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.

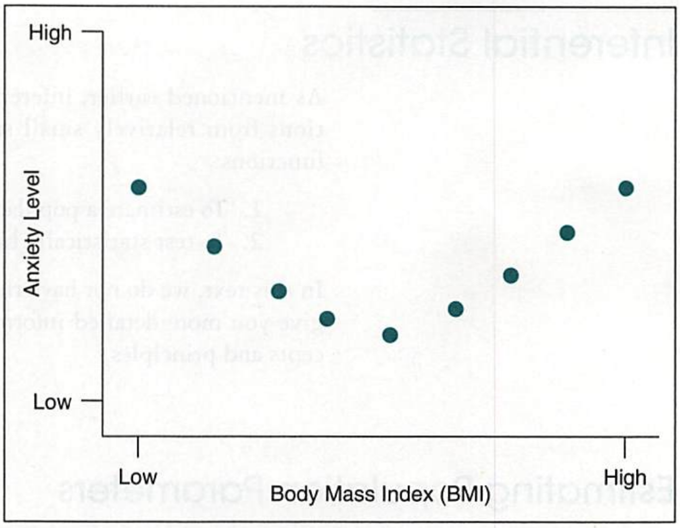
One especially noteworthy statistic in Table 11.4 is the *coefficient of determination,* or R^2. This statistic, which is the square of the Pearson *r*, tells us *how much of the variance is accounted for* by the correlation. Although you will see this expression used frequently in research reports, 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

Statistic	Symbol	Data for Which It's 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	$r_{12 \cdot 3}$	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	$R_{1 \cdot 23}$	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 11.11
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.

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?

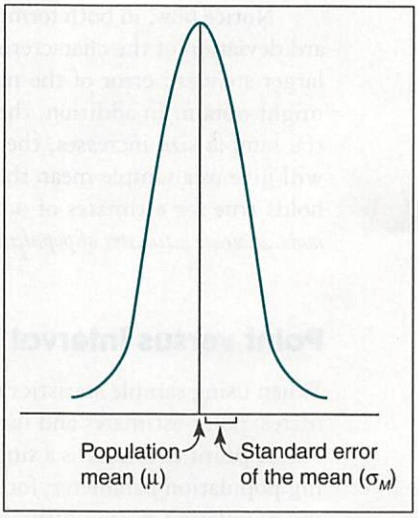
The problem is to determine population parameters on the basis of the sample statistics. From the sample, Jan can estimate the mean and variability of the pin diameters and the proportion of acceptable pins within the population universe. These are the values represented by μ, the σ, and the P.

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.

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

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

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.*

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.*

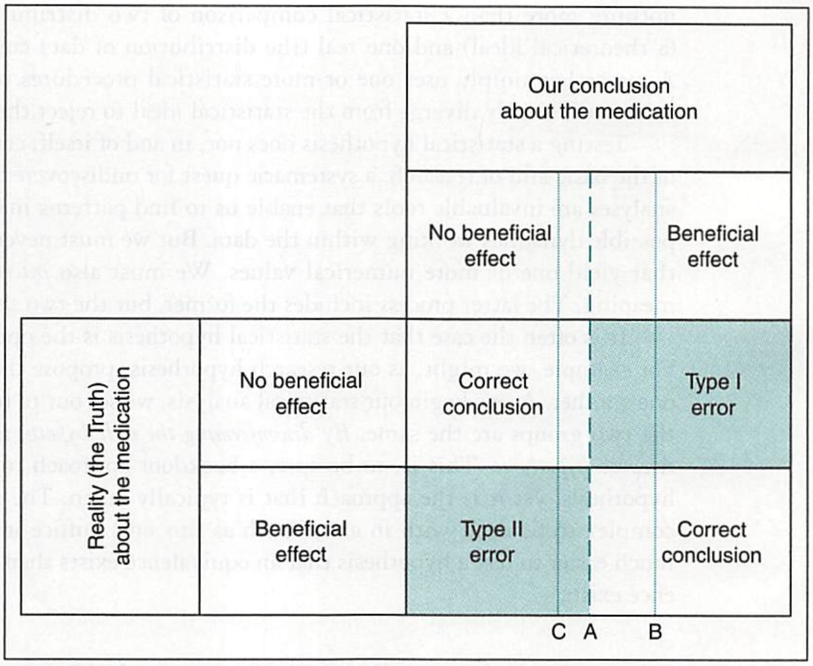
These four possibilities are illustrated in Figure 11.13. The three vertical lines illustrate three hypothetical significance levels we might choose. Imagine that the dashed middle line, Line A, represents a significance level of, say, .05. In this particular situation (such will not always be the case), 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

Statistical Procedure	*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.

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. . . .

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.

[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.=

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.

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.

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.

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

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 non-mathematicians.* 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.

Now go to MyEducationLab at www.myeducationlab.com to take assessments to evaluate your mastery of chapter concepts. Assignments and Activities exercises and Building Skills in Reading Research exercises are also available to help you master concepts and skills. Feedback for many of these exercises is provided so that you can see why your answers are correct or incorrect.