girls had higher scores than the boys, but we would not necessarily have seen the diverging trends depicted in Figure 11.2.

In research questions regarding the physical world, the method for organizing data is apt to be fairly straightforward. Data often come to the scientist prepackaged and prearranged. The sequence of growth rings on a nautilus shell is already there, obvious and nondebatable. But in other disciplines—for instance, in the social sciences, humanities, and education—a researcher may need to give considerable thought to the issue of how best to organize the data.

Organizing Data to Make Them Easier to Think about and Interpret

As we mentioned in Chapter 1, the human mind can think about only so much information at one time. A data set of, say, 5,000 tidbits of information is well beyond a human being's mental capacity to consider simultaneously. In fact, unless the researcher has obtained *very* few pieces of data (perhaps only seven or eight numbers), he or she will want to organize them in one or more ways to make them easier to inspect and think about.

In the preceding example of 11 children and their reading achievement test scores, we experimented with several simple organizational schemes in an effort to find patterns in the data. Let's take another everyday example. Joe is in high school. In February he gets the following quiz grades: 92, 69, 91, 70, 90, 89, 72, 87, 73, 86, 85, 75, 84, 76, 83, 83, 77, 81, 78, 79. Here Joe's grades are listed in a *simple linear sequence*—the order in which Joe earned them. These are the raw numerical facts—the data—derived directly from the situation. As they appear in the preceding list, they don't say very much, except that Joe's performance seems to be inconsistent.

Let's put Joe's grades in a *two-dimensional table* organized by weeks and days:

Grade Record for February

	Monday	*Tuesday*	*Wednesday*	*Thursday*	*Friday*
First week	92	69	91	70	90
Second week	89	72	87	73	86
Third week	85	75	84	76	83
Fourth week	83	77	81	78	79

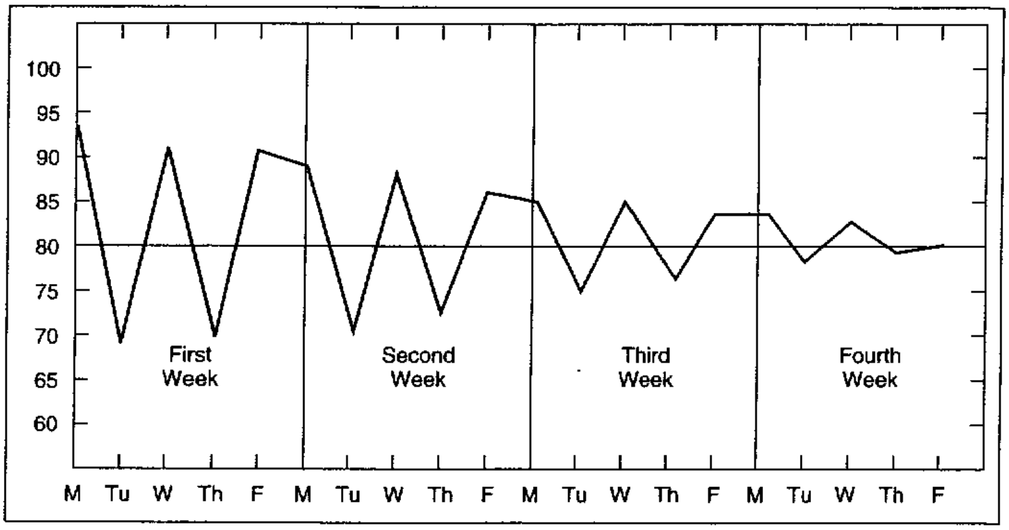
The table reveals some patterns in Joe's grades. If we compare the five columns, we quickly notice that the grades on Mondays, Wednesdays, and Fridays are considerably higher than those on Tuesdays and Thursdays. And if we look at successive numbers in each column, we see that the grades get progressively worse on Mondays, Wednesdays, and Fridays, but progressively better on Tuesdays and Thursdays.

Now let's represent Joe's grades in the form of a simple *line graph,* shown in Figure 11.3. In this graph, we see phenomena that were not readily apparent in the two-dimensional table. It's hard to miss the considerable variability in grades during the first and second weeks, followed by a gradual leveling-out process in the latter part of the month. A profile of this sort should prompt the alert researcher to explore the data further in an attempt to explain the pattern the graph reveals.

Graphing data is often quite useful for revealing patterns in a data set. For example, let's return to a study first described in a Practical Application exercise near the end of Chapter 9:

FIGURE 11.3

Line graph of Joe's daily grades



> Two researchers want to see if a particular training program is effective in teaching horses to enter a horse trailer without misbehaving in the process—that is, without rearing, trying to turn around, or in some other way resisting entry into the trailer. Five horses (Red, Penny, Shadow, Sammy, and Fancy) go through the training, with each horse beginning training on a different day. For each horse, an observer counts the number of misbehaviors every day prior to and during training, with data being collected for a time span of at least 45 days. (Ferguson & Rosales-Ruiz, 2001)

In Chapter 9 we were concerned only with the design of this study, concluding that it was a quasi-experimental (and more specifically, a multiple baseline) study. But now let's look at the results of the study. When the researchers plotted the numbers of five different misbehaviors for each horse before and during training, they constructed the graph presented in Figure 11.4. Was the training effective? Absolutely yes! Once training began, Penny had one really bad day plus another day in which she turned a couple of times, and Shadow and Fancy each tossed their heads during one of their loading sessions. Aside from these four occasions, the horses behaved perfectly throughout the lengthy training period, despite the fact that all five had been quite ornery prior to training. These data have what we might call a ***hit-you-between-the-eyes*** quality: We don't need a fancy statistical analysis to tell us that the training was effective.

Time-series studies often yield data that show clear hit-you-between-the-eyes patterns; for another example, return to Figure 9.3 on page 241. But generally speaking, simply organizing the data in various ways will not, in and of itself, reveal everything the data have to offer. Instead, a quantitative researcher will need to perform statistical analyses to fully discover the patterns and meanings the data hold. Before we turn to the nature of statistics, however, let's briefly look at how a researcher can use computer software to assist with the data organization process.

Using Computer Spreadsheets to Organize and Analyze Data

USING TECHNOLOGY

The process of organizing large amounts of data was once a cumbersome, time-consuming, and tedious task. Fortunately, the advent of computers has made the process much simpler and more efficient. One important tool is an **electronic spreadsheet**, a software program that allows a researcher to enter and then manipulate data in a two-dimensional table. Undoubtedly the best known spreadsheet software is Microsoft's Excel, but other software packages are available as well, including "freeware" you can download without charge from the Internet (e.g., Sphygmic Software Spreadsheet, Simple Spreadsheet, Spread32).

The beauty of electronic spreadsheets is that once you enter data into them, the software can quickly and easily help you organize the data and make simple calculations. For example, you can add several test scores together to create a new column that you might call "Total of Test Scores," or you might divide the numbers in one column by the numbers in another column to get proportions that are potentially meaningful in the context of your study. If you change a data point—for example, perhaps you discover that you miskeyed a test score and so must correct it—all of the relevant calculations are automatically updated. The software typically also lets you copy (or *import*) data from databases, word processing documents, or other spreadsheets into a new spreadsheet.

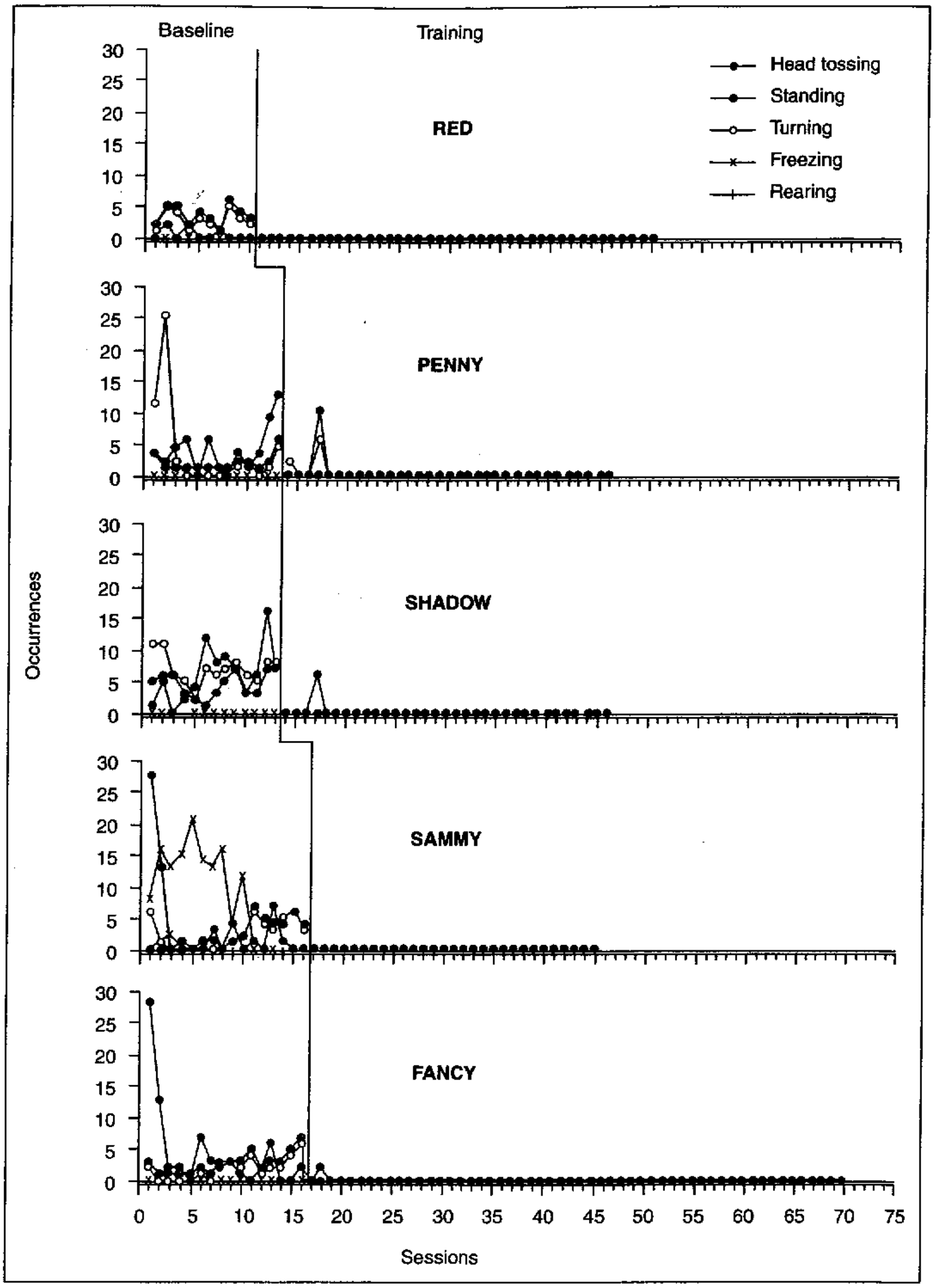
Spreadsheets would be useful to researchers even if they were capable only of listing data and adding up different columns and rows. But in fact, they allow the researcher to do many other things as well:

- *Sorting.* Once the data have been organized into rows and columns, it is possible to reorganize them in any way you wish. For example, suppose you have math test scores for a large number of children of various ages. You originally entered the scores in the order in which you obtained them. But now you decide that you want to consider them on the basis of the children's ages. In a matter of seconds, an electronic spreadsheet can sort the scores by age and list them from youngest to oldest child, or vice versa.
- *Recoding.* A spreadsheet typically allows you to make a new column that reflects a transformation of data in an existing column. For instance, imagine that you have

FIGURE 11.4

Undesirable trailer-loading behaviors for five horses before and after training

Reprinted from "Loading the Problem Loader: The Effects of Target Training and Shaping on Trailer-Loading Behavior of Horses" by D. L. Ferguson & J. Rosalez-Ruiz, 2001, *Journal of Applied Behavior Analysis, 34,* p. 419. Reprinted with permission of the Society for the Experimental Analysis of Behavior, Inc.



reading scores for children from ages 7 to 15. Perhaps you want to compare the scores for children in three different age groups: Group 1 will consist of children who are 7 to 9 years old, Group 2 will include 10- to 12-year-olds, and Group 3 will include 13- to 15-year-olds. You can tell the computer to form a new column called "Group" and to give each child a group number (1, 2, or 3) depending on the child's age.

- *Formulas.* Current spreadsheet programs have the capability to calculate many complex mathematical and statistical formulas. Once the data are organized into rows and columns, you can specify formulas that describe and analyze one or more groups of data. For example, you can enter the formula for computing the average, or mean, of a set of numbers, and the spreadsheet will perform the necessary calculations. Many commonly used formulas are often preprogrammed, so you merely select the statistic or function you need (e.g., you might select "AVERAGE") and highlight the data you wish to include in the calculation. The software does the rest.

- *Graphing.* Most spreadsheet programs have graphing capabilities. After you highlight the appropriate parts of the data, the program will automatically produce a graph from those data. Generally, the type of graph produced is selected from several options (e.g., line graphs, bar graphs, pie charts). Users can select how the axes are labeled, how the legend is created, and how the data points are depicted.
- *"What Ifs."* Thanks to the speed and ease with which an electronic spreadsheet can manipulate and perform calculations on large bodies of data, you can engage in numerous trial-and-error explorations. For example, if you are examining data for a sample of 5,000 people and decide that an additional comparison between certain subgroups might prove interesting, the spreadsheet can complete the comparison in a matter of seconds. This capability allows you to continually ask *what if . . .* ?—for instance, What if the data were analyzed on the basis of gender, rather than on the basis of age? or What if results from administering only one level of a specific medication were analyzed instead of grouping all levels together? This *what-if* capability allows the researcher to explore the data in many possible ways quickly and easily.

In the discussion of Microsoft Excel in Appendix A, you can learn how to use some of the many features that an electronic spreadsheet offers.

We have said enough about organizing a data set. We now turn to one of the most important tools in a researcher's toolbox—statistical analysis.

Choosing Appropriate Statistics

In a single chapter we cannot thoroughly describe the wide variety of statistical procedures available to researchers. Here we must limit ourselves to a description of basic statistical concepts and principles and a brief overview of some of the most commonly used procedures. We are assuming that you have taken or will take at least one course in statistics—better still, take two, three, or even more!—to get a firm foundation in this essential research tool.

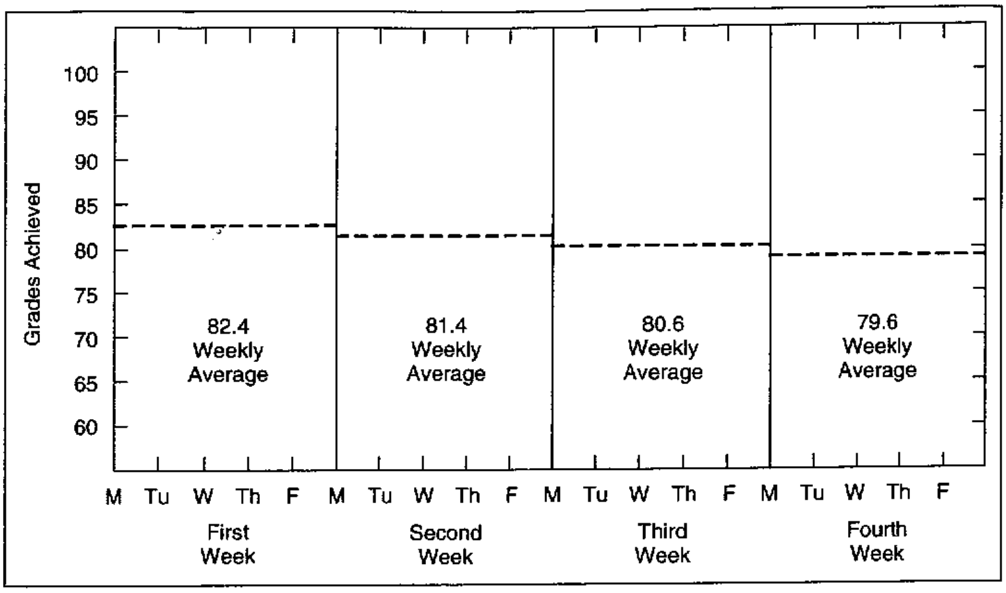
In a preceding section, we looked at Joe's test scores in three ways: a simple linear sequence, a two-dimensional table, and a line graph. All of these depicted Joe's day-to-day performance. Now, instead, let's begin to summarize what we are seeing in the test scores. We can, for example, use a statistic known as a *mean*—in everyday terms, an *average*—to take out the jagged irregularities of Joe's daily performance. In Figure 11.5, we represent Joe's average scores for the four weeks of February with four broken lines. When we do this, we get an entirely new view of Joe's achievement. Whereas Figure 11.3 showed only an erratic zigzagging between daily extremes, with the zigzags becoming less extreme as the weeks went by, the dotted lines in Figure 11.5 show that week by week, very little change actually occurred in Joe's average test performance.

Yet it may be that we also want to summarize how much Joe's grades *vary* each week. The means presented in Figure 11.5 tell us nothing about how consistent or inconsistent Joe's grades are in any given week. We would need a different statistic—perhaps a *range* or a *standard deviation*—to summarize the variability we see each week. (We'll describe the nature of such measures of variability shortly.)

Thus far, we have discovered an important point: *Looking at data in only one way yields incomplete view of those data and thus provides only a portion of the meaning those data hold.* For reason, we have many statistical techniques, each of which is suitable for a different purp

FIGURE 11.5

Line graph of Joe's weekly average grades



Each technique extracts a somewhat different meaning from a particular set of data. Every time you apply a new statistical treatment to your data, you derive new insights and see more clearly and completely the data's underlying dynamics.

In the next few pages, we consider two general functions that statistics can serve. We also discuss the various ways in which the nature of the data may dictate the particular statistical procedures that can be used.

Functions of Statistics

Statistics have two major functions. Some statistics describe what the data look like—where their center or midpoint is, how broadly they are spread, how closely two or more variables within the data are intercorrelated, and so on. Such statistics are, appropriately, called **descriptive statistics.**

Other statistics, known as **inferential statistics,** serve a different purpose: They allow us to draw *inferences* about large populations by collecting data on relatively small samples. For example, imagine that you are an immigration officer. Although you have never been to Egypt, you have met numerous Egyptians as they debark from incoming planes and ships. Perhaps you have even become well acquainted with a small number of Egyptians. From this small sample of the Egyptian population, you might infer what Egyptian people in general are like. (Your inferences may or may not be accurate because your sample, which consists entirely of visitors and immigrants to your own country, is not necessarily representative of the entire population of Egypt. However, that is a sampling problem, not a statistical one.)

More generally, inferential statistics involve using a small sample of a population and then *estimating* the characteristics of the larger population from which the sample has been drawn. For instance, we might estimate a population mean from the mean we obtain for a sample. Or we might determine whether two or more groups of people are actually different, given the differences we observe between samples taken from each of those groups. Inferential statistics provide a way of helping us make reasonable guesses about a large, unknown population by examining a small sample that *is* known. In the process, they also allow us to test hypotheses regarding what is true for that large population.

Statistics as Estimates of Population Parameters

Especially when we use statistics to draw inferences about a population from which a research sample has been drawn, we are using them as *estimates of population parameters.* A **parameter** is a characteristic or quality of a population that, in *concept,* is a constant; however, its *value* is variable.

As an illustration, let's consider a circle. One of the parameters of a circle is its radius. In concept, the radius is a constant: It is the same for every circle—the distance from the center of the circle to the perimeter. In value, it varies, depending on the size of the circle. Large circles have long radii; small circles, short radii. The value—that is, the length of the radius in linear units (centimeters, feet, etc.)—is variable. Thinking of a parameter in this way, we see that each circle has several parameters: The diameter is always twice the radius *(r),* the circumference is always $2\pi r$, and the area is always πr^2. These concepts are constants, even though their particular values vary from one circle to the next.

Within the context of of quantitative data analysis, a parameter is a particular characteristic (e.g., a mean or standard deviation) of the entire population—which is sometimes called a *universe*—about which we want to draw conclusions. In most cases, we can study only a small sample of a population. Any calculation we perform for the sample rather than the population (the sample mean, the sample standard deviation, etc.) is called a **statistic.** Statisticians distinguish between population parameters and sample statistics by using different symbols for each. Table 11.1 presents a few commonly used symbols in statistical notation.

Considering the Nature of the Data

As you begin to think about the statistical procedures that might be most appropriate for your research problem, keep in mind that different statistics are suitable for different kinds of data. In particular, you should consider whether your data

- Have been collected for a single group or, instead, for two or more groups
- Involve continuous or discrete variables
- Represent nominal, ordinal, interval, or ratio scales
- Reflect a normal or non-normal distribution

After we look at each of these distinctions, we will relate them to another distinction—that between parametric and nonparametric statistics.

Single-Group versus Multi-Group Data

In some cases, a research project yields data about a single group of people, objects, or events. In other cases, it may yield parallel sets of data about two or more groups. Analyzing characteristics of a single group will often require different statistical techniques than those for making comparisons among two or more groups.

TABLE 11.1

Conventional statistical notation for population parameters and sample statistics

	The Symbol Used to Designate the Factor	
The Factor in Question	*Population Parameter*	*Sample Statistic*
The mean	μ	$\bar{M}$ or X
The standard deviation	σ	s or SD
Proportion or probability	P	p
Number or total	N	n

Note: The symbol μ is the lowercase form of the Greek letter *mu.* The symbol σ is the lowercase form of the Greek letter *sigma.*

Continuous versus Discrete Variables

In Chapter 2 we define a *variable* as a quality or characteristic in a research investigation that has two or more possible values. Simply put, a variable *varies.* However, it may vary in different ways. A **continuous variable** reflects an infinite number of possible values falling along a particular continuum. A simple example is chronological age. The participants in a research study can be an infinite number of possible ages. Some might be 2 years old, others might be 92, and we might have virtually any age (including fractions of years) in between. Even if the study is limited to a small age range—say, 2- to 4-year-old children—we might have children who are exactly 2 years old, children who are 2 years and 1 month old, children who are 2 years and 2 months old, and so on. We could, in theory, be even more precise, perhaps specifying participants' ages in days, hours, minutes, seconds, or even fractions of a second.

In contrast, a **discrete variable** has a finite and small number of possible values. A simple example is a student's high school grade level. At a four-year high school, a student can be in only one of four grades: 9th, 10th, 11th, or 12th. At most high schools, it isn't possible to be in anything else. One cannot be somewhere between two grade levels, such as in the "9.25th grade."

Nominal, Ordinal, Interval, and Ratio Data

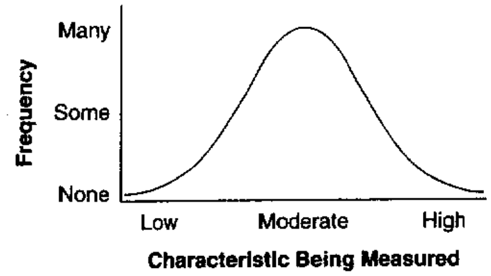
In Chapter 4 we describe four different scales of measurement; these scales, in turn, dictate how we can statistically analyze the numbers we obtain relative to one another. To refresh your memory, we briefly describe each of the scales again.

- *Nominal data* are those for which numbers are used only to identify different categories of people, objects, or other entities; they do not reflect a particular quantity or degree of something. For instance, a researcher might code all males in a data set as 1 and all females as 2. The researcher might also code political affiliation with numbers, perhaps using 1 for Republicans, 2 for Democrats, 3 for Independents, and so on. In neither case do the numbers indicate that participants have more or less of something; girls don't have more "sex" than boys, and Independents don't have more "political affiliation" than Republicans or Democrats.
- *Ordinal data* are those for which the assigned numbers reflect an order or sequence. They tell us the degree to which people, objects, or other entities have a certain quality or characteristic (a variable) of interest. They do not, however, tell us anything about how great the differences are between the people, objects, or other entities. For example, in a group of graduating high school seniors, each student might have a class rank that reflects his or her relative academic standing in the group: A class rank of 1 indicates the highest grade point average (GPA), a rank of 2 indicates the second highest GPA, and so on. These numbers tell us which students surpassed others in terms of GPA, but it does not tell us precisely how similar or different the GPAs of any two students in the sequence are.
- *Interval data* reflect equal units of measurement. As is true for ordinal data, the numbers reflect differences in degree or amount. But in addition, differences between the numbers tell us *how much difference* exists in the characteristic being measured. As an example, scores on intelligence tests (IQ scores) are, because of the way in which they are derived, assumed to reflect an interval scale. Thus, if we take four IQ scores at equal intervals—for instance, 85, 95, 105, and 115—we can assume that the 10-point difference between each pair reflects equivalent differences in intelligence between the people who have obtained those scores. The one limitation of interval data is that a value of zero (0) does *not* necessarily reflect a complete lack of the characteristic being measured. For example, it is sometimes possible to get an IQ score of 0, but such a score does not mean that a person has no intelligence whatsoever.
- *Ratio data* are similar to interval data but have an additional feature: a true zero point. Not only do the numbers reflect equal intervals between values for the characteristic being measured, but in addition a value of 0 tells us that there is a complete absence of that characteristic. An example would be income level: People with an annual income of

$30,000 make $10,000 more than people with an annual income of $20,000, and people with an annual income of $40,000 make $10,000 more than people with an annual income of $30,000. Furthermore, people who make $0 a year have *no* income.

Normal and Non-Normal Distributions

Numerous theorists have proposed that many characteristics of living populations (e.g., populations of maple trees, platypuses, human beings, or a certain subgroup of human beings) reflect a particular pattern, one that looks like this:

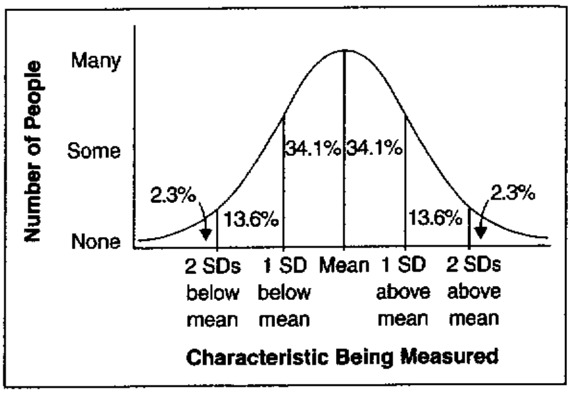


This pattern, commonly called the **normal distribution** or **normal curve**—you may also see the term *bell curve*—has several distinguishing characteristics:

- *It is horizontally symmetrical.* One side is the mirror image of the other.
- *Its highest point is at its midpoint.* More people (or whatever other entities are the focus of investigation) are located at the midpoint than at any other point along the curve. In statistical terms, three widely used measures of central tendency—the mode, the median, and the mean (all to be described shortly)—are equivalent.
- *Predictable percentages of the population lie within any given portion of the curve.* If we divide the curve according to its standard deviation (also to be described shortly), we know that certain percentages of the population lie within each portion. In particular, approximately 34.1% of the population lies between the mean and one standard deviation below the mean, and another 34.1% lies between the mean and one standard deviation above the mean. Approximately 13.6% of the population lies between one and two standard deviations below the mean, with another 13.6% lying between one and two standard deviations above the mean. The remaining 4.6% lies two or more standard deviations away from the mean, with 2.3% at each end of the distribution. This pattern is shown in Figure 11.6. The proportions of the population lying within any particular section of the normal distribution can be found in most introductory statistics books. You can also find them online by using the key words "normal distribution table" in a search engine such as Google or Yahoo!

FIGURE 11.6

Percentages within each portion of the normal distribution



To better understand the normal distribution, take any fortuitous happening and analyze its distribution pattern. For example, let's take the corn production of Iowa farms during a single year. If we could survey the per-acre yield of every farmer in Iowa—the total population, or universe, of the cornfields and corn farmers in Iowa—we would probably find that a few farmers had an unusually poor yield of corn per acre for no discernible reason except that "that's the way it happened." A few other farmers, for an equally unknown reason, likely had unusually heavy yields from their fields. However, most farmers would have had a middle-of-the-road yield, sloping gradually in either direction toward the greater-yield or the lesser-yield direction. The normal curve would describe the Iowa corn production. No one planned it that way; it is simply how nature behaves.

Watch an approaching thunderstorm. An occasional flash of lightning heralds the coming of the storm. Soon the flashes occur more frequently. At the height of the storm, the number of flashes per minute reach a peak. Gradually, with the passing of the storm, the number of flashes subsides. The normal curve is at work once again.

We could think of thousands of situations, only to find that nature often behaves in a way consistent with the normal distribution. The curve is a constant; it is always bell-shaped. In any one situation, the *values* within it vary. The mean is not always the same number, and the overall shape may be more broadly spread or more compressed, depending on the situation.

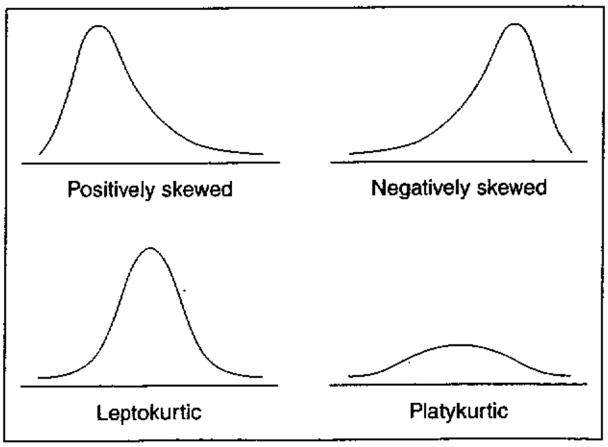
Sometimes, however, a variable doesn't fall in a normal distribution. For instance, its distribution might be lopsided, or **skewed**. If the peak lies to the left of midpoint, the distribution is **positively skewed**; if the peak lies to the right of midpoint, the distribution is **negatively skewed**. Or perhaps a distribution is unusually pointy or flat, such that the percentages within each portion of the distribution are notably different from those depicted in Figure 11.6. Here we are talking about **kurtosis**, with an unusually peaked, or pointy, distribution reflecting a **leptokurtic curve** and an unusually flat one being a **platykurtic curve** (see Figure 11.7).

Of course, some data sets don't resemble a normal distribution, not even a lopsided, pointy, or overly flattened variation of one. In general, ordinal data, by virtue of how they are created, *never* fall in a normal distribution. For instance, a data set might look more like a stairway that progresses upward in regular intervals. Or, take a graduating high school class. If each student is given a class rank according to academic grade point average, Luis might rank first, Janene might rank second, Marietta third, and so on. We don't see a normal distribution in this situation because we have only one student at each academic rank. If we construct a graph that depicts the frequencies of the class ranks, we see a low, flat distribution rather than one that rises upward and peaks in the middle.

Percentile ranks, too, form a flat distribution rather than a bell-shaped curve. Percentile ranks are often used to report performance on scholastic aptitude and achievement tests. To calculate them, a researcher first determines the **raw score**—the number of

FIGURE 11.7

Common departures from the normal distribution



test items correctly answered or number of points accumulated—that each person in the sample earns on a test or other research instrument. Each person's percentile rank is then calculated as follows:

$$\text{Percentile rank} = \frac{\text{Number of other people scoring } \textit{lower} \text{ than the person}}{\text{Total number of people in the sample}}$$

By the very nature of how they are calculated, percentile ranks spread people evenly over the number of possible ranks one might get; for instance, there will be roughly the same number of people earning percentile ranks of 5, 35, 65, and 95. Furthermore, although percentile ranks tell us how people have performed relative to one another, they do not tell us ***how much*** they differ from one another in the characteristic being assessed. In essence, percentile ranks are ***ordinal data*** and must be treated as such.

Choosing between Parametric and Nonparametric Statistics

Your choice of statistical procedures must depend to some degree on the nature of your data and the extent to which they reflect a normal distribution. Some statistics, known as **parametric statistics**, are based on certain assumptions about the nature of the population in question. Two of the most common assumptions are these:

- The data reflect an interval or ratio scale.
- The data fall in a normal distribution (e.g., the distribution has a central high point, and it is not seriously skewed, leptokurtic, or platykurtic).

When either of these assumptions is violated, the results one obtains from parametric statistics may be suspect.

Other statistics, called **nonparametric statistics**, are not based on such assumptions. For instance, some nonparametric statistics are appropriate for data that are ordinal rather than interval in nature. Others may be useful when a population is highly skewed in one direction or the other.

You may be thinking, Why not use nonparametric statistics all the time to avoid having to make (and possibly violate) any assumptions about the data? The reason is simple: Our most complex and powerful inferential statistics are based on parametric statistics. Nonparametric statistics are, by and large, appropriate only for relatively simple analyses.

On an optimistic note, we should point out that some statistical procedures are **robust** with respect to certain assumptions. That is, they yield generally valid results even when an assumption isn't met. For instance, a particular procedure might be as valid with a leptokurtic or platykurtic distribution as it is with a normal distribution; it might even be valid with ordinal rather than interval data. When using any statistical technique, you should consult with a statistics textbook to determine what assumptions are essential for that technique and what assumptions might reasonably be disregarded. Some statistical software packages routinely provide information about whether a particular data set meets or violates certain assumptions and make appropriate adjustments for non-normal distributions.

Descriptive Statistics

As their name implies, descriptive statistics *describe* a body of data. Here we discuss how to determine three things a researcher might want to know about a data set: points of central tendency, amount of variability, and the extent to which two or more variables are associated with one another.

Measures of Central Tendency

A *point of central tendency* is a point around which the data revolve, a middle number around which the data regarding a particular variable seem to hover. In statistical language, we use the term *measures of central tendency* to refer to techniques for finding such a point. Three commonly used measures of central tendency are the mode, the median, and the mean, each of which has its own characteristics and applications.

The **mode** is the single number or score that occurs most frequently. For instance, in this data set

3 4 6 7 7 9 9 9 9 10 11 11 13 13 13 15 15 21 26

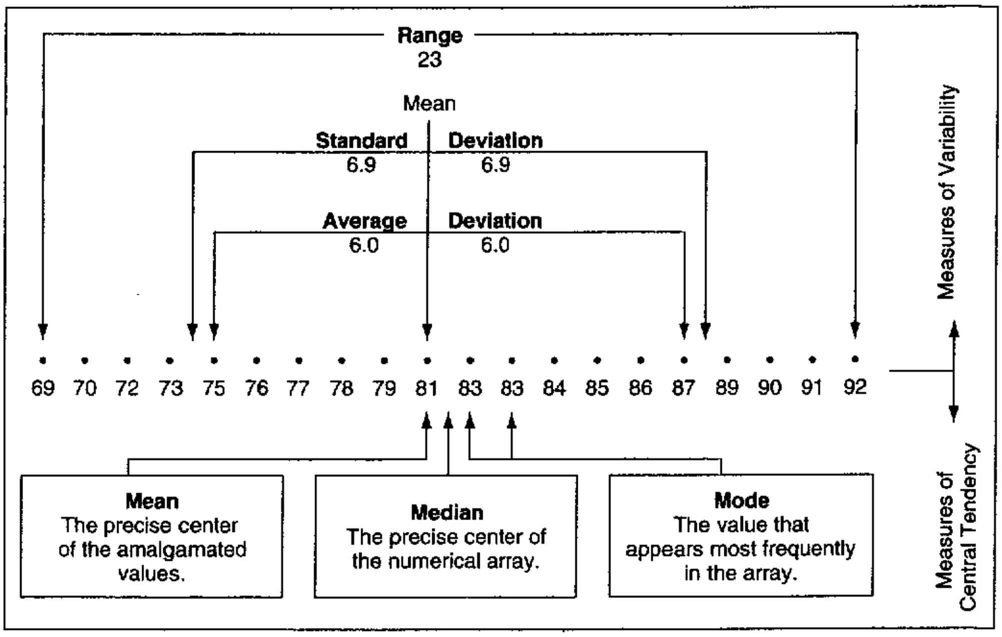
the mode is 9, because 9 occurs more frequently (four times) than any other number. Similarly, if we look at the list of Joe's grades (see p. 273), we see that only one score (83) appears more than once; thus, 83 is the mode. As a measure of central tendency, the mode is of limited value, in part because it doesn't always appear near the middle of the distribution and in part because it isn't very stable from sample to sample. However, the mode is the *only* appropriate measure of central tendency for nominal data.

The **median** is the numerical center of a set of data, with exactly as many scores above it as below it. The word *median* is a derivation of the Latin word for "middle," and so the median score is the one precisely in the middle of the series. Recall that Joe's record has 20 grades. Thus, 10 grades are above the median, and 10 are below it. The median is midway in the series between the 10th and 11th scores, or in this case, midway between the scores of 81 and 83—that is, 82 (see Figure 11.8).

You might think of the mean as the fulcrum point for a set of data: It represents the single point at which the two sides of a distribution "balance." Mathematically, the mean is the *arithmetic average*[1] of the scores within the data set. To find it, we calculate the sum of all the scores (adding each score every time it occurs) and then divide by the total number of scores. If we use

FIGURE 11.8

Measures of central tendency and variability for Joe's grades


Range
23
Mean
Standard
6.9
Deviation
6.9
Average
6.0
Deviation
6.0
69 70 72 73 75 76 77 78 79 81 83 83 84 85 86 87 89 90 91 92
Measures of Variability
Mean
The precise center of the amalgamated values.
Median
The precise center of the numerical array.
Mode
The value that appears most frequently in the array.
Measures of Central Tendency

[1] When the word *arithmetic* is used as an adjective, as it is here, it is typically pronounced with emphasis on the third syllable ("ar-ith-MET-ic").

the symbol X to refer to each score in the data set and the symbol N to refer to the total number of scores, we calculate the mean as follows:

$$M = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N}$$

Statisticians frequently use the symbol Σ (uppercase form of the Greek letter *sigma)* to designate adding all of the numbers related to a particular variable; thus, we can rewrite the formula for a mean as follows:

$$M = \frac{\Sigma X}{N}$$

Using the formula, we find that the mean for Joe's grades is 81, as illustrated in Figure 11.8. (The variation in Joe's grades, depicted in the figure as *measures of variability,* is discussed shortly.)

The mean is the measure of central tendency most commonly used in statistical analyses and research reports. However, it is appropriate only for interval or ratio data, because it makes mathematical sense to compute an average only when the numbers reflect equal intervals along a particular scale.

The median is more appropriate for dealing with ordinal data. The median is also used frequently when a researcher is dealing with a data set that is highly skewed in one direction or the other. As an example, consider this set of scores:

3 4 5 5 6 9 15 17 125

The mean for these scores is 21, a number that doesn't give us a very good idea of the point near which most of the scores are located. The median, which in this case is 6, is a better reflection of central tendency because it isn't affected by the single extreme score of 125. Similarly, medians are often used to reflect central tendency in family income levels, home values, and other such financial variables; most family incomes and home values are clustered at the lower end of the scale, with only a very few extending into the million-dollar range.

Curves Determine Means

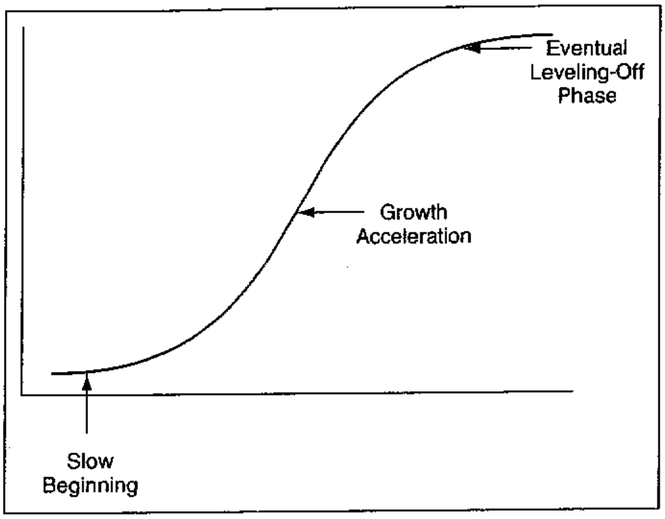
The mean as we have just described it—sometimes known as the *arithmetic* mean[2]—is most appropriate when we have a normal distribution, or at least a distribution that is somewhat symmetrical. But not all phenomena fit a bell-shaped pattern. Growth is one: It often follows an ogive curve that eventually flattens into a plateau (see Figure 11.9).

Growth is a function of geometric progression. As an example, let's consider the work of Thomas Robert Malthus, an English clergyman and economist who theorized about the potential for a population explosion and resulting worldwide famine. In *An Essay on the Principle of Population* (1826/1963), Malthus contended that, when unchecked, a population increases at an exponential rate, in which each successive value depends, multiplicatively, on the preceding value; for example, in the series 2, 4, 8, 16, 32, 64, 128 . . . , each number is twice the preceding number. But Malthus also predicted that the size of the human population would eventually flatten out because there is an upper limit to what Mother Earth can produce in the way of food to sustain the population. Thus, many growth curves resemble an *S*, as illustrated in Figure 11.9.

If we are recording the growth of bean stalks in an agronomy laboratory, we do not find the average growth by assuming a normal distribution and calculating the arithmetic mean. The statistical technique does not fit the natural fact. Instead, we use the **geometric mean**, which is computed by *multiplying* all of the scores together and then finding the *N*th root of the product. In other words, the geometric mean, which we can symbolize as M_g, is calculated as follows:

[2] Again, the emphasis in *arithmetic* is on the third syllable ("ar-ith-MET-ic").

FIGURE 11.9
Typical growth curve



$$M_g = \sqrt[N]{(X_1)\,(X_2)\,(X_3)\ldots(X_N)}$$

For growth phenomena, we use the geometric mean because that is the way things grow. That is the way cells divide—geometrically.

Biologists, physicists, ecologists, and economists all encounter growth phenomena in one form or another. They all witness the same typical aspects of change: a slow beginning (a few settlers in an uninhabited region, a few bacteria on a culture); then, after a period of time, rapid expansion (the boom period of city growth, the rapid multiplication of microorganisms); and finally—sometimes but not always—a leveling-off period (the land becomes scarce and the city sprawl is contained by geographical and economic factors, the bacteria have populated the entire culture). Following are examples of situations in which the application of the geometric mean is appropriate:

- Biological growth
- Population growth
- Increments of money at compound interest
- Decay or simple decelerative situations

In every situation, one basic principle applies: The configuration of the data dictates the measure of central tendency most appropriate for that particular situation. If the data fall in a distribution that approximates a normal curve (as most data do), they call for one measure of central tendency. If they assume an ogive curve configuration (characteristic of a growth situation), they demand another measure. A *polymodal distribution*—one with several distinct peaks—might call for still a third approach; for instance, the researcher might describe it in terms of its two or more modes. Only after careful and informed consideration of the characteristics of the data can the researcher select the most appropriate statistical measure.

Thus, we must emphasize an essential rule for researchers who use statistics in their data analysis: *The nature of the data—the facts of life—governs the statistical technique, not the other way around.* Just as the physician must know what drugs are available for specific diseases and disorders, so the researcher must know what statistical techniques are suited to specific research demands. Table 11.2 presents a summary of the measures of central tendency and their uses, together with the various types of data for which each measure is appropriate.

TABLE 11.2

Using measures of central tendency for different types of data

[illegible]	[illegible] It Is Determined ([illegible] number of scores)	Data for Which It Is Appropriate
Mode	The most frequently occurring score is identified.	• Data on nominal, ordinal, interval, and ratio scales • Multimodal distributions (two or more modes may be identified when a distribution has multiple peaks)
Median	The scores are arranged in order from smallest to largest, and the middle score (when *N* is an odd number) or the midpoint between the two middle scores (when *N* is an even number) is identified.	• Data on ordinal, interval, and ratio scales • Data that are highly skewed
Arithmetic mean	All the scores are added together, and their sum is divided by the total number (*N*) of scores.	• Data on interval and ratio scales • Data that fall in a normal distribution
Geometric Mean	All the scores are multiplied together, and the *N*th root of their product is computed.	• Data on ratio scales • Data that fall in an ogive curve (e.g., growth data)

Measures of Central Tendency as Predictors

Some researchers regard the matter of central tendency from a somewhat different standpoint. They consider it from the perspective of optimal chance: What is the best prediction?

As an example, consider this situation. Suppose you are walking down the street. Suddenly you come to a crowd of people forming in a normal-curve-like manner. Where, based on your best prediction, will you find the cause for the crowd forming? The answer is simple. Where the crowd is deepest, where the greatest number of people are, you will probably find the cause for the gathering. It may be an accident, a street fight, or a person giving away free candy bars. But whatever the occasion, your best guess about the cause of the gathering lies at the point where the human mass is at its peak.

Similarly, we can often make reasonable predictions about a population based on our knowledge of central tendency. When we speak of "the average citizen," "the average student," and "the average wage earner," we are referring to those citizens, students, and wage earners who are huddled around the point of central tendency. In the broad spectrum of possibilities, we are betting on the average as being the best single guess about the nature of the total population.

Measures of Variability: Dispersion and Deviation

Up to this point, we have been discussing the question, What is the best guess? Now we turn to the opposite question: What are the worst odds? This, too, is important to know. The more that the data cluster around the point of central tendency, the greater is the probability of making a correct guess about where any particular data point lies. The farther the data are dispersed from the central axis, the greater the margin of predictive error becomes. Consider, for example, the two curves shown in Figure 11.10. The data are more similar if they cluster about the mean. Scatter them, and they lose some of their uniformity; they become more diverse, more heterogeneous. As specific data points recede farther from the mean, they lose more and more of the quality that makes them "average."

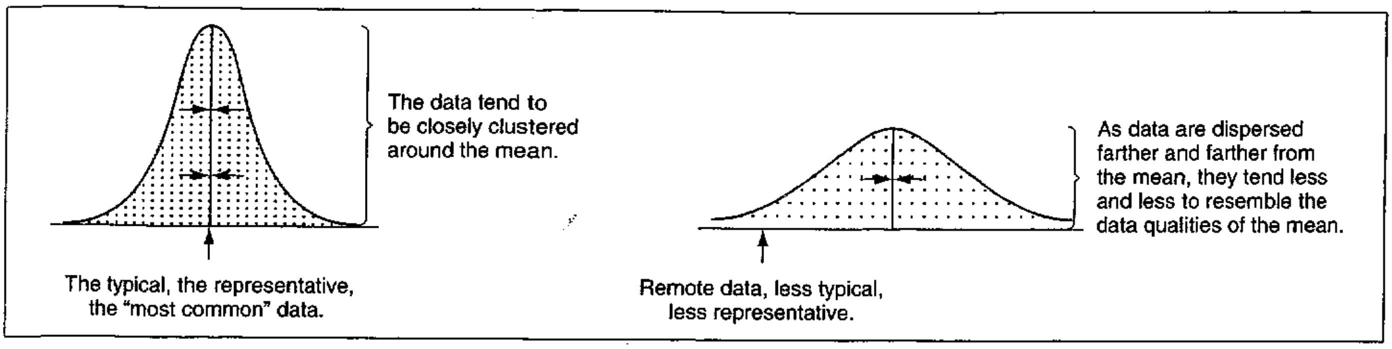

The data tend to be closely clustered around the mean.
The typical, the representative, the "most common" data.
As data are dispersed farther and farther from the mean, they tend less and less to resemble the data qualities of the mean.
Remote data, less typical, less representative.

FIGURE 11.10
Distributions that differ in variability

To derive meaning from data, then, it's important to determine not only their central tendency but also their spread. And it often helps to pin down their spread in terms of one or more quantitative values.

How Great Is the Spread?

The simplest measure of variability is the range, which indicates the spread of the data from lowest to highest value:

Range = Highest score - Lowest score

For instance, the range for Joe's test scores is 92 - 69, or 23 (see Figure 11.8).

Although the range is easy to compute, it has limited usefulness as a measure of variability and may even be misleading if the extreme upper or lower limits are atypical of the other values in the series. Let's take an example. Following are the numbers of children in each of ten families: 1, 3, 3, 3, 4, 4, 5, 5, 6, 15. We might say that the families range from one with a single child to a family of 15 children (a range of 15 – 1, or 14). But this figure is misleading: It suggests that the sample shows a great deal of variability in family size. We give a more realistic estimate of variability in this sample if we say that 80% of the families have from 3 to 6 children.

Other measures of variability use the median or mean as a starting point. One such measure is the **interquartile range**. If we divide the distribution into four equal parts, Quartile 1 lies at a point where 25% of the members of the group are below it. Quartile 2 divides the group into two equal parts and is identical to the median. Quartile 3 lies at a point where 75% of the values are below it.[3] The interquartile range is equal to Quartile 3 (the 75th percentile point) minus Quartile 1 (the 25th percentile point), as follows:

Interquartile range = Quartile 3 - Quartile 1

Thus, the interquartile range gives us the range for the middle 50% of the cases in the distribution. Because quartiles are associated with the median, any researcher employing the median as a measure of central tendency should also consider the quartile deviation as a possible statistical measure for variability.

But now let's instead use the mean as a starting point. Imagine that we determine how far away from the mean each piece of data is in the distribution. That is, we calculate the *difference* between each score and the mean score (we call this difference the *deviation*). If we were to add all of these differences (ignoring the plus and minus signs) and then divide the sum by the *number* of scores (which reflects the number of score–mean differences as well), we get the *average* of the

[3] If, instead of dividing the data into 4 equal parts, we divide them into 10 equal parts, each part is called a *decile*; if into 100 equal parts, each part is called a *percentile*.

would simply multiply the z by the new scale's standard deviation (s_{new}) and then add the new scale's mean (M_{new}) to the product obtained, as follows:

$$\text{New standard score} = (z \times s_{\text{new}}) + M_{\text{new}}$$

Let's take an example. One common standard-score scale is the IQ scale, which uses a mean of 100 and a standard deviation of 15. (As you might guess, this scale is the one on which intelligence test scores are typically based.) If we were to convert Mary's extroversion score to the IQ scale, we would plug her z-score of 2, plus a standard deviation of 15 and a mean of 100, into the preceding formula, as follows:

$$\text{IQ score} = (2 \times 15) + 100 = 130$$

Thus, using the IQ scale, Mary's score on the extroversion test would be 130.

Another commonly used standard-score scale is the **stanine**. Stanines have a mean of 5 and a standard deviation of 2. Mary's stanine would be 9, as we can see from the following calculation:

$$\text{Stanine} = (2 \times 2) + 5 = 9$$

Stanines are *always* a whole number from 1 to 9. If our calculations gave us a number with a fraction or decimal, we would round it off to the nearest whole number. If some of our calculations resulted in numbers of 0 or less, or 10 or more, we would change those scores to 1 and 9, respectively.

Standard scores take a variety of forms, each with a prespecified mean and standard deviation; z-scores, IQs, and stanines are just three examples.[4] But in general, standard scores give us a context that helps us interpret the scores: If we know the mean and standard deviation on which the scores are based, then we also know where in the distribution any particular score lies. For instance, an IQ score of 70 is two standard deviations (30 points) below the mean of 100, and a stanine score of 6 is one half of a standard deviation (1 is half of 2) above the mean of 5.

Converting data to standard scores does not change the shape of the distribution; it merely changes the mean and standard deviation of that distribution. But imagine that, instead, we *do* want to change the nature of the distribution. Perhaps we want to change a skewed distribution into a more balanced, normally distributed one. Perhaps, in the process, we also want to change ordinal data into interval data. Several procedures exist for doing such things, but describing them would divert us from the basic nature and functions of statistics that we need to focus on here. You can find discussions of *normalizing* a data set in many basic statistics textbooks; another good resource is Harwell and Gatti (2001).

Keeping Measures of Central Tendency and Variability in Perspective

Statistics related to central tendency and variability help us summarize our data. But—so as not to lose sight of our ultimate goal in conducting research—we should remind ourselves that statistical manipulation of the data is *not,* in and of itself, research. Research goes one step further and demands *interpretation* of the data. In finding medians, means, interquartile ranges, or standard deviations, we have not interpreted the data, nor have we extracted any *meaning* from them. We have merely described the center and spread of the data. We have attempted only to see what the data look like. After learning their basic nature, we should then look for conditions that are forcing the data to behave as they do. For example, if we toss a pair of dice 100 times and one particular die yields a 5 in 80 of those tosses, we will have a distribution for that die much different from what we would expect. This may suggest to us that a reason lurks behind the particular results we have obtained. For example, perhaps we are playing with a loaded die!

[4] A standard score gaining increasing popularity for reporting academic achievement test results is the *NCE score,* which has a mean of 50 and a standard deviation of 21.06. With this particular (and seemingly very odd) standard deviation, an NCE score of 1 is equivalent to a percentile score of 1 and, likewise, an NCE score of 99 is equivalent to a percentile score of 99.

Measures of Association: Correlation

The statistics described so far—measures of central tendency and variability—involve only a single variable. Oftentimes, however, we also want to know whether two or more variables are in some way associated with one another. For example, relationships exist between age and reading ability (as illustrated in Figure 8.1 in Chapter 8), between emotional state and physical health, between the amount of rainfall and the price of vegetables in the marketplace. Consider, too, the relationships between temperature and pressure, between the intensity of light and the growth of plants, between the administration of a certain medication and the resulting platelet agglutination in the blood. Relationships among variables are everywhere. One function of statistics is to capture the nature and strength of such relationships.

The statistical process by which we discover whether two or more variables are in some way associated with one another is called *correlation.*The resulting statistic, called a **correlation coefficient**, is a number between -1 and $+1$; most correlation coefficients are decimals (either positive or negative) somewhere between these two extremes. A correlation coefficient for two variables simultaneously tells us two different things about the relationship between those variables:

- *Direction.* The direction of the relationship is indicated by the *sign* of the correlation coefficient—in other words, by whether the number is a positive or negative one. A positive number indicates a **positive correlation**: As one variable increases, the other variable also increases. For example, there is a positive correlation between self-esteem and school achievement: Students with higher self-esteem achieve at higher levels (e.g., Marsh, Gerlach, Trautwein, Lüdtke, & Brettschneider, 2007). In contrast, a negative number indicates an inverse relationship, or **negative correlation**: As one variable increases, the other variable *de*creases. For example, there is a negative correlation between the number of friends children have and the likelihood that they'll be victims of bullying: Children who have many friends are *less* likely to be bullied by their peers than are children who have few or no friends (e.g., Espelage & Swearer, 2004).
- *Strength.* The strength of the relationship is indicated by the *size* of the correlation coefficient. A correlation of $+1$ or -1 indicates a *perfect* correlation: If we know the degree to which one characteristic is present, we know exactly how much of the other characteristic exists. For example, if we know the length of a horseshoe crab in inches, we also know—or at least we can quickly calculate—exactly what its length is in centimeters. A number close to either $+1$ or -1 (e.g., $+.89$ or $-.76$) indicates a *strong* correlation: The two variables are closely related, such that knowing the level of one variable allows us to predict the level of the other variable with considerable accuracy. For example, we often find a strong relationship between two intelligence tests taken at the same time: People tend to get similar scores on both tests, especially if both tests cover similar kinds of content (e.g., McGrew, Flanagan, Zeith, & Vanderwood, 1997). In contrast, a number close to 0 (e.g., $+.15$ or $-.22$) indicates a *weak* correlation: Knowing the level of one variable allows us to predict the level of the other variable, but we cannot predict with much accuracy. For example, there is a weak relationship between intellectual giftedness and emotional adjustment: Generally speaking, people with higher IQ scores show greater emotional maturity than people with lower scores (e.g., Janos & Robinson, 1985), but many people are exceptions to this rule. Correlations in the middle range (for example, those in the .40s and .50s, positive or negative) indicate a *moderate* correlation.

The most widely used statistic for determining correlation is the Pearson product moment correlation, sometimes called the Pearson r. But there are numerous other correlation statistics as well. As in the case of the central tendency, the nature of the data determines the technique that is most appropriate for calculating correlation. In Table 11.4, we present several parametric and nonparametric correlational techniques and the kinds of data for which they are appropriate.

One especially noteworthy statistic in Table 11.4 is the *coefficient of determination,* or R^2. This statistic, which is the square of the Pearson r, tells us *how much of the variance is accounted for* by the correlation. Although you will see this expression used frequently in research reports, researchers usually don't stop to explain what it means. By *variance,* we are specifically referring to a particular

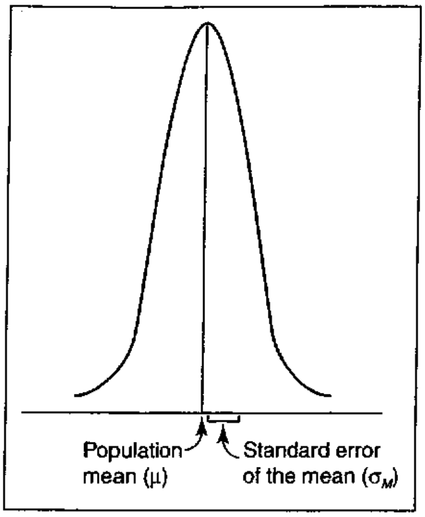
TABLE 11.4

Examples of correlational statistics

[illegible]	[illegible]	[illegible] Which It Is Appropriate
		Parametric Statistics
Pearson product moment correlation	r	Both variables involve continuous data.
Coefficient of determination	R^2	This is the square of the Pearson product moment correlation; thus, both variables involve continuous data.
Point biserial correlation	r_{pb}	One variable is continuous; the other involves discrete, dichotomous, and perhaps nominal data (e.g., Democrats vs. Republicans, males vs. females).
Biserial correlation	r_b	Both variables are continuous, but one has been artificially divided into an either-or dichotomy (e.g., "above freezing" vs. "below freezing," "pass" vs. "fail").
Phi coefficient	ϕ	Both variables are true dichotomies.
Triserial correlation	r_{tri}	One variable is continuous; the other is a trichotomy (e.g., "low," "medium," "high").
Partial correlation	$r_{12\cdot3}$	The relationship between two variables exists, in part, because of their relationships with a third variable, and the researcher wants to "factor out" the effects of this third variable (e.g., what is the relationship between motivation and student achievement when IQ is held constant statistically?).
Multiple correlation	$R_{1\cdot23}$	One variable is related to two or more variables; here the researcher wants to compute the first variable's *combined* relationship with the others.
		Nonparametric Statistics
Spearman rank order correlation (Spearman's rho)	ρ	Both variables involve rank-ordered data and so are ordinal in nature.
Kendall coefficient of concordance	W	Both variables involve rankings (e.g., rankings made by independent judges regarding a particular characteristic) and hence are ordinal data, and the researcher wants to determine the degree to which the rankings are similar.
Contingency coefficient	C	Both variables involve nominal data.
Kendall's tau correlation	τ	Both variables involve ordinal data; the statistic is especially useful for small sample sizes (e.g., $N < 10$).

measure of variability mentioned earlier: the square of the standard deviation, or s^2. For example, if we find that, in our data set, the R^2 between Variable 1 and Variable 2 is .30, we know that 30% of the variability in Variable 1 is reflected in its relationship with Variable 2. This knowledge will allow us to control for—and essentially *reduce*—some of the variability in our data set through such statistical procedures as partial correlation and analysis of covariance (described in Table 11.4 and later in Table 11.5, respectively). It is important to note, too, that the correlation statistics presented in Table 11.4 are all based on an important assumption: that the relationship between the two variables is a *linear* one—that is, as one variable continues to increase, the other continues to increase (for a positive correlation) or decrease (for a negative correlation). Not all relationships take a linear form, however. For example, consider the relationship between body mass index (a general measure of a person's body fat; often abbreviated as BMI) and anxiety. In one recent study (Scott, McGee, Wells, & Browne, 2008), researchers found that anxiety was highest in people who were either very underweight or very overweight; anxiety was lowest for people of relatively *average* weight. Such a relationship is known as a *U-shaped relationship* (see Figure 11.11). U-shaped and other nonlinear relationships can be detected through scatter plots and other graphic techniques, as well as through certain kinds of statistical analyses (e.g., see B. Thompson, 2008).

FIGURE 11.12
Distribution of sample means



Random samples from populations—please note the word *random* here—display roughly the same characteristics as the populations from which they were selected. Thus, we should expect the mean height for our sample to be approximately the same as the mean for the overall population. It will not be *exactly* the same, however. In fact, if we were to collect the heights for a second random sample of 200 boys, we would be likely to compute a slightly different mean than we had obtained for the first sample.

Different samples—even when each has been randomly selected from the same population—will almost certainly yield slightly different estimates of the overall population. The difference between the population mean and a sample mean constitutes an *error* in our estimation. Because we don't know what the exact population mean is, we also don't know how much error is in our estimate. We *do* know three things, however:

1. The means we might obtain from an infinite number of random samples form a normal distribution.
2. The *mean of this distribution of sample means* is equal to the mean of the population from which the samples have been drawn (μ). In other words, the population mean equals the average, or mean, of all the sample means.
3. The standard deviation of this distribution of sample means is directly related to the standard deviation of the characteristic in question for the overall population.

This situation is depicted in Figure 11.12.

The third characteristic just listed—the standard deviation for the distribution of sample means—is called the **standard error of the mean**. This index tells us how much the particular mean we calculate is likely to vary from one sample to another *when all samples are the same size and are drawn randomly from the same population.* Statistically, when all of the samples are of a particular size *(n),* the standard error of the mean is represented as

$$\sigma_M = \frac{\sigma}{\sqrt{n}}$$

Here we are faced at once with a problem. The formula we just presented involves using the population standard deviation (σ), but the purpose of using the sample was to *avoid* having to measure the entire population. Fortunately, statisticians have devised a way to estimate the standard error of the mean from the standard deviation of a *sample* drawn from the population. This formula is

$$\text{Estimated } \sigma_M = \frac{s}{\sqrt{n-1}}$$

Inferential Statistics

As mentioned earlier, inferential statistics allow us to draw inferences about large populations from relatively small samples. More specifically, inferential statistics have two main functions:

1. To estimate a population parameter from a random sample
2. To test statistically based hypotheses

In this text, we do not have the space to venture too far into these areas; statistics textbooks can give you more detailed information. However, we comment briefly about several general concepts and principles.

Estimating Population Parameters

When we conduct research, more often than not we use a sample to learn about the larger population from which the sample has been drawn. Typically we compute various statistics for the sample we have studied. Inferential statistics can tell us how closely these sample statistics approximate parameters of the overall population. For instance, we often want to estimate population parameters related to central tendency (the mean, or μ), variability (the standard deviation, or σ), and proportion (*P*). These values in the population compare with the *M* or $\overline{X}$, the *s*, and the *p* of the sample (see Table 11.1, page 278).

To show you what we mean by estimation, we use a simple illustration. Jan is a production manager for a large corporation. The corporation manufactures a piece of equipment that requires a connecting-rod pin, which the corporation also manufactures. The pin fits snugly into a particular joint in the equipment, permitting a metal arm to swivel within a given arc. The pin's diameter is critical: If the diameter is too small, the arm will wobble while turning; if it is too large, the arm will stick and refuse to budge. Jan has received complaints from customers that some of the pins are faulty. She decides to estimate, on the basis of a random sample of the connecting-rod pins, how many units of the equipment may have to be recalled in order to replace their faulty pins. From this sample, Jan wants to know three facts about the thousands of equipment units that have been manufactured and sold:

1. What is the average diameter of the pins?
2. How widely do the pins vary in diameter?
3. What proportion of the pins are acceptable in the equipment units already sold?

The problem is to determine population parameters on the basis of the sample statistics. From the sample, Jan can estimate the mean and variability of the pin diameters and the proportion of acceptable pins within the population universe. These are the values represented by μ, the σ, and the *P*.

Statistical estimates of population parameters are based on the assumption that *the sample is randomly chosen and representative of the total population.* Only when we have a random, representative sample can we make reasonable guesses about how closely our statistics estimate population parameters. To the extent that a sample is nonrandom and therefore nonrepresentative—to the extent that the sample's selection has been *biased* in some way—our statistics may be poor reflections of the population from which it has been drawn.

An Example: Estimating a Population Mean

Imagine that we want to estimate the average (mean) height of 10-year-old boys in the state of New Hampshire. Measuring the heights of the entire population would be incredibly time-consuming, so we decide to measure the heights of a random and presumably representative sample of, let's say, 200 boys.

Notice how, in both formulas, the standard error of the mean is directly related to the standard deviation of the characteristic being measured: More variability in the population leads to a larger standard error of the mean—that is, to greater variability in the sample means that we might obtain. In addition, the standard error is *inversely* related to *n,* the size of the sample. As the sample size increases, the standard error of the mean decreases. Thus, a larger sample size will give us a sample mean that more closely approximates the population mean. This principle holds true for estimates of other population parameters as well. In general, *larger samples yield more accurate estimates of population parameters.*

Point versus Interval Estimates

When using sample statistics to estimate population parameters, we can make two types of estimates: point estimates and interval estimates.

A **point estimate** is a single statistic that is used as a reasonable estimate of the corresponding population parameter; for instance, we might use a sample mean as a close approximation to the population mean. Although point estimates have the seeming benefit of being precise, in fact this precision is illusory. A point estimate typically does *not* correspond exactly with its equivalent in the population. Let's return to our previous example of the connecting-rod pins. Perhaps the company has produced 500,000 pins, and Jan has selected a sample of 100 of them. When she measures the diameters of these pins, she finds that the mean diameter is 0.712 centimeter, and the standard deviation is 0.020 centimeter. She guesses that the mean and standard deviation of the diameters of *all* of the pins are also 0.712 and 0.020, respectively. Her estimates will probably be close—and they are certainly better than nothing—but they won't necessarily be dead-on.

A more accurate approach—although still not 100% dependable—is to identify **interval estimates** of parameters. In particular, we specify a range within which a population parameter probably lies, and we state the probability that it actually lies there. Such an interval is often called a **confidence interval** because it attaches a certain level of probability to the estimate—a certain level of *confidence* that the estimated range includes the population parameter.

As an example, Jan might say that she is 95% certain that the mean of the 500,000 connecting-rod pin diameters her company has produced is somewhere between 0.708 and 0.716. What Jan has done is to determine that the standard error of the mean is 0.002 (see the previously presented formula for estimated σM). Jan knows that sample means fall in a normal distribution (look once again at Figure 11.12). She also knows that normal distributions have predictable proportions within each section of the curve (look once again at Figure 11.6). In particular, Jan knows that about 68% (34.1% + 34.1%) of the sample means lie within one standard error of the population mean, and that about 95% (13.6% + 34.1% + 34.1% + 13.6%) lie within two standard errors of the population mean. What she has done, then, is to go two standard errors (2 × 0.002, or 0.004) to either side of her sample mean (0.712) to arrive at her 95% confidence interval of 0.708 to 0.716.

We have said enough about estimation for you to appreciate its importance. To venture further would get us involved in specific statistical procedures that are not the province of this text. For additional guidance, we urge you to consult one or more statistics textbooks, such as those listed in the "For Further Reading" list at the end of the chapter.

Testing Hypotheses

The second major function of inferential statistics is to test hypotheses. At the outset, we should clarify our terminology. The term *hypothesis* can confuse you unless you understand that it has two different meanings in research literature. The first meaning relates to a *research hypothesis;* the second relates to a *statistical hypothesis.*

TABLE 11.3

Using measures of variability for different types of data

		Data for Which It's Appropriate
Range	The difference between the highest and lowest scores in the distribution	• Data on ordinal, interval, and ratio scales*
Interquartile range	The difference between the 25th and 75th percentiles	• Data on ordinal, interval, and ratio scales • Especially useful for highly skewed data
Standard deviation	$s = \sqrt{\frac{\Sigma(X-M)^2}{N}}$	• Data on interval and ratio scales • Most appropriate for normally distributed data
Variance	$s^2 = \frac{\Sigma(X-M)^2}{N}$	• Data on interval and ratio scales • Most appropriate for normally distributed data • Especially useful in inferential statistical procedures (e.g., analysis of variance)

* Measures of variability are usually inappropriate for nominal data. Instead, frequencies or percentages of each number are reported.

make them meaningful. For instance, if we say that Mary has gotten a score of 35 on a test of extroversion (i.e., on a test assessing her tendency to be socially outgoing), you might ask: What does that score *mean?* Is it high? Low? Somewhere in the middle? Without a context, a score of 35 has no meaning. We have no idea how extroverted or introverted Mary is.

Sometimes researchers provide context by converting raw scores to **norm-referenced scores**, scores that reflect where each person is positioned relative to other members of the person's group. We have already seen one example of a norm-referenced score: A *percentile rank* is the percentage of people in the group that a particular individual has scored *better than.* For example, if Mary scores at the 95th percentile on a test of extroversion, then we know that she is quite outgoing—more so than 95% of the people who have taken the test. But as noted earlier in the chapter, percentile ranks have a definite limitation: They are ordinal data rather than interval data, and so we cannot perform even such basic arithmetic operations as addition and subtraction on them. Accordingly, we will be very limited in the statistical procedures we can use with percentile ranks.

More useful in statistical analyses are standard scores. Simply put, a **standard score** tells us how far an individual's performance is from the mean with respect to standard deviation units. The simplest standard score is a *z*-score, which is calculated by using an individual's raw score (which we will symbolize as *X*), along with the mean and standard deviation for the entire group, as follows:

$$z = \frac{X - M}{s}$$

As an illustration, let's return to Mary's score of 35 on the extroversion test. If the mean of the scores on this test is 25, and if the standard deviation is 5, we would calculate Mary's *z*-score as follows:

$$z = \frac{35 - 25}{5} = \frac{10}{5} = 2$$

When we calculate *z*-scores for an entire group, we get a distribution that has a mean of 0 and a standard deviation of 1.

Because about half of the *z*-scores for any group of people will be a negative number—as just noted, the *mean* for the group is 0—researchers sometimes change *z*-scores into other standard-score scales that yield only positive numbers. To convert a *z*-score to another scale, we

Most of our discussions of hypotheses in earlier chapters have involved the first meaning of the word *hypothesis* (e.g., recall Chapter 1's discussion of the homeowner who speculates about why a table lamp may have failed). In forming a research hypothesis, a researcher speculates about how the research problem or one of its subproblems might be resolved. A research hypothesis is a reasonable conjecture, an educated guess, a theoretically or empirically based prediction. Its purpose is a practical one: It provides a temporary objective, an operational target, a logical framework that guides a researcher as he or she collects and analyzes data.

When we encounter the phrase "testing a hypothesis," however, the matter is entirely different. Here the word *hypothesis* refers to a statistical hypothesis, usually a *null* hypothesis. A **null hypothesis** (often symbolized as H_0) postulates that any result observed is the result of chance alone. For instance, if we were to compare the means of two groups, our null hypothesis would be that both groups are parts of the same population and that any differences between them—including any difference we see between their means—are strictly the result of the fact that *any* two samples from the population will yield slightly different estimates of a population parameter.

Now let's say that we look at the *probability* that our result is due to chance alone. If, for example, we find that a difference between two group means would, if due entirely to chance, occur *only one time in a thousand,* we could reasonably conclude that the difference is *not* due to chance—that, instead, something in the situation we are studying (perhaps an experimental treatment we have imposed) is systematically leading to a difference in the groups' means. This process of comparing observed data with the results that we would expect from chance alone is called *testing the null hypothesis.*

At what point do researchers decide that a result has *not* occurred by chance alone? One common cutoff is a 1-in-20 probability: Any result that would occur by chance only 5% of the time—that is, a result that would occur, on average, only one time in every 20 times—probably is *not* due to chance but instead to another, systematic factor that is influencing the data. Other researchers use a more rigorous 1-in-100 criterion: The observed result would occur by chance only one time in 100. The probability that researchers use as their cutoff point, whether .05, .01, or some other figure, is the **significance level**, or **alpha** (α). A result that, based on this criterion, we deem *not* to be due to chance is called a **statistically significant** result. When we decide that a result is due to something other than chance, we *reject the null hypothesis.*

In the "Results" section of a research report, you will often see the researcher's alpha level implied in parentheses. For example, imagine that a researcher reports that "a t-test revealed significantly different means for the two treatment groups ($p < .01$)." The "$p < .01$" here means that the difference in means for the two groups would occur by chance less than one time in 100 *if* the two groups had been drawn from the same population. Sometimes, instead, a researcher will state the actual probability with which a result might occur by chance alone. For example, a researcher might report that "a t-test revealed significantly different means for the two treatment groups ($p = .003$)." The "$p = .003$" here means that a difference this large would occur only three times in 1,000 for two groups that come from the same population. In this situation, then, chances are good that the two groups come from *different* populations—a round-about way of saying that the two treatments differentially affected the outcome.

When we reject the null hypothesis, we must look to an alternative hypothesis—which may, in fact, be the *research hypothesis*—as being more probable. For example, if our null hypothesis is that two groups are the same and we then obtain data that lead us to reject this hypothesis, we indirectly support the opposite hypothesis: that the two groups are *different.*

In brief, we permit a certain narrow margin of variation within our data, which we deem to be natural and the result of pure chance. Any variation within this statistically permissible range is not considered to be important enough to claim our attention. Whatever exceeds these limits, however, is considered to be the result of some determinative factor other than chance, and so the influence is considered to be an important one. The term *significant,* in the statistical sense in which we have been using it, is close to its etymological meaning—namely, "giving a signal" that certain dynamics are operating within the data and merit attention.

Making Errors in Hypothesis Testing

It is possible, of course, that we may make a mistake when we decide that a particular result is not the result of chance alone. In fact, *any* result could conceivably be due to chance; our sample, although selected randomly, may be a fluke that displays atypical characteristics simply through the luck of the draw. If we erroneously conclude that a result was not due to chance when in fact it *was* due to chance—if we incorrectly reject the null hypothesis—we are making a **Type I error** (also called an *alpha error*).

Yet in another situation, we might conclude that a result is due to chance when in fact it is *not*. In such a circumstance, we have failed to reject a null hypothesis that is actually false—something known as a **Type II error** (also called a *beta error*). For example, imagine that we are testing the relative effects of a new medication versus the effects of a placebo in lowering blood cholesterol. Perhaps we find that people who have been taking the new medication have, on average, a lower cholesterol level than people taking the placebo, but the difference is a small one. We might find that such a difference could occur 25 times out of 100 due to chance alone, and so we *retain the null hypothesis*. If, in actuality, the medication does reduce cholesterol more than a placebo does, we have made a Type II error.

Statistical hypothesis testing is all a matter of probabilities, and there is always the chance that we could make either a Type I or Type II error. We can decrease the odds of making a Type I error by lowering our level of significance, say, from .05 to .01, or perhaps to an even lower level. In the process of doing so, however, we increase the likelihood that we will make a Type II error—that we will fail to reject a null hypothesis that is, in fact, incorrect. To decrease the probability of a Type II error, we would have to increase our significance level (α), which, because it increases the odds of rejecting the null hypothesis, also increases the probability of a Type I error. Obviously, then, there is a trade-off between Type I and Type II errors: Whenever you decrease the risk of making one, you increase the risk of making the other.

To illustrate this trade-off, we return to our study of the potentially cholesterol-reducing medication. There are four possibilities:

1. We correctly conclude that the medication reduces cholesterol.
2. We correctly conclude that it does not reduce cholesterol.
3. We mistakenly conclude that it is effective when it *isn't*.
4. We mistakenly conclude that it isn't effective when it *is*.

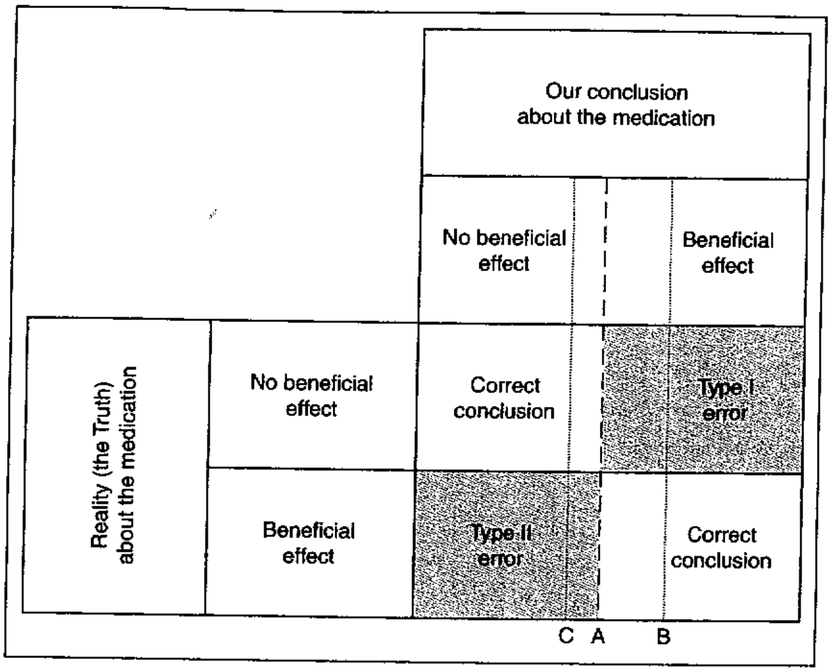
These four possibilities are illustrated in Figure 11.13. The three vertical lines illustrate three hypothetical significance levels we might choose. Imagine that the dashed middle line, Line A, represents a significance level of, say, .05. In this particular situation (such will not always be the case), we have a slightly greater chance of making a Type I error (represented by the upper shaded area) than of making a Type II error (represented by the lower shaded area). But the significance level we choose is an arbitrary one. We could reduce our chance of a Type I error by decreasing our significance level to, say, .03. Line B to the right of Line A in the figure represents such a change; notice how it would create a smaller box (lower probability) for a Type I error but create a larger box (greater probability) for a Type II error. Alternatively, if we raise the significance level to, say, .06 (as might be represented by Line C, to the left of Line A in the figure), we decrease the probability of a Type II error but increase the probability of a Type I error.

There is perhaps nothing more frustrating for the novice researcher than obtaining insignificant results—those that, from a statistical standpoint, could have been due to chance alone. Following are three suggestions for decreasing the likelihood of making a Type II error and thereby increasing the likelihood of correctly rejecting an incorrect null hypothesis. In other words, these are suggestions for increasing the **power** of a statistical test:

- *Use as large a sample size as is reasonably possible.* The larger the sample, the less the statistics you compute will diverge from actual population parameters.[5]

[5] Formulas exist for computing the power of statistical procedures for varying sample sizes. For example, see Lipsey (1990) or Murphy, Myors, and Wolach (2009).

FIGURE 11.13
The trade-off between Type I and Type II errors



- *Maximize the validity and reliability of your measures.* Measures of variables in a research study rarely have perfect (100%) validity and reliability, but some measures are more valid and reliable than others. Research projects that use measures with high validity and reliability are more likely to yield statistically significant results. (Again we refer you to the section "Validity and Reliability in Measurement" in Chapter 4.)
- *Use parametric rather than nonparametric statistics whenever possible.* As a general rule, nonparametric statistical procedures are less powerful than parametric techniques. By "less powerful," we mean that nonparametric statistics typically require larger samples to yield results that enable the researcher to reject the null hypothesis. When characteristics of the data meet the assumptions for parametric statistics, then, we urge you to use these statistics. (Look once again at the section "Choosing between Parametric and Nonparametric Statistics" earlier in this chapter.)

It is important—in fact, critical—to keep in mind that *whenever we test more than one statistical hypothesis, we increase the probability of making at least one Type I error.* Let's say that, for a particular research project, we have set the significance level at .05, such that we will reject the null hypothesis whenever we obtain results that would be due to chance alone only 1 time in 20. And now let's say that as we analyze our data, we perform 20 different statistical tests, always setting α at .05. In this situation, although we won't necessarily make a Type I error, the odds are fairly high that we will.[6]

Another Look at Statistical Hypotheses versus Research Hypotheses

Novice researchers sometimes become so wrapped up in their statistical analyses that they lose track of their overall research problem or hypothesis. In fact, testing a null hypothesis involves

[6] When testing 20 hypotheses at a .05 significance level, the probability of making at least one Type I error is .642—in other words, chances are better than 50-50 that at least one Type 1 error is being made. In general, the probability of making a Type I error when conducting multiple statistical tests is $1 - (1 - \alpha)^n$, where α (alpha) is the significance level and n is the number of tests conducted.

nothing more than a statistical comparison of two distributions of data—one hypothetical (a theoretical ideal) and one real (the distribution of data collected from a research sample). A researcher simply uses one or more statistical procedures to determine whether calculated values sufficiently diverge from the statistical ideal to reject the null hypothesis.

Testing a statistical hypothesis does not, in and of itself, contribute much to the fulfillment of the basic aim of research: a systematic quest for undiscovered knowledge. Certainly statistical analyses are invaluable tools that enable us to find patterns in the data and thus help us detect possible dynamics working within the data. But we must never stop with statistical procedures that yield one or more numerical values. We must also *interpret* those values and give them meaning. The latter process includes the former, but the two should never be confused.

It is often the case that the statistical hypothesis is the opposite of the research hypothesis. For example, we might, as our research hypothesis, propose that two groups are different from one another. As we begin our statistical analysis, we set out to test the statistical hypothesis that the two groups are the same. *By disconfirming the null hypothesis, we indirectly find support for our research hypothesis.* This is, to be sure, a backdoor approach to finding evidence for a research hypothesis, yet it is the approach that is typically taken. The reasons for this approach are too complex to be dealt with in a text such as this one. Suffice it to say that it is mathematically much easier to test a hypothesis that an equivalence exists than to test a hypothesis that a difference exists.

Examples of Statistical Techniques for Testing Hypotheses

Table 11.5 lists many commonly used parametric and nonparametric statistical techniques for testing hypotheses. We hope it will help you make decisions about the techniques that are most appropriate for your own research situation. As you can see in the table, however, nonparametric techniques exist only for relatively simple statistical analyses, such as comparing measures of central tendency or testing the statistical significance of correlations. When your research problem calls for a relatively sophisticated analysis (e.g., multiple regression or structural equation modeling), parametric statistical procedures—and the underlying assumptions about the data these procedures require—are your only viable option.

We urge you to consult one or more statistics texts to learn as much as you can about whatever statistical procedures you use. Better still, enroll in one or more statistics courses! You can successfully solve your research problem only if you apply statistical procedures appropriately and thereby conduct accurate analyses of your data.

Meta-Analysis

Occasionally researchers use inferential statistics not to analyze and draw conclusions from data they have collected but instead to analyze and draw conclusions about *other researchers' statistical analyses.* Such analysis of analyses is known as **meta-analysis**. A meta-analysis is most useful when many studies have already been conducted on a particular topic or research problem and another researcher wants to pull all of the results together into a neat and mathematically concise package.

The traditional approach to synthesizing previous studies related to a particular research question is simply to describe them all, pointing out which studies yield which conclusions, which studies contradict others, and so on. In a meta-analysis, however, the researcher integrates the studies statistically rather than verbally. After pinning down the research problem, the researcher:

1. *Conducts a fairly extensive search for relevant studies.* The researcher does not choose arbitrarily among studies that have been reported about the research problem. Instead, he or she uses some systematic and far-reaching approach (e.g., searching in several prespecified professional journals, using certain keywords in a search of online databases) to identify studies that have addressed the topic of interest.

TABLE 11.5

Examples of inferential statistical procedures and their purposes

	Purpose
	Parametric Statistics
Student's *t*-test	To determine whether a statistically significant difference exists between two means. A *t*-test takes slightly different forms depending on whether the two means come from separate, independent groups (an *independent-samples t*-test) or, instead, from a single group or two interrelated groups (a *dependent-samples t*-test).
Analysis of variance (ANOVA)	To examine differences among three or more means by comparing the variances (s^2) both within and across groups. As is true for *t*-tests, ANOVAs take slightly different forms for separate, independent groups and for a single group; in the latter case, a *repeated-measures* ANOVA is called for. If an ANOVA yields a significant result (i.e., a significant value for *F*), you should follow up by comparing various pairs of means using a *post hoc comparison of means.*
Analysis of covariance (ANCOVA)	To look for differences among means while controlling for the effects of a variable that is correlated with the dependent variable (the former variable is called a *covariate*). This technique can be statistically more powerful than ANOVA (i.e., it decreases the probability of a Type II error).
t-test for a correlation coefficient	To determine whether a Pearson product moment correlation coefficient (*r*) is larger than would be expected from chance alone.
Regression	To examine how effectively one or more variables allow(s) you to predict the value of another (dependent) variable. A *simple linear regression* generates an equation in which a single independent variable yields a prediction for the dependent variable. A *multiple linear regression* yields an equation in which two or more independent variables are used to predict the dependent variable.
Factor analysis	To examine the correlations among a number of variables and identify clusters of highly interrelated variables that reflect underlying themes, or *factors,* within the data.
Structural equation modeling (SEM)	To examine the correlations among a number of variables—often with different variables measured for a single group of people at different points in time—in order to identify possible causal relationships (*paths*) among the variables. SEM encompasses such techniques as *path analysis* and *confirmatory analysis* and is typically used to test a previously hypothesized model of how variables are causally interrelated. SEM enables a researcher to identify a *mediator* in a relationship: a third variable that may help explain why Variable A seemingly leads to Variable B (i.e., Variable A affects the mediating variable, which in turn affects Variable B). SEM also enables a researcher to identify a *moderator* of a relationship: a third variable that alters the nature of the relationship between Variables A and B (e.g., Variables A and B might be correlated when the moderating variable is high but not when it is low, or vice versa). (Mediating and moderating variables are discussed in more detail in Chapter 2.) When using SEM, the researcher must keep in mind that the data are *correlational* in nature; thus, any conclusions about cause-and-effect relationships are speculative at best.
	Nonparametric Statistics
Mann-Whitney *U*	To compare the medians of two groups when the data are ordinal rather than interval in nature. This procedure is the nonparametric counterpart of the independent-samples *t*-test in parametric statistics.
Kruskal-Wallis test	To compare three or more group medians when the data are ordinal rather than interval in nature. This procedure is the nonparametric counterpart of ANOVA.
Wilcoxon signed-rank test	To compare the medians of two correlated variables when the data are ordinal rather than interval in nature. This procedure is a nonparametric equivalent of a dependent-samples *t*-test in parametric statistics.
Chi-square (χ^2) goodness-of-fit test	To determine how closely observed frequencies or probabilities match expected frequencies or probabilities. A chi-square can be computed for nominal, ordinal, interval, or ratio data.
Odds ratio	To determine whether two dichotomous nominal variables (e.g., smokers vs. nonsmokers and presence vs. absence of heart disease) are significantly correlated. This is one nonparametric alternative to a *t*-test for Pearson's *r*.
Fisher's exact test	To determine whether two dichotomous variables (nominal or ordinal) are significantly correlated when the sample sizes are quite small (e.g., $n < 30$). This is another nonparametric alternative to a *t*-test for Pearson's *r*.

2. *Identifies appropriate studies to include in the meta-analysis.* The researcher limits the chosen studies to those that involve a particular experimental treatment (in experimental studies), pre-existing condition (in ex post facto studies), or other variable that is the focus of the meta-analysis. He or she may further restrict the chosen studies to those that involve particular populations, settings, assessment instruments, or other factors that may impact a study's outcome.
3. *Converts each study's results to a common statistical index.* Previous researchers may possibly have used different statistical procedures to analyze their data. For example, if each researcher has compared two or more groups that received two or more different experimental interventions, one investigator may have used a *t*-test, another may have conducted an analysis of variance, and a third may have conducted a multiple regression. The meta-analytic researcher's job is to find a common denominator here. Typically, when an experimental intervention has been studied, an **effect size** is calculated for each study; that is, the researcher determines how much of a difference the intervention makes (in terms of standard deviation units) in each study. The effect sizes of all of the studies are then used to compute an average effect size for that intervention.

The statistical procedures used in meta-analyses vary widely, depending, in part, on the research designs of the included studies; for instance, correlational studies require different meta-analytic procedures than experimental studies. We must point out, too, that meta-analyses, although they can make an important contribution to the knowledge bases of many disciplines, are not for the mathematically fainthearted. If you are interested in conducting a meta-analysis, several of the resources listed in the "For Further Reading" section at the end of this chapter should prove helpful.

Using Statistical Software Packages

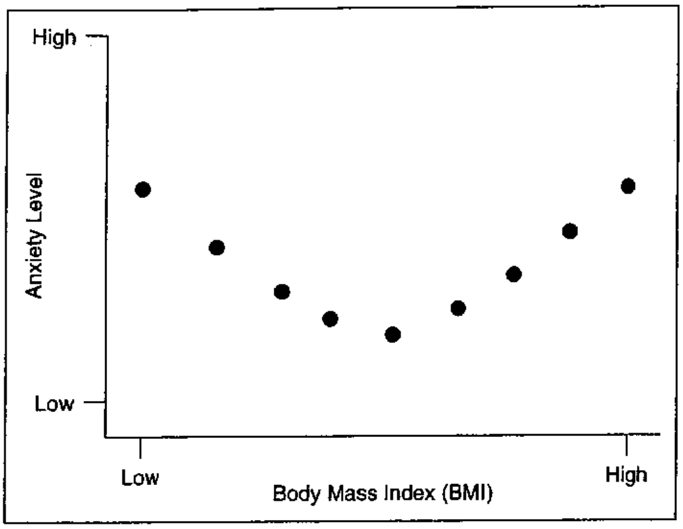
USING TECHNOLOGY

Earlier in the chapter, we mentioned that general-purpose spreadsheet programs can be used to describe and analyze sets of quantitative data. However, many spreadsheets are limited in their statistical analysis capabilities. As an alternative, you may want to consider using one of the several statistical software packages now widely available for use on personal computers (e.g., SPSS, SAS, SYSTAT, Minitab, Statistica). Such packages have several advantages:

- *Increased user-friendliness.* As statistical software programs become increasingly powerful, they also become more user-friendly. In most cases, the programs are logical and easy to follow, and results are presented in easy-to-read table format. Selection of the proper statistics and interpretation of the results, however, are still left to the researcher.
- *Range of available statistics.* Many of these programs include a wide variety of statistical procedures, and they can easily handle large data sets, multiple variables, and missing data points.
- *Assumption testing.* A common feature of statistical software packages is to test for characteristics (e.g., skewness, kurtosis) that might violate the assumptions on which a parametric statistical procedure is based.
- *Speed of completion.* As always, a major benefit of using the computer is the speed with which it accomplishes tasks. Even relatively simple statistical procedures might take several hours if executed by hand; more complex analyses are, for all practical purposes, impossible for a researcher to conduct using only paper, pencil, and a hand-held calculator.
- *Graphics.* Many statistical programs allow the researcher to summarize and display data in tables, pie charts, bar graphs, or other graphics.

FIGURE 11.11
U-shaped relationship between body mass index (BMI) and anxiety

Based on Scott et al., 2008.



Always keep in mind that the nature of the data governs the correlational procedure that is appropriate for those data. Don't forget the cardinal rule: *Look at the data!* Determine their nature, scrutinize their characteristics, and then select the correlational technique suitable for the type of data with which you are working.

How Validity and Reliability Affect Correlation Coefficients

Beginning researchers should be aware that the extent to which one finds a statistical correlation between two characteristics depends, in part, on how well those characteristics have been measured. Even if there really *is* a correlation between two variables, a researcher won't necessarily find one if the measurement instruments he or she uses have poor validity and reliability. For instance, we are less likely to find a correlation between age and reading level if the reading test we use is neither a valid (accurate) nor reliable (consistent) measure of reading achievement.

Over the years, we authors have had many students find disappointingly low correlation coefficients between two variables that they hypothesized would be highly correlated. By looking at the correlation coefficient alone, a researcher cannot determine the reason for a low correlation any more than he or she can determine the reason for a high one. Yet one thing is certain: *You will find substantial correlations between two characteristics only if you can measure both characteristics with a reasonable degree of validity and reliability.* We refer you back to the section "Validity and Reliability in Measurement" in Chapter 4, where you can find strategies for determining and enhancing both of these essential qualities of sound measurement.

A Reminder about Correlation

Whenever you find evidence of a correlation within your data, you must remember one important point: *Correlation does not necessarily indicate causation.* For example, if you find a correlation between self-esteem and classroom achievement, you cannot necessarily conclude that students' self-esteem *influences* their achievement. Only experimental studies, such as those described in Chapter 9, allow you to draw definitive conclusions about the extent to which one thing causes or influences another.

Finding a correlation in a data set is equivalent to discovering a signpost. That signpost points to the fact that two variables are associated, and it reveals the nature of the association (positive or negative, strong or weak). It should then lead you to wonder, What is the underlying reason for the association? But the statistic alone will not be able to answer that question.

In Appendix B, we show you some of the basics of one statistical software program, SPSS, and use a small data set to illustrate some of the ways you might use it.[7]

For frugal researchers—especially those whose research problems require small data sets and relatively simple statistical procedures (e.g., computing standard deviations, correlation coefficients, or chi-squares)—online statistics calculators provide another option. Two examples are www.easycalculation.com (www.easycalculation.com/statistics/statistics.php) and GraphPad Software's QuickCalcs (www.graphpad.com/quickcalcs). A Google or Yahoo! search for "online statistics calculator" can identify other helpful websites as well.

Yet we must caution you: *A computer cannot and should not do it all for you.* You may be able to perform sophisticated calculations related to dozens of statistical tests and present the results in a variety of ways, but if you do not understand how the results relate to your research problem, or if you cannot otherwise make logical, theoretical, or pragmatic sense of what your analyses have revealed, then all your efforts have been for naught. Powerful statistical software programs make it all too easy to conduct studies so large and complex that the researcher loses sight of the initial research question. In the words of Krathwohl (1993), the researcher eventually behaves "like a worker in a laboratory handling radioactive material, . . . manipulating mechanical hands by remote control from a room outside a sealed data container. With no sense of the data, there is little basis for suspecting an absurd result, and we are at the mercy of the computer printout" (p. 608).

Ultimately *you* must be in control of your analyses; you must know what calculations are being performed and why. Only by having an intimate knowledge of the data can you derive true meaning from the statistics computed and use them to address your research problem.

Interpreting the Data

To the novice researcher, statistics can be like the voice of a bevy of sirens. For those who have never studied or have forgotten the works of Homer, the *Odyssey* describes the perilous straits between Scylla and Charybdis. On these treacherous rocks resided a group of Sirens—svelte maidens who, with enticing songs, lured sailors in their direction and, by so doing, caused ships to drift and founder on the jagged shores.

For many beginning researchers, statistics hold a similar appeal. Subjecting data to elegant statistical routines may lure novice researchers into thinking they have made a substantial discovery, when in fact they have only calculated a few numbers. Behind every statistic lies a sizable body of data; the statistic may summarize these data in a particular way, but it cannot capture all the nuances of the data. The entire body of data collected—not any single statistic calculated—is what ultimately must be used to resolve the research problem. There is no substitute for the task the researcher ultimately faces: to discover the meaning of the data and its relevance to the research problem. Any statistical process you may employ is only ancillary to this central quest.

At the beginning of the chapter, we presented a hypothetical data set for 11 school children and discovered that the 5 girls in the sample had higher reading achievement test scores than the 6 boys. Shortly thereafter, we presented actual data about growth marks on the shells of the chambered nautilus. Perhaps these examples piqued your curiosity. For instance, perhaps you wondered about questions such as these:

- Why were all of the girls' scores higher than those of the boys?
- Why were the intervals between each of the scores equidistant for both boys and girls?
- What caused the nautilus to record a growth mark each day of the lunar month?
- Is the relationship between the forming of the partitions and the lunar cycle singular to the nautilus, or are there other similar occurrences in nature?

[7] At the instructor's request, this book can be packaged with the Student Version of SPSS at a discount; the CD for the software provides versions for both Windows and Macintosh users. Please contact your local Pearson representative if you are an instructor who is interested in setting up such a package for your students.

Knowledge springs from questions like these. But we must be careful not to make snap judgments about the data we have collected. It is all too easy to draw hasty and unwarranted conclusions. Even the most thorough research effort can go astray at the point of drawing conclusions from the data.

For example, from our study of 11 children and their reading achievement scores, we might conclude that girls read better than boys. But if we do so, we are not thinking carefully about the data. Reading is a complex and multifaceted skill. The data *do not* say that girls read better than boys. The data *do* say that, on a particular test given on a particular day to a particular group of 11 children, all girls' scores were higher than all boys' scores and that, for both boys and girls, the individual scores differed by intervals of 4. The apparent excellence of the girls over the boys was limited to test performance in those reading skills that were specifically measured by the test. Honesty and precision dictate that all conditions in the situation be considered and that we make generalizations only in strict accordance with the data. On the following day, the same test given to another 11 children might yield different results.

In general, interpreting the data means several things:

1. *Relating the findings to the original research problem and to the specific research questions and hypotheses.* Researchers must eventually come full circle to their starting point—why they conducted a research study in the first place and what they hoped to discover—and relate their results to their initial concerns and questions.
2. *Relating the findings to pre-existing literature, concepts, theories, and research studies.* To be useful, research findings must in some way be connected to the larger picture—to what people already know or believe about the topic in question. Perhaps the new findings confirm a current theoretical perspective, perhaps they cast doubt on common "knowledge," or perhaps they simply raise new questions that must be addressed before humankind can truly understand the phenomenon in question.
3. *Determining whether the findings have practical significance as well as statistical significance.* Statistical significance is one thing; practical significance—whether findings are actually useful—is something else altogether. For example, let's return to that new medication for lowering blood cholesterol level mentioned earlier in the chapter. Perhaps we randomly assign a large sample of individuals to one of two groups; one is given the medication, and the other is given a placebo. At the end of the study, we measure cholesterol levels for the two groups and then conduct a *t*-test to compare the group means. If our sample size is quite large, the standard error of the mean will be very small, and we may therefore find that even a minor difference in the cholesterol levels of the two groups is statistically significant. Is the difference *practically* significant as well? That is, do the benefits of the medication outweigh its costs and any unpleasant side effects? A calculation of *effect size*—how different the cholesterol levels are for the treatment and control groups relative to the standard deviation for one or both groups—can certainly help us as we struggle with this issue. But ultimately a statistical test cannot, in and of itself, answer the question. Only the human mind—the researcher, practitioners in the field of medicine, and so on—can answer it.
4. *Identifying limitations of the study.* Finally, interpreting the data involves outlining the weaknesses of the study that yielded them. No research study can be perfect, and its imperfections inevitably cast at least a hint of doubt on its findings. Good researchers know—and also report—the weaknesses along with the strengths of their research.

A Sample Dissertation

To illustrate this final step in the research process—interpretation of the data—we present excerpts from Kimberly Mitchell's doctoral dissertation in psychology conducted at the University of Rhode Island (Mitchell, 1998). The researcher was interested in identifying possible causal factors leading to eating disorders and substance abuse, and she hypothesized that family dynamics and child abuse might be among those factors. She drew on three theoretical perspectives that potentially had relevance to her research question: problem behavior theory, social cognitive theory, and the theory of cognitive adaptation. She administered several surveys to a large sample of undergraduate students and obtained a large body of correlational data about the students' childhoods, eating habits, drug use, and so on. She then used ***structural equation modeling*** (described briefly in Table 11.5) as a means of revealing possible—we must emphasize the word *possible*—cause-and-effect relationships in her data set.

The dissertation refers to several psychological theories and concepts with which many of our readers may not be familiar. Nevertheless, as you read the excerpts, you should be able to see how the author frequently moves back and forth between her results and the broader theoretical framework. We pick up the dissertation at the point where Mitchell begins to summarize and interpret her results.

DISSERTATION ANALYSIS 8

DISCUSSION

Summary of Results and Integration

The purpose of this study was to integrate several theories that are beneficial for understanding health-risk behaviors. Problem Behavior Theory (Jessor, 1987), Social Cognitive Theory. . . . (Bandura, 1977a), and the Theory of Cognitive Adaptation (Taylor, 1983) are similar in that they all pose a cognitive component within the individual that is crucial to overcome the potential negative consequences of life stressors. . . . This study supports these three theories, as well as previous research in the field. It extends the research by linking these theories into a single comprehensible framework for understanding the link between the childhood stressors of sexual abuse and negative family functioning and adult substance misuse of alcohol, illicit drugs, and eating.

A series of structural equation models revealed the powerful impact individuals' perceptions of their confidence and their interactions with their environment play on health-risk behavior. The first three models examined various ways childhood stressors (sexual abuse and family functioning) could predict current health-risk behaviors (alcohol use, illicit drug use, and binge eating). Examination of the first three models (Full, Direct, and Mediational) and chi-square difference tests revealed that the mediators (self-efficacy, life satisfaction, and coping) are extremely important in predicting health-risk behaviors. This [finding] supports Jessor's (1987) theory that problem behavior is the result of the interaction of the personality system, perceived environment, and the behavioral system. The personality system is measured by the cognitive mediator constructs; the perceived environment by the family functioning construct; and the behavioral system by the outcome constructs. . . . [T]he socialization an individual encounters throughout childhood through interactions with family members appears to influence both how the individual perceives the self and the environment around him/her. These factors seem to propel individuals to behave in ways that may or may not be risky for their health.

Comments

The author capitalizes the names of the three theories. More often, researchers use lowercase letters when referring to particular theoretical perspectives. Either approach is acceptable as long as the author is consistent.

Notice how the author begins with a "grand conclusion" of sorts, which she supports in subsequent paragraphs. She also explains how she has expanded on existing theories by integrating them to explain the phenomenon she has studied.

The "models" she refers to here are multivariable flowcharts that reflect how some variables may influence other variables, perhaps directly or perhaps indirectly through additional, mediator variables.

Self-efficacy *refers to people's confidence in their ability to perform a task (e.g., resist the temptation to abuse alcohol) successfully. It is a central concept in Bandura's social cognitive theory, one the three theoretical frameworks on which the author bases her study.*

Furthermore, Jessor (1987) suggests that problem behaviors in which adolescents engage are interrelated and co-vary. Donovan and Jessor (1985) suggest that diverse problem behavior, such as alcohol abuse, risky sexual behavior, and drug use constitute a single behavioral syndrome. The current study supports this notion. All of the structural models revealed a positive relationship between alcohol and drug use, as well as a positive relationship between drug use and binge eating. Although the relationship between alcohol use and binge eating was not found to be significant, they are indirectly related through drug use. Such relationships support the idea that these health-risk behaviors constitute a single behavioral syndrome. Future research with a longitudinal design is needed to see if there is a linear trend among these variables....

[The author continues with a discussion of more specific aspects of her findings and their relevance to the three theoretical frameworks. We pick up her discussion again when she summarizes her conclusions.]

Summary of Conclusions

There are several conclusions that can be drawn from this study. First, in support of Problem Behavior Theory (Jessor, 1987), health-risk behaviors may be part of a single behavioral syndrome. The consistent relationships found throughout the models between alcohol use and drug use, as well as [between] drug use and binge eating, reveal the presence of a higher order behavioral syndrome.=

Second, there is a complex relationship between child sexual abuse and family functioning in terms of their ability to predict life satisfaction, coping, and self-efficacy. While child sexual abuse was found to significantly predict coping and life satisfaction, the inclusion of family functioning into the model made these paths disappear. The initial finding indicates a confounding of child sexual abuse and family functioning rather than sexual abuse itself. Furthermore, the constant relationship between child sexual abuse and family functioning shows that, although child sexual abuse does not directly predict the mediator constructs, it plays a role in the prediction indirectly.

Third, family functioning and cognitive mediators interact in specific and consistent ways to determine health-risk behaviors. Those students with high levels of family functioning are likely to have high life satisfaction, more effective coping strategies, and higher self-efficacy for alcohol use, drug use, and eating. In turn, these cognitive factors interact to predict health-risk behavior.

[The author continues with additional conclusions, and then turns to the limitations of her study.]

Study Limitations

The present study offers several important findings to the literature. Yet, there are some limitations to the study as well. First, the design was cross-sectional rather than longitudinal. Structural equation modeling is a multivariate technique that is well utilized with longitudinal data (Maruyama, 1998). By incorporating longitudinal data into the overall design, one can begin to establish causality in the results. The use of cross-sectional data with this sample does not allow the researcher to make causal statements about the findings. For example, the data cannot tell us whether self-efficacy for

Notice how the author continually connects her findings with the theoretical frameworks she is using.

Here the author points out both what she has found and what she has not found.

Although the author has previously presented each of her conclusions, she summarizes them all here. Such a summary is typical of lengthy research reports. It is quite helpful to readers, who might easily lose track of some important conclusions as they read earlier portions of a report.

The author makes the point that two of her independent (predictor) variables, child sexual abuse and family functioning, are highly interrelated. Their strong correlation is reflected in the models identified through her structural equation modeling procedures.

The author's use of the term cross-sectional *is somewhat different from our use of it in Chapter 8. She simply means that she collected all data from her sample at one time, rather than following the sample over a lengthy period and collecting data at two or more times. As the author states, a longitudinal design would have better enabled her to identify important factors that preceded—and so may have had a causal effect on—other factors.*

alcohol use comes before actual alcohol use or vice versa. Furthermore, the study asks the participant to answer a portion of the survey retrospectively, such as [is true for] the child sexual abuse and family functioning items. This brings up problems with how reliable the responses are due to the length of time that has passed between the incident(s) in question and the time of the study.....

The author points out a problem with using surveys to learn about people's prior life experiences: Human memory is not always accurate. Her use of the word reliable *here refers to accuracy and dependability (i.e., validity) of the results, rather than to reliability as we have previously defined the term.*

A second limitation to this study is the nature of the sample itself. Although the sample size is excellent (n=469), there were disproportionate numbers of men and women (125 and 344, respectively). Furthermore, the sample was extremely homogeneous (87% White; 91% freshman or sophomore; 74% with family income over $35,000; and 73% Catholic or Protestant). This degree of similarity among participants limits the generalizability of the study results to other populations. Yet the results are still important because this is a population at high risk for alcohol use, drug use, and bulimia-related binge eating.

The author explains ways in which her sample was not completely representative of the overall population of older adolescents and young adults but also makes a good case for the value of studying this sample.

Another limitation to this study is the lack of response to the probing sexual abuse questions. Approximately one half of the 91 students who reported sexual abuse did not respond to the in-depth questions regarding the abuse experience(s) (e.g., degree of trust with perpetrator, frequency of abuse). This could be due to the nature of the survey itself or [to] the environment in which students filled out the survey. In terms of the nature of the survey, once students responded to the overall sexual abuse questions geared to determine whether they were abuse survivors or not, they were instructed to skip the next five questions if their responses to the previous seven questions were all "Never." It is possible that students who did not respond "Never" to the seven questions skipped the follow-up questions anyway in a desire to finish the survey quickly. The second possibility to the lack of response is the environment in which students took the survey. Students were asked to sign up for a designated one-hour time slot to participate in the study. It is highly likely that students signed up for the same time slots as their friends in class and subsequently sat next to each other while filling out the survey. Due to the close proximity and the sensitive nature of the questions, some sexual abuse survivors may not have wanted to fill out additional questions in fear that their friends might see. Better procedures in the future would be to have all students fill out all questions, whether they are abuse survivors or not, and/or to allow them to have more privacy while taking the survey.....

The author identifies gaps (missing data) in her survey data and suggests plausible explanations for them. At the end of the paragraph, she offers suggestions for how future research might minimize such gaps.

A final limitation of the study is the use of self-report data only. Self-report data may be fraught with problems derived from memory restrictions and perception differences. A more comprehensive design would include actual physical ways to measure the outcome variables. For example, the researcher could have strengthened the design by taking blood or urine samples to examine drug use. The problem here is that [the latter] method requires a great deal of time and money to undertake.

By perception differences, *the author is presumably referring to how different participants may have interpreted their prior experiences and/or items on the survey. An additional weakness of self-report data is that some participants may have intentionally misrepresented their prior experiences and/or current behaviors.*

[The researcher concludes the discussion by talking about potential implications of her findings for clinical practice and social policy.]

NOTE: Excerpt is from *Childhood Sexual Abuse and Family Functioning Linked with Eating and Substance Misuse: Mediated Structural Models* (pp. 92–94, 114–119) by K. J. Mitchell, 1998, unpublished doctoral dissertation, University of Rhode Island, Kingston. Reprinted with permission.

PRACTICAL APPLICATION Analyzing Data in a Quantitative Study

You can gain a clearer understanding of statistics and statistical procedures by reading about them in research reports and using them in actual practice. If your research project involves quantitative data, the following checklist can help you clarify which statistical analyses might be most appropriate for your situation.

✔ CHECKLIST

Questions to Consider When Choosing a Statistical Procedure

CHARACTERISTICS OF THE DATA

_____ 1. Are the data _____ continuous or _____ discrete?

_____ 2. What scale do the data reflect? Are they _____ nominal, _____ ordinal, _____ interval, or _____ ratio?

_____ 3. What do you want to do with the data?

_____ Calculate central tendency? If so, with which measure? _____________

_____ Calculate variability? If so, with which measure? _____________

_____ Calculate correlation? If so, with which measure? _____________

_____ Estimate population parameters? If so, which ones? _____________

____

_____ Test a null hypothesis? If so, at what confidence level? _____________

_____ Other? (specify) _____________________________

_____ 4. State your rationale for processing the data as you have just indicated you intend to do.

____

____

____

INTERPRETATION OF THE DATA

_____ 5. After you have treated the data statistically to analyze their characteristics, what will you then have?

____

____

_____ 6. From a research standpoint, what will your interpretation of the data consist of? How will the statistical analyses help you solve any part of your research problem?

____

____

_____ 7. What remains to be done before your problem (or any one of its subproblems) can be resolved?

____

____

_____ 8. What is your plan for carrying out this further interpretation of the data?

____

____

____

____

For Further Reading

Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences* (4th ed.). Upper Saddle River, NJ: Pearson.

Arthur, W., Jr., Bennett, W., Jr., & Huffcutt, A. I. (2001). *Conducting meta-analysis using SAS.* Mahwah, NJ: Erlbaum.

Azen, R., & Walker, C. M. (2011). *Categorical data analysis for the behavioral and social sciences.* New York: Routledge.

Bechhofer, R. E., Santner, T. J., & Goldsman, D. M. (1995). *Design and analysis of experiments for statistical selection, screening, and multiple comparisons.* New York: Wiley.

Bruning, J. L., & Kintz, B. L. (1997). *Computational handbook of statistics* (4th ed.). Boston: Allyn & Bacon.

Coladarci, T., Cobb, C. D., Minium, E. W., & Clarke, R. C. (2011). *Fundamentals of statistical reasoning in education with CD* (3rd ed.). New York: Wiley.

Coolidge, F. L. (2006). *Statistics: A gentle introduction* (2nd ed.). Thousand Oaks, CA: Sage.

Cooper, H. (2009). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.

Cramer, D. (1998). *Fundamental statistics for social research: Step-by-step calculations and computer techniques Using SPSS for Windows.* New York: Routledge.

Cummings, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170–180.

Fennessey, J. (1968). The general linear model: A new perspective on some familiar topics. *American Journal of Sociology, 74,* 1–27.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage.

Field, A., & Miles, J. (2010). *Discovering statistics using SAS.* Thousand Oaks, CA: Sage.

Fowler, F. J., Jr. (2008). *Survey research methods* (4th ed.). Thousand Oaks, CA: Sage. [See Chapter 10.]

Gonzalez, R. (2009). *Data analysis for experimental design.* New York: Guilford.

Gravetter, F. J., & Wallnau, L. B. (2011). *Essentials of statistics for the behavioral sciences* (7th ed.). Belmont, CA: Wadsworth/Cengage.

Haskins, L., & Jeffrey, K. (1990). *Understanding quantitative history.* New York: McGraw-Hill.

Heiman, G. W. (2006). *Basic statistics for the behavioral sciences* (5th ed.). Boston: Houghton Mifflin.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings.* Thousand Oaks, CA: Sage.

Kirk, R. E. (2008). *Statistics: An introduction* (5th ed.). Belmont, CA: Wadsworth/Cengage.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford.

Kranzler, J. H. (2011). *Statistics for the terrified* (5th ed.). Upper Saddle River, NJ: Pearson.

Lind, D. A., Marchal, W. G., & Wathen, S. A. (2010). *Basic statistics for business and economics* (14th ed.). New York: McGraw-Hill.

Phillemer, D. B. (1994). One- versus two-tailed hypothesis tests in contemporary educational research. *Educational Researcher, 20*(9), 13–17.

Phillips, J. L., Jr. (2000). *How to think about statistics* (6th ed.). New York: Henry Holt.

Rosner, B. (2011). *Fundamentals of biostatistics* (7th ed.). Monterey, CA: Brooks/Cole/Cengage.

Rowntree, D. (2004). *Statistics without tears: A primer for nonmathematicians.* Boston: Allyn & Bacon.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford, England: Oxford University Press.

Spector, P. E. (2001). *SAS programming for researchers and social scientists* (2nd ed.). Thousand Oaks, CA: Sage.

Sweet, S. A., & Grace-Martin, K. (2012). *Data analysis with SPSS: A first course in applied statistics* (4th ed.). Upper Saddle River, NJ: Pearson.

Thompson, B. (2008). *Foundations of behavioral statistics.* New York: Guilford.

Vogt, W. P., & Johnson, R. B. (2011). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences* (4th ed.). Thousand Oaks, CA: Sage.

Wilkinson, L., & the Task Force on Statistic Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wood, P. (2000). Meta-analysis. In G. M. Breakwell, S. Hammond, & C. Fife-Schaw (Eds.), *Research methods in psychology* (2nd ed., pp. 414–425). Thousand Oaks, CA: Sage.

Wright, D. B. (2006). The art of statistics: A survey of modern techniques. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 879–901). Mahwah, NJ: Erlbaum.