MODULE 9

Conducting Correlational Research

LEARNING OBJECTIVES

- Describe the difference between strong, moderate, and weak correlation coefficients.
- Draw and interpret scatterplots.
- Explain negative, positive, curvilinear, and no relationship between variables.
- Explain how assuming causality and directionality, the third-variable problem, restrictive ranges, and curvilinear relationships can be problematic when interpreting correlation coefficients.
- Explain how correlations allow us to make predictions.

When conducting correlational studies, researchers determine whether two naturally occurring variables (for example, height and weight or smoking and cancer) are related to each other. Such studies assess whether the variables are "co-related" in some way: Do tall people tend to weigh more than people of average height, or do those who smoke tend to have a higher-than-normal incidence of cancer? As we saw in Module 2, the correlational method is a type of nonexperimental method that describes the relationship between two measured variables. In addition to describing a relationship, correlations allow us to make predictions from one variable to another. If two variables are correlated, we can predict from one variable to the other with a certain degree of accuracy. Thus knowing that height and weight are correlated allows us to estimate, within a certain range, an individual's weight based on knowing the person's height.

Correlational studies are conducted for a variety of reasons. Sometimes it is impractical or ethically impossible to do an experimental study. For instance, it would be ethically impossible to manipulate smoking and assess whether it causes cancer in humans. How would you as a participant in an experiment like to be randomly assigned to the smoking condition and be told that you have to smoke a pack of cigarettes a day? Obviously this approach is not a viable experiment; however, one means of assessing the relationship between smoking and cancer is through correlational studies. In this type of study we can examine people who have already chosen to smoke and assess the degree of relationship between smoking and cancer.

Sometimes researchers choose to conduct correlational research because they are interested in measuring many variables and assessing the relationships between them. For example, they might measure various aspects of personality and assess the relationship between dimensions of personality.

MAGNITUDE, SCATTERPLOTS, AND TYPES OF RELATIONSHIPS

magnitude: An indication of the strength of the relationship between two variables.

Correlations vary in their **magnitude**, the strength of the relationship. Sometimes there is no relationship between variables, or the relationship may be weak; other relationships are moderate or strong. Correlations can also be represented graphically in a scatterplot or scattergram. In addition, relationships are of different types: positive, negative, none, or curvilinear.

Magnitude

The magnitude, or strength, of a relationship is determined by the correlation coefficient describing the relationship. As we saw in Module 6, a correlation coefficient is a measure of the degree of relationship between two variables; it can vary between −1.00 and +1.00. The stronger the relationship between the variables, the closer the coefficient is to either −1.00 or +1.00. The weaker the relationship between the variables, the closer the coefficient is to 0. You may recall from Module 6 that we typically discuss correlation coefficients as assessing a strong, moderate, or weak relationship, or no relationship at all. Table 9.1 provides general guidelines for assessing the magnitude of a relationship, but these ranges do not necessarily hold for all variables and all relationships.

A correlation coefficient of either −1.00 or +1.00 indicates a perfect correlation—the strongest relationship possible. For example, if height and weight were perfectly correlated (+1.00) in a group of 20 people, this coefficient would mean that the person with the highest weight was also the tallest person, the person with the second-highest weight was the second-tallest person, and so on down the line. In addition, in a perfect relationship each individual's score on one variable goes perfectly with his or her score on the other variable. For instance, this might mean that for every increase (decrease) in height of 1 inch, there is a corresponding increase (decrease) in weight of 10 pounds. If height and weight had a perfect negative correlation (−1.00), this coefficient would mean that the person with the highest weight was the shortest, the person with the second-highest weight was the second shortest, and so on, and that height and weight increased (decreased) by a set amount for each individual. It is very unlikely that you will ever observe a perfect correlation between two variables, but you may observe some very strong relationships between variables (±.70−.99). To sum up, whereas a correlation coefficient of ±1.00 represents a perfect relationship, a coefficient of 0 indicates no relationship between the variables.

Scatterplots

scatterplot: A figure that graphically represents the relationship between two variables.

A **scatterplot**, or scattergram, is a figure showing the relationship between two variables that graphically represents a correlation coefficient. Figure 9.1 presents a scatterplot of the height and weight relationship for 20 adults.

TABLE **9.1**

Estimates for Weak, Moderate, and Strong Correlation Coefficients

Correlation Coefficient	Strength of Relationship
±.70−1.00	Strong
±.30−.69	Moderate
±.00−.29	None (.00) to weak

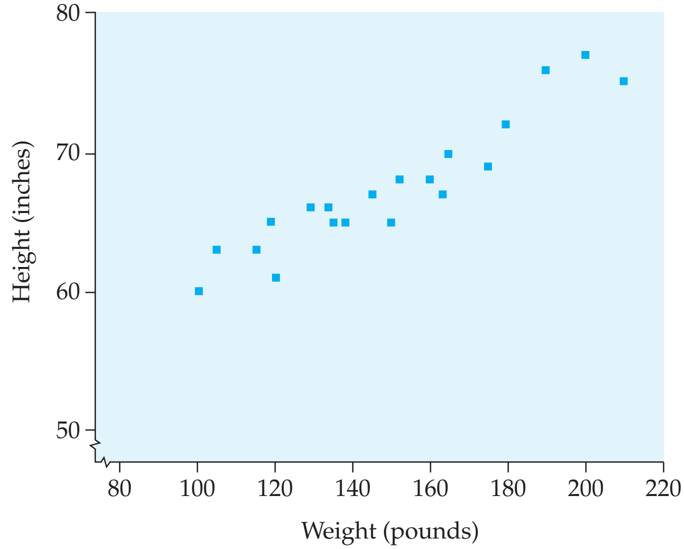

FIGURE **9.1** Scatterplot for height and weight

In a scatterplot two measurements are represented for each participant by the placement of a marker. In Figure 9.1 the horizontal *x*-axis shows the participant's weight, and the vertical *y*-axis shows height. The two variables could be reversed on the axes, and it would make no difference in the scatterplot. This scatterplot shows an upward trend, and the points cluster in a linear fashion. The stronger the correlation is, the more tightly the data points cluster around an imaginary line through their center. When there is a perfect correlation (± 1.00), the data points all fall on a straight line. In general, a scatterplot may show four basic patterns: a positive relationship, a negative relationship, no relationship, or a curvilinear relationship.

Positive Relationships

The relationship represented in Figure 9.2a shows a positive correlation, one in which there is a direct relationship between the two variables: An increase in one variable is related to an increase in the other, and a decrease in one is related to a decrease in the other. Notice that this scatterplot is similar to the one in Figure 9.1. The majority of the data points fall along an upward angle (from the lower left corner to the upper right corner). In this example a person who scored low on one variable also scored low on the other, an individual with a mediocre score on one variable had a mediocre score on the other, and anyone who scored high on one variable also scored high on the other. In other words, an increase (decrease) in one variable is accompanied by an increase (decrease) in the other; as variable *x* increases (or decreases), variable *y* does the same. If the data in Figure 9.2a represented height and weight measurements, we could say that those who are taller tend to weigh more, whereas

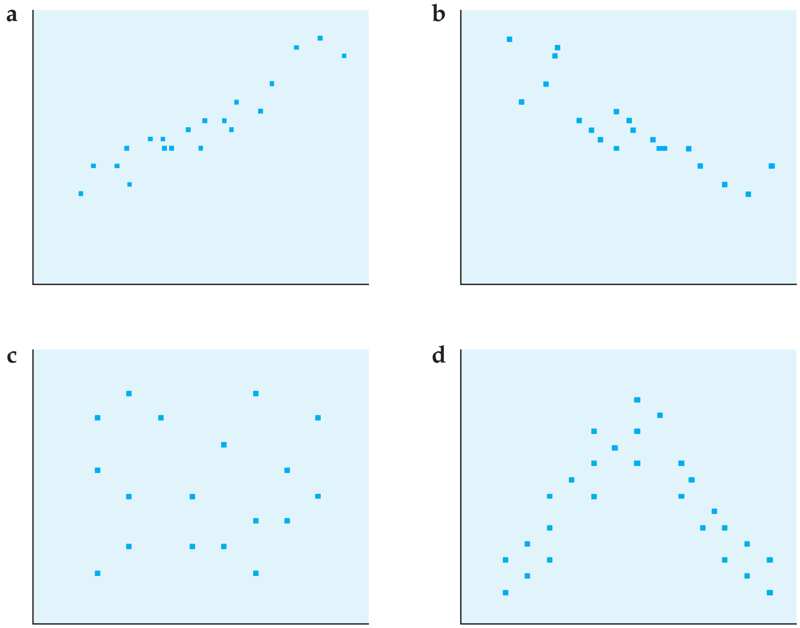

FIGURE **9.2** Possible types of Correlational relationships: (a) positive; (b) negative; (c) none; (d) curvilinear

those who are shorter tend to weigh less. Notice also that the relationship is linear: We could draw a straight line representing the relationship between the variables, and the data points would all fall fairly close to that line.

Negative Relationships

Figure 9.2b represents a negative relationship between two variables. Notice that in this scatterplot the data points extend from the upper left to the lower right. This negative correlation indicates that an increase in one variable is accompanied by a *decrease* in the other variable. This correlation represents an inverse relationship: The more of variable x that we have, the less we have of variable y. Assume that this scatterplot represents the relationship between age and eyesight. As age increases, the ability to see clearly tends to decrease—a negative relationship.

No Relationship

As shown in Figure 9.2c, it is also possible to observe no meaningful relationship between two variables. In this scatterplot the data points are scattered randomly. As you would expect, the correlation coefficient for these data is very close to 0 ($-.09$).

Curvilinear Relationships

A correlation coefficient of 0 indicates no meaningful relationship between two variables. However, it is also possible for a correlation coefficient of 0 to indicate a curvilinear relationship, as illustrated in Figure 9.2d. Imagine that this graph represents the relationship between psychological arousal (the *x*-axis) and performance (the *y*-axis). Individuals perform better when they are moderately aroused than when arousal is either very low or very high. The correlation coefficient for these data is also very close to 0 (−.05). Think about why this strong curvilinear relationship leads to a correlation coefficient close to 0. The strong positive relationship depicted in the left half of the graph essentially cancels out the strong negative relationship in the right half of the graph. Although the correlation coefficient is very low, we would not conclude that there is no relationship between the two variables. As the figure shows, the variables are very strongly related to each other in a curvilinear manner, with the points being tightly clustered in an inverted U shape.

Correlation coefficients only tell us about linear relationships. Thus even though there is a strong relationship between the two variables in Figure 9.2d, the correlation coefficient does not indicate this relationship because it is curvilinear. For this reason it is important to examine a scatterplot of the data in addition to calculating a correlation coefficient. Alternative statistics (beyond the scope of this text) can be used to assess the degree of curvilinear relationship between two variables.

IN REVIEW Relationships Between Variables

	Type of Relationships			
	Positive	**Negative**	**None**	**Curvilinear**
Description of Relationship	Variables increase and decrease together	As one variable increases, the other decreases in an inverse relationship	Variables are unrelated and do not move together in any way	Variables increase together up to a point and then as one continues to increase, the other decreases
Description of scatterplot	Data points are clustered in a linear pattern extending from lower left to upper right	Data points are clustered in a linear pattern extending from upper left to lower right	There is no pattern to the data points they are scattered all over the graph	Data points are clustered in a curved linear pattern forming a U shape or an inverted U shape
Example of variables related in this manner	Smoking and cancer	Mountain elevation and temperature	Intelligence and weight	Memory and age

CRITICAL THINKING CHECK 9.1

1. Which of the following correlation coefficients represents the weakest relationship between two variables?
   −.59 +.10 −1.00 +.76
2. Explain why a correlation coefficient of 0 or close to 0 may not mean that there is no relationship between the variables.
3. Draw a scatterplot representing a strong negative correlation between depression and self-esteem. Make sure you label the axes correctly.

MISINTERPRETING CORRELATIONS

Correlational data are frequently misinterpreted, especially when presented by newspaper reporters, talk show hosts, and television newscasters. Here we discuss some of the most common problems in interpreting correlations. Remember, a correlation simply indicates that there is a weak, moderate, or strong relationship (either positive or negative) or no relationship between two variables.

The Assumptions of Causality and Directionality

The most common error made when interpreting correlations is assuming that the relationship observed is causal in nature: that a change in variable A *causes* a change in variable B. Correlations simply identify relationships; they do not indicate causality. For example, a commercial recently appeared on television sponsored by an organization promoting literacy. The statement was made at the beginning of the commercial that a strong positive correlation had been observed between illiteracy and drug use in high school students (those high on the illiteracy variable also tended to be high on the drug use variable). The commercial concluded with a statement along the lines of "Let's stop drug use in high school students by making sure they can all read." Can you see the flaw in this conclusion? The commercial did not air for very long, probably because someone pointed out the error.

causality: The assumption that a correlation indicates a causal relationship between two variables.

directionality: The inference made with respect to the direction of a causal relationship between two variables.

This commercial made the twin errors of assuming causality and directionality. **Causality** refers to the assumption that the correlation between two variables indicates a causal relationship, and **directionality** refers to the inference made with respect to the direction of a causal relationship between two variables. The commercial assumed that illiteracy was causing drug use; it claimed that if illiteracy were lowered, then drug use would also be lowered. As we know, a correlation between two variables indicates only that they are related, that is, they vary together. Although it is possible that one variable causes changes in the other, we cannot draw this conclusion from correlational data.

Research on smoking and cancer illustrates this limitation of correlational data. For research with humans we have only correlational data indicating a positive correlation between smoking and cancer. Because the data are correlational, we cannot conclude that there is a causal relationship. In this situation it is probable that the relationship is causal. However, based solely on correlational data, we cannot draw that conclusion, nor can we assume the direction of the relationship. Thus the tobacco industry could argue that, yes, there is a correlation between smoking and cancer, but maybe cancer causes smoking, or maybe individuals predisposed to cancer are more attracted to smoking cigarettes. Even though experimental data based on research with laboratory animals indicate that smoking causes cancer, the tobacco industry questions whether the research is applicable to humans and for years continued to state that no research had produced evidence of a causal link between smoking and cancer in humans.

A classic example of the assumption of causality and directionality with correlational data occurred when researchers observed a strong negative correlation between eye movement patterns and reading ability in children. Poor readers tended to make more erratic eye movements than normal, more movements from right to left, and more stops per line of text. Based on this correlation, some researchers assumed causality and directionality: They presumed that poor oculomotor skills caused poor reading and proposed programs for "eye movement training." Many elementary school students who were poor readers spent time in such training, supposedly developing oculomotor skills in the hope that these skills would improve their reading ability. Experimental research later provided evidence that the relationship between eye movement patterns and reading ability is indeed causal, but that the direction of the relationship is the reverse: poor reading causes more erratic eye movements! Children who are having trouble reading need to go back over the information more and stop and think about it more. When children improve their reading skills (i.e., improve recognition and comprehension), their eye movements become smoother (Olson & Forsberg, 1993). Because of the errors of assuming causality and directionality, many children never received the appropriate training to improve their reading ability.

The Third-Variable Problem

When we interpret a correlation, it is important to remember that although the correlation between the variables may be very strong, the relationship may be the result of a third variable that influences both of the measured variables. The **third-variable problem** results when a correlation between two variables is dependent on another (third) variable.

third-variable problem: The problem of a correlation between two variables being dependent on another (third) variable.

A good example of the third-variable problem is a well-cited study conducted by social scientists and physicians in Taiwan (Li, 1975). The researchers attempted to identify the variables that best predicted the use of birth control; a question of interest to the researchers because of overpopulation problems in Taiwan. They collected data on various behavioral and environmental variables and found that the variable most strongly correlated with contraceptive use was the number of electrical appliances (yes, electrical appliances—stereos, toasters, televisions, and so on) in the home. If we take this correlation at face value, it means that individuals who use many electrical appliances tend also to use contraceptives, whereas those with fewer electrical appliances tend to use contraceptives less.

It should be obvious that this relationship is not causal (buying electrical appliances does not cause individuals to use birth control, nor does using birth control cause individuals to buy electrical appliances). Thus we probably do not have to worry about people assuming either causality or directionality when interpreting this correlation. The problem is a third variable. In other words, the relationship between electrical appliances and contraceptive use is not really a meaningful relationship; other variables are tying them together. Can you think of other ways in which individuals who use contraceptives and who have a large number of appliances might be similar? Education is a possible third variable. Individuals with a higher education level tend to be better informed about contraceptives and also tend to have a higher socioeconomic status (they get better paying jobs). Their higher socioeconomic status allows them to buy more "things," including electrical appliances.

partial correlation: A correlational technique that involves measuring three variables and then statistically removing the effect of the third variable from the correlation of the remaining two.

It is possible statistically to determine the effects of a third variable by using a correlational procedure known as **partial correlation**, which involves measuring all three variables and then statistically removing the effect of the third variable from the correlation of the remaining two. If the third variable (in this case, education) is responsible for the relationship between electrical appliances and contraceptive use, then the correlation should disappear when the effect of education is removed, or partialed out.

Restrictive Range

restrictive range: A variable that is truncated and has limited variability.

The idea behind measuring a correlation is that we assess the degree of relationship between two variables. Variables by definition must vary. When a variable is truncated, we say that it has a **restrictive range**, that is, the variable does not vary enough. Look at Figure 9.3a, which represents a scatterplot of SAT scores and college GPAs for a group of students. SAT scores and GPAs are positively correlated. Neither of these variables is restricted in range (for this group of students, SAT scores vary from 400 to 1600 and GPAs vary from 1.5 to 4.0), so we have the opportunity to observe a relationship between the variables. Now look at Figure 9.3b, which represents the correlation between the same two variables, except the range on the SAT variable is restricted to those who scored between 1000 and 1150. The SAT variable has been restricted, or truncated, and does not "vary" very much. As a result the opportunity to observe a correlation has been diminished. Even if there were a strong relationship between these variables, we could not observe it because of the restricted range of one of the variables. Thus when interpreting and using correlations, beware of variables with restricted ranges.

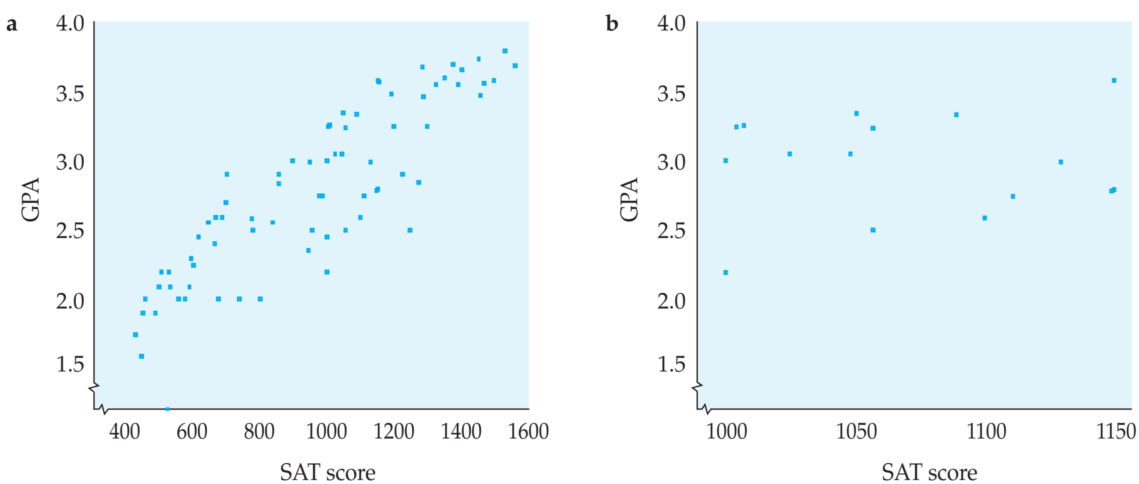

FIGURE **9.3** Restrictive range and correlation

Curvilinear Relationships

Curvilinear relationships and the caution in interpreting them were discussed earlier in the module. Because correlations are a measure of linear relationships, when a relationship is curvilinear, a correlation coefficient does not adequately indicate the degree of relationship between the variables. If necessary, look back over the previous section on curvilinear relationships in order to refresh your memory concerning them.

IN REVIEW Misinterpreting Correlations

	Types of Misinterpretations			
	Causality and Directionality	**Third Variable**	**Restrictive Range**	**Curvilinear Relationship**
Description of Misinterpretation	We assume that the correlation is causal and that one variable causes changes in the other.	Other variables are responsible for the observed correlation.	One or more of the variables is truncated or restricted, and the opportunity to observe a relationship is minimized.	The curved nature of the relationship decreases the observed correlation coefficient.
Examples	We assume that smoking causes cancer or that illiteracy causes drug abuse because a correlation has been observed.	We find a strong positive relationship between birth control and the number of electrical appliances.	If SAT scores are restricted (limited in range), the correlation between SAT and GPA appears to decrease.	As arousal increases, performance increases up to a point; as arousal continues to increase, performance decreases.

CRITICAL THINKING CHECK 9.2

1. "I have recently observed a strong negative correlation between depression and self-esteem." Explain what this statement means. Make sure you avoid the misinterpretations described in the text.
2. General State University officials recently investigated the relationship between SAT scores and GPAs (at graduation) for its senior class. They were surprised to find a weak correlation between these two variables. They know they have a grade inflation problem (the whole senior class graduated with GPAs of 3.0 or higher), but they are unsure how this might help account for the low correlation observed. Can you explain?

Prediction and Correlation

Correlation coefficients not only describe the relationship between variables, but they also allow us to make predictions from one variable to another. Correlations between variables indicate that when one variable is present at a certain level, the other also tends to be present at a certain level. Notice the wording. The statement is qualified by the phrase "tends to." We are not saying that a prediction is guaranteed or that the relationship is causal but simply that the variables seem to occur together at specific levels. Think about some of the examples used in this module. Height and weight are positively correlated. One is not causing the other; nor can we predict an individual's weight exactly based on height (or vice versa). But because the two variables are correlated, we can predict with a certain degree of accuracy what an individual's approximate weight might be if we know the person's height.

Let's take another example. We have noted a correlation between SAT scores and college freshman GPAs. Think about the purpose of the SAT. College admissions committees use the test as part of the admissions procedure because there is a positive correlation between SAT scores and college freshman GPAs. Individuals who score high on the SAT tend to have higher college freshman GPAs; those who score lower on the SAT tend to have lower college freshman GPAs. Therefore knowing students' SAT scores can help predict, with a certain degree of accuracy, their freshman GPAs and their potential for success in college. At this point some of you are probably saying, "But that isn't true for me. I scored poorly (or very well) on the SAT, and my GPA is great (or not so good)." Statistics tell us only the trend for most people in the population or sample. There are always outliers—the few individuals who do not fit the trend. Most people, however, are going to fit the pattern.

Think about another example. There is a strong positive correlation between smoking and cancer, but you may know someone who has smoked for 30 or 40 years and does not have cancer or any other health problems. Does this one individual negate the fact that there is a strong relationship between smoking and cancer? No. To claim that it does would be a classic **person-who argument**, that is, arguing that a well established statistical trend is invalid because we know a "person who" went against the trend (Stanovich, 2007). A counterexample does not change the existence of a strong statistical relationship between the variables nor that you are increasing your chance of getting cancer if you smoke. Because of the correlation between the variables, we can predict (with a fairly high degree of accuracy) who might get cancer based on knowing a person's smoking history.

person-who argument: Arguing that a well-established statistical trend is invalid because we know a "person who" went against the trend.

SUMMARY

After reading this module, you should have an understanding of the correlational research method, which allows researchers to observe relationships between variables, and of correlation coefficients, the statistics that assess the relationship. Correlations vary in type (positive, negative, none, or curvilinear) and magnitude (weak, moderate, or strong). The pictorial representation of a correlation is a scatterplot. A scatterplot allows us to see the relationship, facilitating its interpretation.

Several errors are commonly made when interpreting correlations, including assuming causality and directionality, overlooking a third variable, having