Correlation and Prediction

Chapter Outline

We now look at some descriptive statistics for the relationship between two or more variables. To give you an idea of what we mean, let's consider some common real-world examples. Among students, there is a relationship between high school grades and college grades. It isn't a perfect relationship, but generally speaking, students with better high school grades tend to get better grades in college. Similarly, there is a relationship between parents' heights and the adult height of their children. Taller parents tend to give birth to children who grow up to be taller

TIP FOR SUCCESS

Before beginning this chapter, be sure you have mastered the concepts of mean, standard deviation, and *Z* scores.

From Chapter 3 of *Statistics for the Behavioral & Social Sciences: A Brief Course*, Fifth Edition. Arthur Aron, Elliot J. Coups, Elaine N. Aron. Copyright © 2011 by Pearson Education, Inc. Published by Pearson Prentice Hall. All rights reserved.

than the children of shorter parents. Again, the relationship isn't perfect, but the general pattern is clear. Now we'll look at an example in detail.

One hundred thirteen married people in the small college town of Santa Cruz, California, responded to a questionnaire in the local newspaper about their marriage. (This was part of a larger study reported by Aron, Norman, Aron, McKenna, & Heyman, 2000.) As part of the questionnaire, they answered the question, "How exciting are the things you do together with your partner?" using a scale from 1 = *not exciting at all* to 5 = *extremely exciting*. The questionnaire also included a standard measure of marital satisfaction (that included items such as "In general, how often do you think that things between you and your partner are going well?").

The researchers were interested in finding out the relationship between doing exciting things with a marital partner and the level of marital satisfaction people reported. In other words, they wanted to look at the relationship between two groups of scores: the group of scores for doing exciting things and the group of scores for marital satisfaction. As shown in Figure 1, the relationship between these two groups of scores can be shown very clearly using a graph. The horizontal axis is for people's answers to the question, "How exciting are the things you do together with your partner?" The vertical axis is for the marital satisfaction scores. Each person's score on the two variables is shown as a dot.

The overall pattern is that the dots go from the lower left to the upper right. That is, lower scores on the variable "exciting activities with partner" more often go with lower scores on the variable "marital satisfaction," and higher with higher. So,

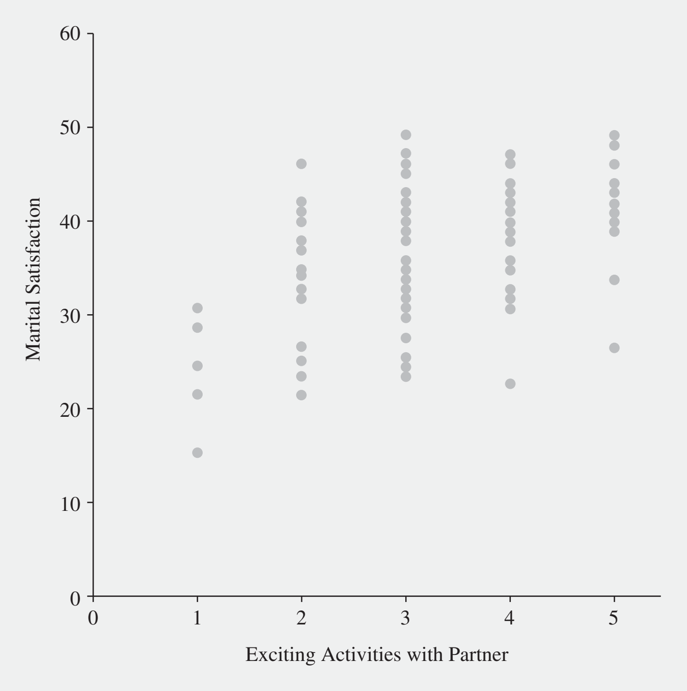


Figure 1 Scatter diagram showing the correlation for 113 married individuals between doing exciting activities with their partner and their marital satisfaction. (Data from Aron et al., 2000.)

in general, this graph shows that the more that people did exciting activities with their partners, the more satisfied they were in their marriages. Even though the pattern is far from one to one, you can see a general trend. This general pattern is of high scores on one variable going with high scores on the other variable, low scores going with low scores, and moderate with moderate. This is an example of a **correlation.**

A correlation describes the relationship between two variables. More precisely, the usual measure of a correlation describes the relationship between *two equal-interval numeric variables.* The difference between values for an equal-interval numeric variable corresponds to differences in the underlying thing being measured. (Most behavioral and social scientists consider scales like a 1 to 10 rating scale as approximately equal-interval scales.) There are countless examples of correlations: In children, there is a correlation between age and coordination skills; among students, there is a correlation between amount of time studying and amount learned; in the marketplace, we often assume that a correlation exists between price and quality—that high prices go with high quality and low with low.

This chapter explores correlation, including how to describe it graphically, different types of correlations, how to figure the *correlation coefficient* (which gives a number for the degree of correlation), issues about how to interpret a correlation coefficient, and how you can use correlation to predict the score on one variable from knowledge of a person's score on another correlated variable (such as predicting college grades from high school grades).

Graphing Correlations: The Scatter Diagram

Figure 1 shows the correlation between exciting activities and marital satisfaction. This kind of diagram is an example of a **scatter diagram** (also called a *scatterplot* or *scattergram*). A scatter diagram shows the pattern of the relationship between two variables at a glance.

How to Make a Scatter Diagram

There are three steps to making a scatter diagram.

❶ **Draw the axes and decide which variable goes on which axis.** Often, it doesn't matter which variable goes on which axis. However, sometimes the researchers are thinking of one of the variables as predicting or causing the other. In that case, the variable that is doing the predicting or causing goes on the horizontal axis and the variable that is being predicted about or caused goes on the vertical axis. In Figure 1, we put exciting activities on the horizontal axis and marital satisfaction on the vertical axis. This was because the study was based on a theory that the more exciting activities that a couple does together, the more the couple is satisfied with their marriage. (We will have more to say about prediction later in the chapter.)

❷ **Determine the range of values to use for each variable and mark them on the axes.** Your numbers should go from low to high on each axis, starting from where the axes meet. Usually, your low value on each axis is 0. However, you can use a higher value to start each axis if the lowest value your measure can possibly have in the group you are studying is a lot higher than 0. For example, if a variable is age and you are studying college students, you might start that axis with 16 or 17, rather than 0.

correlation Association between scores on two variables.

scatter diagram Graph showing the relationship between two variables: the values of one variable are along the horizontal axis and the values of the other variable are along the vertical axis; each score is shown as a dot in this two-dimensional space.

Each axis should continue to the highest value your measure can possibly have. When there is no obvious highest possible value, make the axis go to a value that is as high as people ordinarily score in the group of people of interest for your study.

In Figure 1, the horizontal axis is for the question about exciting activities, which was answered on a scale of 1 to 5. We start the axis at 0, because this is standard, even though the lowest possible score on the scale is 1. We went up to 5, because that is the highest possible value on the scale. Similarly, the vertical axis goes from 0 to 60, since the highest possible score on marital satisfaction was 60. (There were 10 marital satisfaction items, each answered on a 1 to 6 scale.) Note also that scatter diagrams are usually made roughly square, with the horizontal and vertical axes being about the same length (a 1:1 ratio).

❸ **Mark a dot for each pair of scores.** Find the place on the horizontal axis for the first pair of scores on the horizontal-axis variable. Next, move up to the height for the score for the first pair of scores on the vertical-axis variable. Then mark a clear dot. Continue this process for the remaining people. Sometimes the same pair of scores occurs twice (or more times). This means that the dots for these people would go in the same place. When this happens, you can put a second dot as near as possible to the first—touching, if possible—but making it clear that there are in fact two dots in the one place. Alternatively, you can put the number 2 in that place.

An Example. Suppose a researcher is studying the relationship of sleep to mood. As an initial test, the researcher asks six students in her morning seminar two questions:

1. How many hours did you sleep last night?
2. How happy do you feel right now on a scale from 0 = *not at all happy* to 8 = *extremely happy*?

The (fictional) results are shown in Table 1. (In practice, a much larger group would be used in this kind of research. We are using an example with just six students to keep things simple for learning. In fact, we have done a real version of this study. Results of the real study are similar to what we show here, except not as strong as the ones we made up to make the pattern clear for learning.)

Table 1 Hours Slept Last Night and Happy Mood Example (Fictional Data)

Hours Slept	Happy Mood
5	2
7	4
8	7
6	2
6	3
10	6

❶ **Draw the axes and decide which variable goes on which axis.** Because sleep comes before mood in this study, it makes most sense to think of sleep as the predictor. (However, it is certainly possible that people who are in general in a good mood are able to get more sleep.) Thus, as shown in Figure 2a, we put hours slept on the horizontal axis and happy mood on the vertical axis.

❷ **Determine the range of values to use for each variable and mark them on the axes.** For the horizontal axis, we start at 0 as usual. We do not know the maximum possible, but let us assume that students rarely sleep more than 12 hours. The vertical axis goes from 0 to 8, the lowest and highest scores possible on the happiness question (see Figure 2b.)

❸ **Mark a dot for each pair of scores.** For the first student, the number of hours slept last night was 5. Move across to 5 on the horizontal axis. The happy mood rating for the first student was 2, so move up to the point across from the 2 on the vertical axis. Place a dot at this point, as shown in Figure 2c. Do the same for each of the other five students. The result should look like Figure 2d.

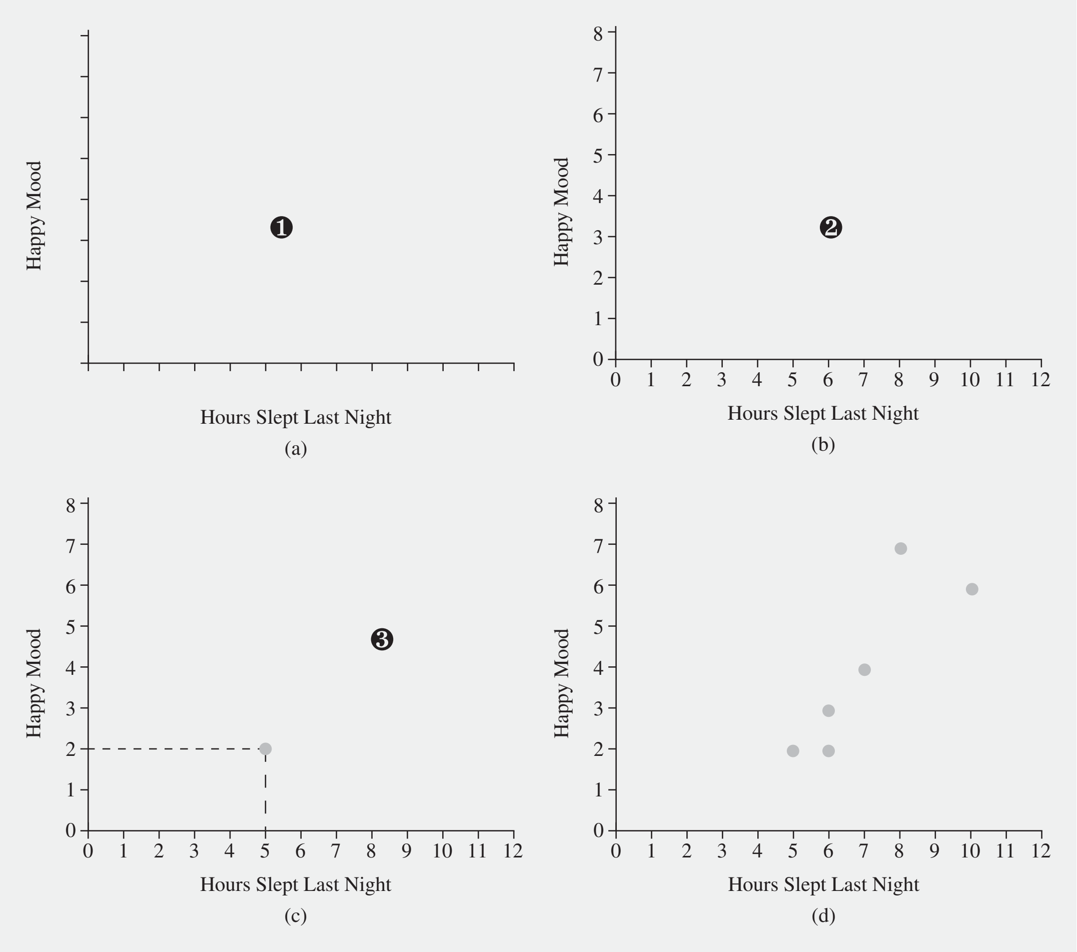


Figure 2 Steps for making a scatter diagram. (a) ❶ Draw the axes and decide which variable goes on which axis—the predictor variable (Hours Slept Last Night) on the horizontal axis, the other (Happy Mood) on the vertical axis. (b) ❷ Determine the range of values to use for each variable and mark them on the axes. (c) ❸ Mark a dot for the pair of scores for the first student. (d) ❸ continued: Mark dots for the remaining pairs of scores.

How are you doing?

1. What does a scatter diagram show, and what does it consist of?
2. (a) When it is the kind of study in which one variable can be thought of as predicting another variable, which variable goes on the horizontal axis? (b) Which variable goes on the vertical axis?
3. Make a scatter diagram for the following scores for four people who were each tested on two variables, *X* and *Y*. *X* is the variable we are predicting from; it can have scores ranging from 0 to 6. *Y* is the variable being predicted; it can have scores from 0 to 7.

Person	*X*	*Y*
A	3	4
B	6	7
C	1	2
D	4	6

Answers

1. A scatter diagram is a graph that shows the relation between two variables. One axis is for one variable; the other axis, for the other variable. The graph has a dot for each pair of scores. The dot for each pair is placed above the score for that pair on the horizontal axis variable and directly across from the score for that pair on the vertical axis variable.
2. (a) The variable that is doing the predicting goes on the horizontal axis. (b) The variable that is being predicted goes on the vertical axis.
3. See Figure 3.

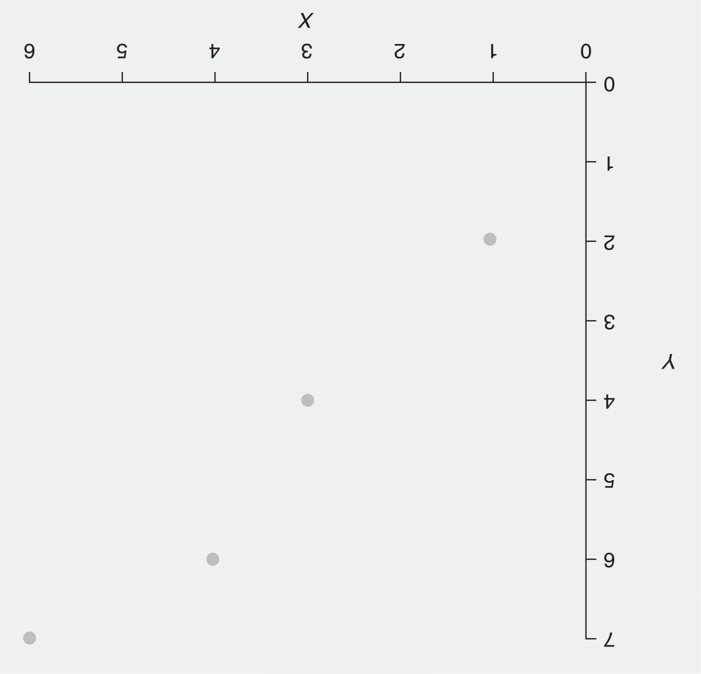


Figure 3 Scatter diagram for scores in "How Are You Doing?" question 3.

Patterns of Correlation

Linear and Curvilinear Correlations

In each example so far, the pattern in the scatter diagram very roughly approximates a straight line. Thus, each is an example of a **linear correlation**. In the scatter diagram for the study of exciting activities and marital satisfaction (Figure 1), you could draw a line showing the general trend of the dots, as we have done in Figure 4. Similarly, you could draw such a line in the happy mood and sleep study example,

linear correlation Relationship between two variables that shows up on a scatter diagram as the dots roughly following a straight line.

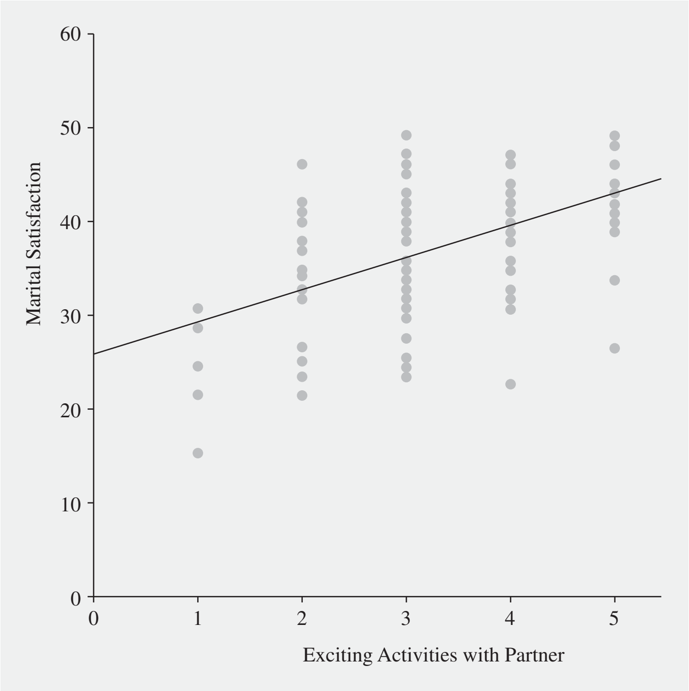


Figure 4 The scatter diagram of Figure 1 with a line drawn in to show the general trend. (Data from Aron et al., 2000.)

as shown in Figure 5. Notice that the scores do not all fall right on the line, far from it in fact. Notice, however, that the line does describe the general tendency of the scores.

Sometimes, however, the general relationship between two variables does not follow a straight line at all, but instead follows the more complex pattern of a

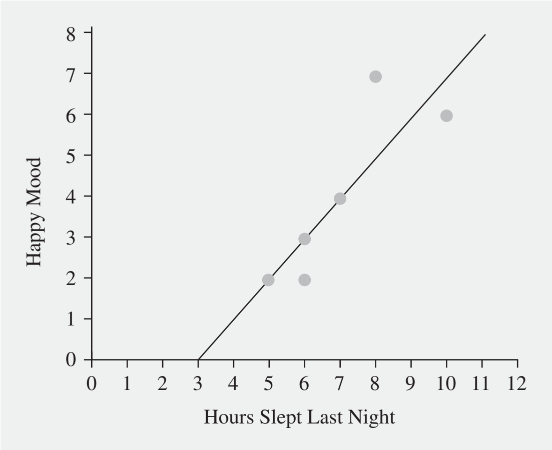


Figure 5 Scatter diagram from Figure 2d with a line drawn to show the general trend.

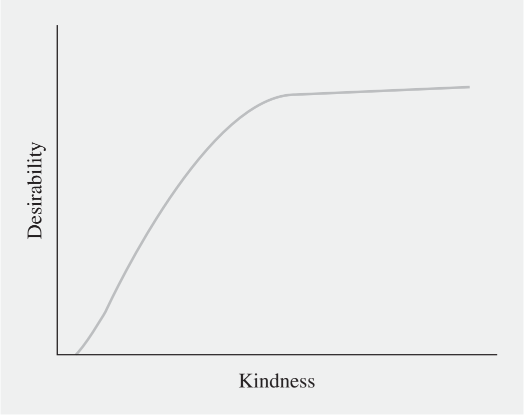


Figure 6 Example of a curvilinear relationship: Desirability and kindness.

curvilinear correlation. Consider, for example, the relationship between a person's level of kindness and the degree to which that person is desired by others as a potential romantic partner. There is evidence suggesting that, up to a point, a greater level of kindness increases a person's desirability as a romantic partner. However, beyond that point, additional kindness does little to increase desirability (Li et al., 2002). This particular curvilinear pattern is shown in Figure 6. Notice that you could not draw a straight line to describe this pattern. There are many different curvilinear patterns. For example, the pattern can look like the letter U, the letter V, the letter C, or in fact any systematic pattern that is not a straight line.

The usual way of figuring the correlation (the one you learn shortly in this chapter) gives the degree of *linear* correlation. If the true pattern of association is curvilinear, figuring the correlation in the usual way could show little or no correlation. Thus, it is important to look at scatter diagrams to identify these richer relationships rather than automatically figuring correlations in the usual way, assuming that the only possible relationship is a straight line.

curvilinear correlation Relationship between two variables that shows up on a scatter diagram as dots following a systematic pattern that is not a straight line; any association between two variables other than a linear correlation.

no correlation No systematic relationship between two variables.

positive correlation Relationship between two variables in which high scores on one go with high scores on the other, mediums with mediums, and lows with lows; on a scatter diagram, the dots roughly follow a straight line sloping up and to the right.

negative correlation Relationship between two variables in which high scores on one go with low scores on the other, mediums with mediums, and lows with highs; on a scatter diagram, the dots roughly follow a straight line sloping down and to the right.

No Correlation

It is also possible for two variables to be essentially unrelated to each other. For example, if you were to do a study of creativity and shoe size, your results might appear as shown in Figure 7. The dots are spread everywhere, and there is no line, straight or otherwise, that is any reasonable representation of a trend. There is simply **no correlation.**

Positive and Negative Linear Correlations

In the examples so far of linear correlations, such as exciting activities and marital satisfaction, high scores generally go with high scores, lows with lows, and mediums with mediums. This situation is called a **positive correlation.** (One reason for the term "positive" is that in geometry, the slope of a line is positive when it goes up and to the right on a graph like this. Notice that in Figures 4 and 5 the positive correlation is shown by a line that goes up and to the right.)

Sometimes, however, high scores tend to go with low scores and low scores with high scores. This is called a **negative correlation.** For example, in the newspaper survey about marriage, the researchers also asked about boredom with the relationship.

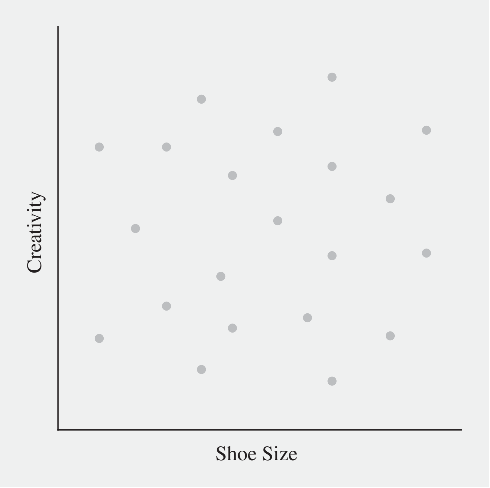


Figure 7 Two variables with no association with each other: creativity and shoe size (fictional data).

Not surprisingly, the more bored a person was, the *lower* was the person's marital satisfaction. Similarly, the less bored a person was, the higher the marital satisfaction. That is, high scores on one variable went with low scores on the other. This is shown in Figure 8, where we also put in a line to emphasize the general trend. You can see that as it goes from left to right, it slopes downward. (Compare this to the result for the relation of exciting activities and marital satisfaction shown in Figure 4, which slopes upward.)

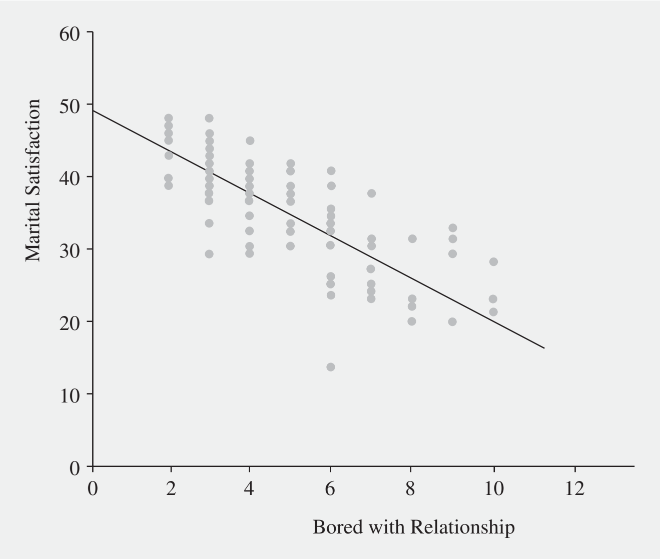


Figure 8 Scatter diagram with the line drawn in to show the general trend for a negative correlation between two variables: greater boredom with the relationship goes with lower marital satisfaction. (Data from Aron et al., 2000.)

Another study (Mirvis & Lawler, 1977) also illustrates a negative correlation. That study found that absenteeism from work had a negative linear correlation with job satisfaction. That is, the higher the level of job satisfaction, the lower the level of absenteeism. Put another way, the *lower* the level of job satisfaction is, the *higher* the absenteeism becomes. Research on this topic has continued to show this pattern all over the world (e.g., Punnett et al., 2007) and the same pattern is found for university classes: the *more* satisfied students are, the *less* they miss class (Yorges, Bloom, & Defonzo, 2007).

Strength of the Correlation. What we mean by the *strength of the correlation* is how much there is a clear pattern of some particular relationship between two variables. For example, we saw that a positive linear correlation is when high scores go with highs, mediums with mediums, lows with lows. The strength of such a correlation, then, is *how much* highs go with highs, and so on. Similarly, the strength of a negative linear correlation is how much the highs on one variable go with the lows on the other, and so forth. To put this another way, the strength of a linear correlation can be seen in a scatter diagram by how close the dots fall to a simple straight line.

Importance of Identifying the Pattern of Correlation

The procedure you learn in the next main section is for figuring the direction and strength of linear correlation. As we suggested earlier, the best approach to such a problem is *first* to make a scatter diagram and use it to identify the pattern of correlation. If the pattern is curvilinear, then you would not go on to figure the linear correlation. This is important because figuring the linear correlation when the true correlation is curvilinear would be misleading. (For example, you might conclude that there is little or no correlation when in fact there is a quite strong relationship; it is just not linear.) You should assume that the correlation is linear, unless the scatter diagram shows a curvilinear correlation. We say this because when the linear correlation is small, the dots will fall far from a straight line. In such situations, it can sometimes be hard to imagine a straight line that roughly shows the pattern of dots.

If the correlation appears to be linear, it is also important to "eyeball" the scatter diagram a bit more. The idea is to note the direction (positive or negative) of the linear correlation and also to make a rough guess as to the strength of correlation. There is a "small" (or "weak") correlation when you can barely tell there is a correlation at all—the dots fall far from a straight line. There is a "large" (or "strong") correlation if the dots fall very close to a straight line. The correlation is "moderate" if the pattern of dots is somewhere between a small and a large correlation. Some examples of scatter diagrams with varying directions and strengths of correlation are shown in Figure 9. You can see that Figure 9a is a very strong positive correlation (the dots go up and to the right and all fall close to what would be a simple straight line), 9b is a very strong negative correlation, 9c seems more of a moderate positive correlation, 9d seems more of a small to moderate negative correlation, and 9e and 9f appear either very small or no correlation. Using a scatter diagram to examine the direction and approximate strength of correlation is important because it lets you check to see whether you have made a major mistake when you then do the figuring you learn in the next section.

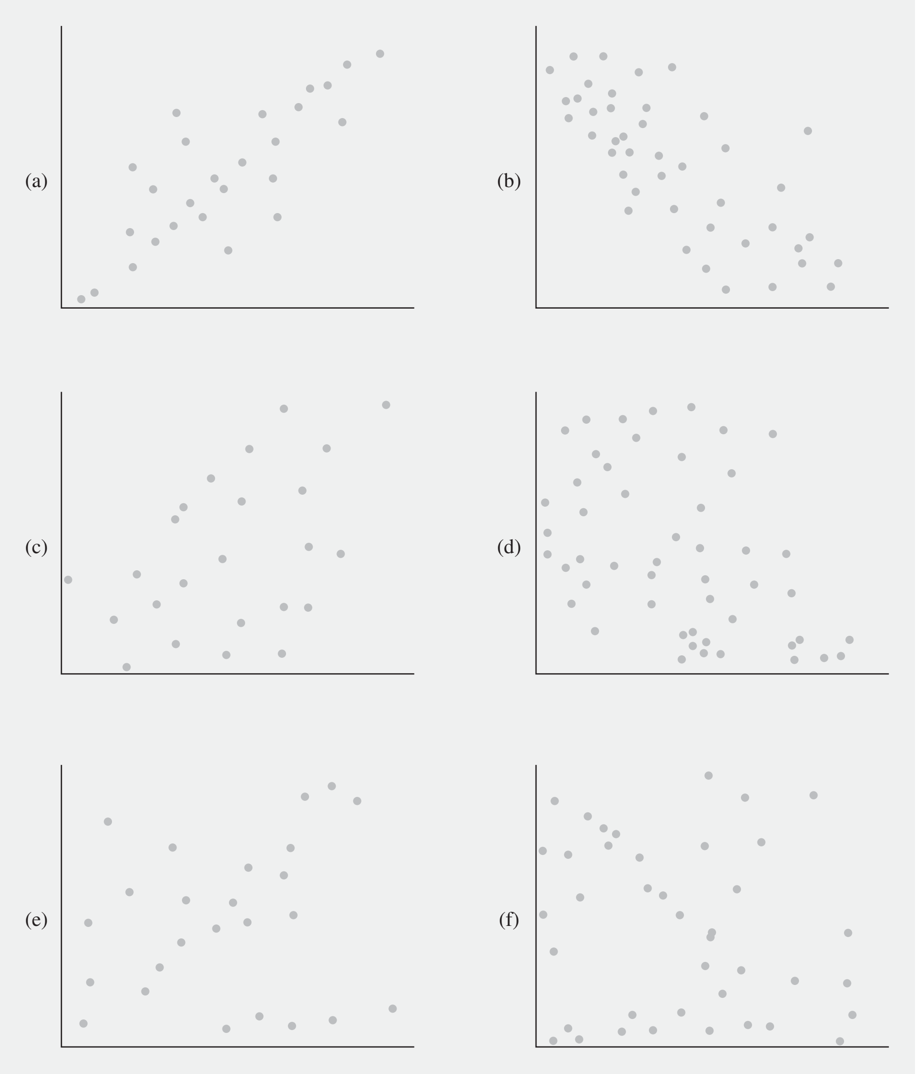


Figure 9 Examples of scatter diagrams with different degrees of correlation.

How are you doing?

1. What is the difference between a linear and curvilinear correlation in terms of how they appear in a scatter diagram?
2. What does it mean to say that two variables have no correlation?
3. What is the difference between a positive and negative linear correlation? Answer this question in terms of (a) the patterns in a scatter diagram and (b) what those patterns tell you about the relationship between the two variables.
4. For each of the scatter diagrams shown in Figure 10, say whether the pattern is roughly linear, curvilinear, or no correlation. If the pattern is roughly linear, also say if it is positive or negative, and whether it is large, moderate, or small.

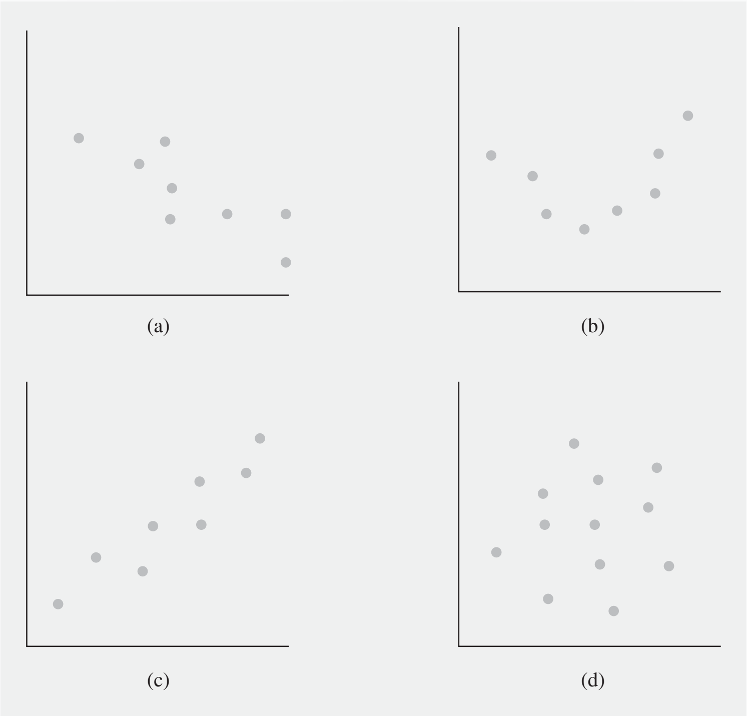


Figure 10 Scatter diagrams for "How Are You Doing?" question 4.

5. Give two reasons why it is important to identify the pattern of correlation in a scatter diagram before going on to figure the linear correlation.

Answers

1. In a linear correlation, the pattern of dots roughly follow a straight line (although with a small correlation, the dots will be spread widely around a straight line); in a curvilinear correlation, there is a clear systematic pattern to the dots, but it is not a straight line.
2. The two variables have no correlation when there is no pattern of relationship between them.
3. (a) In a scatter diagram for a positive linear correlation, the line that roughly describes the pattern of dots goes up and to the right; in a negative linear correlation, the line goes down and to the right. (b) In a positive linear correlation, the basic pattern is that high scores on one variable go with high scores on the other, mediums go with mediums, and lows go with lows; in a negative linear correlation, high scores on one variable generally go with low scores on the other, mediums go with mediums, and lows go with highs.
4. Figure 10: (a) linear, negative, large; (b) curvilinear; (c) linear, positive, large; (d) no correlation (it may at first look one way or the other to you; but try seeing it as the other direction and you will see that too!).
5. Identifying whether the pattern of correlation in a scatter diagram is linear tells you whether it is appropriate to use the standard procedures for figuring a linear correlation. If it is linear, identifying the direction and approximate strength of correlation before doing the figuring lets you check the results of your figuring when you are done.

The Correlation Coefficient

Looking at a scatter diagram gives a *rough idea* of the type and strength of relationship between two variables. But it is not a very precise approach. What you need is a number that gives the *exact correlation* (in terms of its direction and strength). For example, we saw that a positive linear correlation is when high scores go with highs, mediums with mediums, lows with lows. The exact correlation tells us precisely how much highs go with highs, and so on. As you just learned, in terms of a scatter diagram, a large linear correlation means that the dots fall close to a straight line (the line sloping up or down depending on whether the linear correlation is positive or negative). A *perfect linear correlation* means all the dots fall *exactly* on the straight line.

Logic of Figuring the Exact Linear Correlation

The first thing you need in figuring the linear correlation is some way to measure what is a high score and what is a low score. This means comparing scores on different variables in a consistent way. You can solve this problem of comparing apples and oranges by using *Z* scores.

TIP FOR SUCCESS

Don't be surprised if you find yourself having to read this section of the chapter a few times in order to fully understand the logic of figuring the exact correlation coefficient.

To review, a *Z* score is the number of standard deviations a score is from the mean. Whatever the range of values of the variable, if you change your raw scores to *Z* scores, a raw score that is high (i.e., above the mean of the scores on that variable) will always have a positive *Z* score. Similarly, a raw score that is low (below the mean) will always have a negative *Z* score. Furthermore, regardless of the particular measure used, *Z* scores tell you in a very standard way just how high or low each score is. A *Z* score of 1 is always exactly 1 standard deviation above the mean, and a *Z* score of 2 is twice as many standard deviations above the mean. *Z* scores on one variable are directly comparable to *Z* scores on another variable.

There is an additional reason why *Z* scores are so useful when figuring the exact correlation. It has to do with what happens if you multiply a score on one variable by a score on the other variable, which is called a *cross-product.* When using *Z* scores, this is called a **cross-product of *Z* scores.** If you multiply a high *Z* score by a high *Z* score, you will always get a positive cross-product. This is because no matter what the variable, scores above the mean are positive *Z* scores, and any positive number multiplied by any positive number has to be a a positive number. Furthermore—and here is where it gets interesting—if you multiply a low *Z* score by a low *Z* score, you also always get a positive cross-product. This is because no matter what the variable, scores below the mean are negative *Z* scores, and a negative multiplied by a negative gives a positive.

If highs on one variable go with highs on the other, and lows on one go with lows on the other, the cross-products of *Z* scores always will be positive. Considering a whole distribution of scores, suppose you take each person's *Z* score on one variable and multiply it by that person's *Z* score on the other variable. The result of doing this when highs go with highs and lows with lows is that the multiplications all come out positive. If you add up these cross-products of *Z* scores, which are all positive, for all the people in the study, you will end up with a large positive number.

On the other hand, with a negative correlation, highs go with lows and lows with highs. In terms of *Z* scores, this would mean positives with negatives and negatives with positives. Multiplied out, that gives all negative cross-products. If you add all these negative cross-products together, you get a large negative number.

Finally, suppose there is no linear correlation. In this situation, for some people highs on one variable would go with highs on the other variable (and some lows would go with lows), making positive cross-products. For other people, highs on one variable would go with lows on the other variable (and some lows would go with highs),

cross-product of *Z* scores The result of multiplying a person's *Z* score on one variable by the person's *Z* score on another variable.

making negative cross-products. Adding up these cross-products for all the people in the study would result in the positive cross-products and the negative cross-products cancelling each other out, giving a result of 0 (or close to 0).

In each situation, we changed all the scores to *Z* scores, multiplied each person's two *Z* scores by each other, and added up these cross-products. The result was a *large positive number* if there was a *positive linear correlation*, a *large negative number* if there was a *negative linear correlation*, and *a number near 0* if there was *no linear correlation*.

However, you are still left with the problem of figuring the *strength* of a positive or negative correlation. The larger the number, the bigger the correlation. But how large is large, and how large is not very large? You can't judge the strength of the correlation from the sum of the cross-products alone, because it gets bigger just by adding the cross-products of more people together. (That is, a study with 100 people would have a larger sum of cross-products than the same study with only 25 people.) The solution is to divide this sum of the cross-products by the number of people in the study. That is, you figure the *average of the cross-products of Z scores*. It turns out that because of the nature of *Z* scores, this average can never be more than +1, which would be a positive linear **perfect correlation.** It can never be less than −1, which would be a negative linear perfect correlation. In the situation of no linear correlation, the average of the cross-products of *Z* scores is 0.

TIP FOR SUCCESS

If you figure a correlation coefficient to be larger than +1 or less than −1, you have made a mistake in your figuring!

For a positive linear correlation that is not perfect, which is the usual situation, the average of the cross-products of *Z* scores is between 0 and +1. To put it another way, if the general trend of the dots is upward and to the right, but they do not fall exactly on a single straight line, this number is between 0 and +1. The same rule holds for negative correlations: They fall between 0 and −1.

The Correlation Coefficient

The average of the cross-products of *Z* scores is called the **correlation coefficient (*r*).** It is also called the *Pearson correlation coefficient* (or the *Pearson product–moment correlation coefficient*, to be very traditional). It is named after Karl Pearson. Pearson, along with Francis Galton (see Box 1), played a major role in developing the correlation coefficient. The correlation coefficient is abbreviated by the letter ***r***, which is short for *regression*, an idea closely related to correlation. (We discuss regression later in the chapter.)

The sign (+ or −) of a correlation coefficient tells you the general trend, in terms of whether it is a positive correlation (the dots on the scatter diagram go up and to the right) or a negative correlation (the dots go down and to the right). The actual value of the correlation coefficient—from a low of 0 to a high of 1, ignoring the sign of the correlation coefficient—tells you the strength of the linear correlation. So, a correlation coefficient of +.85 is a stronger linear correlation than a correlation of +.42. Similarly, a correlation of −.90 is a stronger linear correlation than +.85 (since .90 is bigger than .85). Another way of thinking of this is that in a scatter diagram, the closer the dots are to falling on a single straight line, the stronger the linear correlation. Figure 11 shows the scatter diagrams from Figure 9, with the correlation coefficient shown for each scatter diagram. To be sure you understand this, check that the correlation coefficient for each scatter diagram agrees roughly with the correlation coefficient you would expect based on the pattern of dots.

perfect correlation Relation between two variables that shows up on a scatter diagram as the dots exactly following a straight line; correlation of $r = 1$ or $r = -1$; situation in which each person's *Z* score on one variable is exactly the same as that person's *Z* score on the other variable.

correlation coefficient (*r*) Measure of the degree of linear correlation between two variables, ranging from −1 (a perfect negative linear correlation) through 0 (no correlation) to +1 (a perfect positive linear correlation); average of the cross-products of *Z* scores of two variables.

r Correlation coefficient.

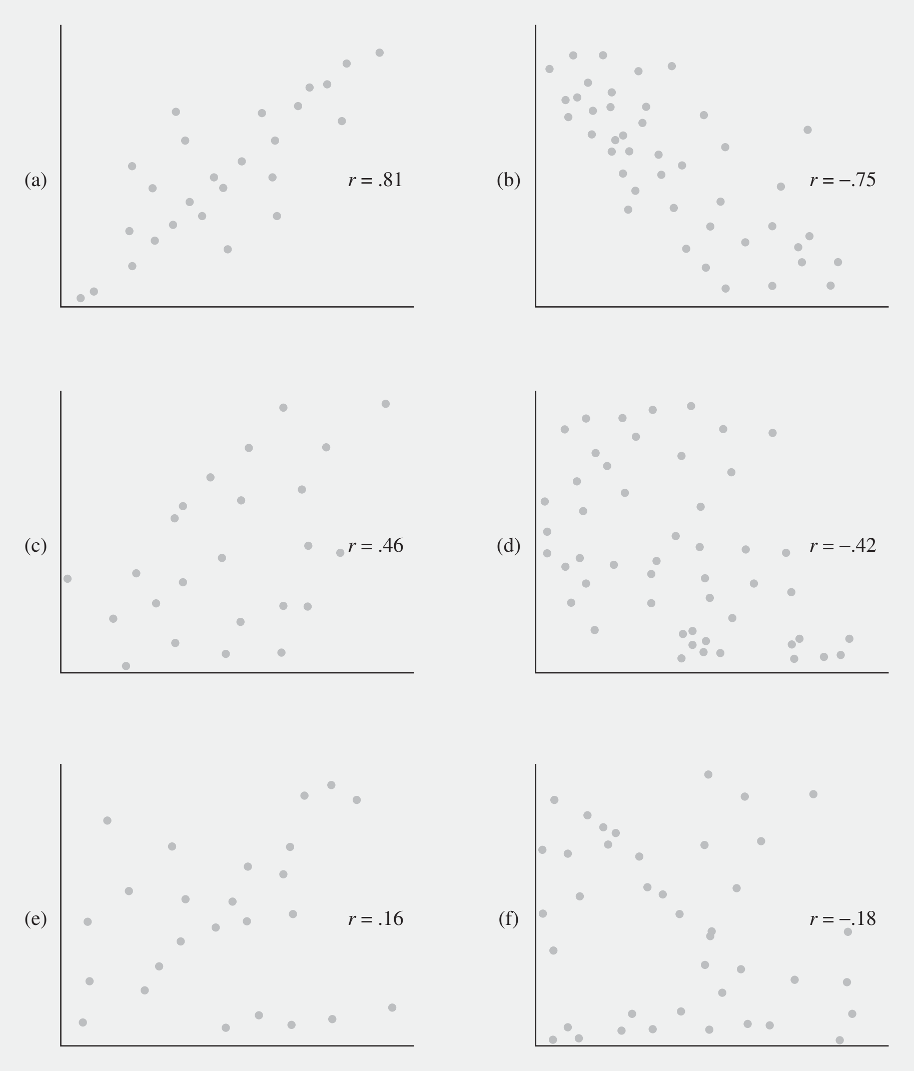


Figure 11 Examples of scatter diagrams and correlation coefficients for different degrees of correlation.

Formula for the Correlation Coefficient

The correlation coefficient, as we have seen, is the average of the cross-products of Z scores. Put as a formula,

$$r = \frac{\sum Z_X Z_Y}{N} \qquad (1)$$

The correlation coefficient is the sum, over all the people in the study, of the product of each person's two Z scores, then divided by the number of people.

r is the correlation coefficient. $\boldsymbol{Z_X}$ is the Z score for each person on the X variable and $\boldsymbol{Z_Y}$ is the Z score for each person on the Y variable. $Z_X Z_Y$ is Z_X multiplied by Z_Y (the cross-product of the Z scores) for each person and $\sum Z_X Z_Y$ is the sum of

Z_X Z score for variable X.

Z_Y Z score for variable Y.

the cross-products of *Z* scores over all the people in the study. *N* is the number of people in the study.[1]

Steps for Figuring the Correlation Coefficient

Here are the four steps for figuring the correlation coefficient.

❶ **Change all scores to *Z* scores.** This requires figuring the mean and the standard deviation of each variable, then changing each raw score to a *Z* score.
❷ **Figure the cross-product of the *Z* scores for each person.** That is, for each person, multiply the person's *Z* score on one variable by the person's *Z* score on the other variable.
❸ **Add up the cross-products of the *Z* scores.**
❹ **Divide by the number of people in the study.**

An Example

Let us try these steps with the sleep and mood example.

TIP FOR SUCCESS

When changing the raw scores to *Z* scores in Step ❶, you will make fewer mistakes if you do all the *Z* scores for one variable and then all the *Z* scores for the other variable. Also, to make sure you have done it correctly, when you finish all the *Z* scores for a variable, add them up—they should add up to 0 (within rounding error).

TIP FOR SUCCESS

When figuring the cross-products of the *Z* scores, pay careful attention to the sign of each *Z* score. As you know, a negative score multiplied by a negative score gives a positive score. Mistakes in this step are common, so do your figuring carefully!

❶ **Change all scores to *Z* scores.** Starting with the number of hours slept last night, the mean is 7 (sum of 42 divided by 6 students), and the standard deviation is 1.63 (sum of squared deviations, 16, divided by 6 students, for a variance of 2.67, the square root of which is 1.63). For the first student, the number of hours slept is 5. The *Z* score for this person is $(5 - 7)/1.63$, which is -1.23. Thus the first score is a *Z* score of -1.23. We figured the rest of the *Z* scores for the number of hours slept in the same way and you can see them in the appropriate column in Table 2. We also figured the *Z* scores for the happy mood scores and they are also shown in Table 2. For example, you will see that the happy mood *Z* score for the first student is -1.04.
❷ **Figure the cross-product of the *Z* scores for each person.** For the first student, multiply -1.23 by -1.04, which gives 1.28. The cross-products for all the students are shown in the last column of Table 2.
❸ **Add up the cross-products of the *Z* scores.** Adding up all the cross-products of *Z* scores, as shown in Table 2, gives a sum of 5.09.
❹ **Divide by the number of people in the study.** Dividing 5.09 by 6 (the number of students in the study) gives a result of .848, which rounds off to .85. This is the correlation coefficient.

In terms of the correlation coefficient formula,

$$r = \frac{\sum Z_X Z_Y}{N} = \frac{5.09}{6} = .85.$$

[1] There is also a "computational" version of this formula, which is mathematically equivalent and thus gives the same result:

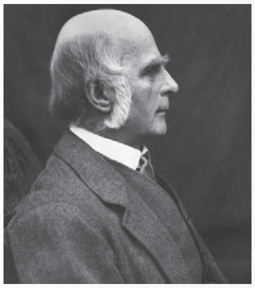
$$r = \frac{N\sum(XY) - \sum X \sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}} \qquad (2)$$

This formula is easier to use when figuring by hand when you have a large number of people in the study, because you don't have to first figure all the *Z* scores. However, researchers rarely use computational formulas like this anymore because most of the actual figuring is done by a computer. As a student learning statistics, it is better to use the definitional formula (1). This is because when solving problems using the definitional formula, you are strengthening your understanding of what the correlation coefficient means. In all examples in this chapter, we use the definitional formula and we urge you to use it in doing the chapter's practice problems.

Table 2 Figuring the Correlation Coefficient for the Sleep and Mood Study (Fictional Data)

Number of Hours Slept (X)				Happy Mood (Y)				Cross-Products
Deviation		Dev Squared	Z Scores	Deviation		Dev Squared	Z Scores	
X	$X - M$	$(X - M)^2$	Z_X	Y	$Y - M$	$(Y - M)^2$	Z_Y	$Z_X Z_Y$
5	−2	4	−1.23	2	−2	4	−1.04	1.28
7	0	0	0	4	0	0	0	0
8	1	1	.61	7	3	9	1.56	.95
6	−1	1	−.61	2	−2	4	−1.04	.63
6	−1	1	−.61	3	−1	1	−.52	.32
10	3	9	1.84	6	2	4	1.04	1.91
$\Sigma = 42$		$\Sigma(X - M)^2 = 16$		$\Sigma = 24$		$\Sigma(Y - M)^2 = 22$		$\Sigma Z_X Z_Y = 5.09$
$M = 7$		$SD^2 = 16/6 = 2.67$		$M = 4$		$SD^2 = 22/6 = 3.67$		$r = 5.09/6 = .85$
		$SD = 1.63$				$SD = 1.92$		

BOX 1 Galton: Gentleman Genius



Corbis/Bettman

Francis Galton is credited with inventing the correlation coefficient. (Karl Pearson worked out the formulas, but Pearson was a student of Galton and gave Galton all the credit.) Statistics at this time (around the end of the 19th century) was a tight little British club. In fact, most of science was an only slightly larger club. Galton also was influenced greatly by his own cousin, Charles Darwin.

Galton was a typical eccentric, independently wealthy gentleman scientist. Aside from his work in statistics, he possessed a medical degree, invented glasses for reading underwater, experimented with stereoscopic maps, dabbled in meteorology and anthropology, devised a system for classifying fingerprints that is still in use, and wrote a paper about receiving intelligible signals from the stars.

Above all, Galton was a compulsive counter. Some of his counts are rather infamous. Once while attending a lecture he counted the fidgets of an audience per minute, looking for variations with the boringness of the subject matter. While twice having his picture painted, he counted the artist's brushstrokes per hour, concluding that each portrait required an average of 20,000 strokes. While walking the streets of various towns in the British Isles, he classified the beauty of the female inhabitants using a recording device in his pocket to register "good," "medium," or "bad."

Galton's consuming interest, however, was the counting of geniuses, criminals, and other types in families. He wanted to understand how each type was produced so that science could improve the human race by encouraging governments to enforce eugenics—selective breeding for intelligence, proper moral behavior, and other qualities—to be determined, of course, by the eugenicists. (Eugenics has since been generally discredited.) The concept of correlation came directly from his first simple efforts in this area, the study of the relation of the height of children to their parents.

You can learn more about Galton on the following Web page: http://www-history.mcs.st-andrews.ac.uk/Biographies/Galton.html.

Sources: Peters (1987), Salsburg (2001), Tankard (1984).

Because this correlation coefficient is positive and near 1, the highest possible value, this is a *very strong positive linear correlation*.

An Example of Graphing and Figuring a Correlation

In this example, we put together the steps of making a scatter diagram and computing the correlation coefficient.

Table 3 Average Class Size and Achievement Test Scores in Five Elementary Schools (Fictional Data)

Elementary School	Class Size	Achievement Test Score
Main Street	25	80
Casat	14	98
Lakeland	33	50
Shady Grove	28	82
Jefferson	20	90

Suppose that an educational researcher knows the average class size and average achievement test score from the five elementary schools in a particular small school district, as shown in Table 3. (Again, it would be very rare in actual research practice to do a study or figure a correlation with only five cases. We have kept the numbers low here to make it easier for you to follow the steps of the example.) The question he then asks is, What is the relationship between these two variables?

The first thing he must do is to make a scatter diagram. This requires three steps.

❶ **Draw the axes and decide which variable goes on which axis.** Because it seems more reasonable to think of class size as affecting achievement test scores rather than the other way around, we will draw the axes with class size along the bottom.

❷ **Determine the range of values to use for each variable and mark them on the axes.** We will assume that the achievement test scores go from 0 to 100. Class size has to be at least 1 (thus close to zero so we can use the standard of starting our axes at zero) and in this example we guessed that it would be unlikely to be more than 50.

❸ **Mark a dot for each pair of scores.** The completed scatter diagram is shown in Figure 12.

As you learned earlier in the chapter, before figuring the correlation coefficient it is a good idea to look at the scatter diagram to be sure that the pattern is not curvilinear. In this example, the pattern of dots does not appear to be curvilinear (in fact, it is roughly a straight line), so you can assume that the pattern is linear. It is also wise to make a rough estimate of the direction and strength of correlation. This serves as a check against making a major mistake in figuring. In this example, the basic pattern is one in which

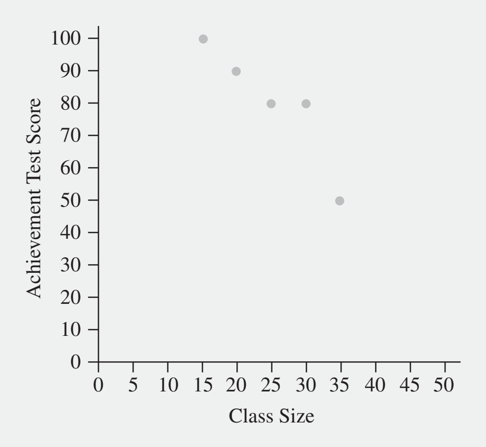


Figure 12 Scatter diagram for the scores in Table 3.

Table 4 Figuring the Correlation Coefficient for Average Class Size and Achievement Test Scores in Five Elementary Schools (Fictional Data)

School	Class Size		Achievement Test Score		Cross-Products
	X	Z_X ❶	Y	Z_Y ❶	Z_XZ_Y ❷
Main Street	25	.15	80	.00	.00
Casat	14	−1.53	98	1.10	−1.68
Lakeland	33	1.38	50	−1.84	−2.54
Shady Grove	28	.61	82	.12	.07
Jefferson	20	−.61	90	.61	−.37
Σ:	120		400		−4.52 ❸
M:	24		80		$r = -.90$ ❹
SD =	$\sqrt{214/5} = 6.54$		$\sqrt{1{,}328/5} = 16.30$		

the dots go down and to the right fairly consistently. This suggests a strong negative correlation. Now we can figure the correlation coefficient, following the usual steps.

❶ **Change all scores to *Z* scores.** The mean for class size is 24, and the standard deviation is 6.54. The *Z* score for the first class size, of 25, is .15. That is, $(25 - 24)/6.54 = .15$. All of the *Z* scores are shown in the appropriate columns in Table 4.

❷ **Figure the cross-product of the *Z* scores for each person.** For the first cross-product, multiply .15 by 0, which is 0. The second is −1.53 multiplied by 1.10, which is −1.68. All of the cross-products of *Z* scores are shown in the last column in Table 4.

❸ **Add up the cross-products of the *Z* scores.** The total is −4.52.

❹ **Divide by the number of people in the study.** The sum of the cross-products of *Z* scores (-4.52) divided by the number of schools (5) is −.90. That is, $r = -.90$.

In terms of the correlation coefficient formula,

$$r = \frac{\sum Z_X Z_Y}{N} = \frac{-4.52}{5} = -.90.$$

This correlation coefficient of −.90 agrees well with our original estimate of a strong negative correlation.

How are you doing?

1. Give two reasons why we use *Z* scores for figuring the exact linear correlation between two variables, thinking of correlation as how much high scores go with high scores and lows go with lows (or vice versa for negative correlations).
2. When figuring the correlation coefficient, why do you divide the sum of cross-products of *Z* scores by the number of people in the study?
3. Write the formula for the correlation coefficient and define each of the symbols.
4. Figure the correlation coefficient for the *Z* scores shown below for three people who were each tested on two variables, *X* and *Y*.

Person	Z_X	Z_Y
K	.5	−.7
L	−1.4	−.8
M	.9	1.5

Answers

1. First, Z scores put both variables on the same scale of measurement so that a high or low score (and how much it is high or low) means the same thing for both variables. Second, high Z scores are positive and low Z scores are negative. Thus, if highs go with highs and lows with lows, the cross-products of the Z scores will all be positive. Similarly, with a negative correlation where highs go with lows and lows with highs, the cross-products will all be negative.
2. You divide the sum of cross-products of the Z scores by the number of people in the study, because otherwise the more people in the study, the bigger the sum of the cross-products, even if the strength of correlation is the same. Dividing by the number of people corrects for this. (Also, when using Z scores, after dividing the sum of the cross-products by the number of people, the result has to be in a standard range of -1 to 0 to $+1$.)
3. The formula for the correlation coefficient is: $r = (\sum Z_X Z_Y)/N$. r is the correlation coefficient. $\sum$ is the symbol for sum of—add up all the scores that follow (in this formula, you add up all the cross-products that follow). Z_X is the Z score for each person's raw score on one of the variables (the one labeled X) and Z_Y is the Z score for each person's raw score on the other variable (labeled Y). N is the number of people in the study.
4. $r = (\sum Z_X Z_Y)/N = [(.5)(-.7) + (-1.4)(-.8) + (.9)(1.5)]/3$
   $= [-.35 + 1.12 + 1.35]/3 = 2.12/3 = .71.$

Issues in Interpreting the Correlation Coefficient

There are some subtle cautions in interpreting a correlation coefficient.

Causality and Correlation

If two variables have a clear linear correlation, we normally assume that there is something causing them to go together. However, you can't know the **direction of causality** (what is causing what) just from the fact that the two variables are correlated.

Three Possible Directions of Causality

Consider the example from the start of the chapter, the correlation between doing exciting activities with your partner and satisfaction with the relationship. There are three possible directions of causality for these two variables:

1. It could be that doing exciting activities together causes the partners to be more satisfied with their relationship.
2. It could also be that people who are more satisfied with their relationship choose to do more exciting activities together.
3. Another possibility is that something like having less pressure (versus more pressure) at work makes people happier in their marriage and also gives them more time and energy to do exciting activities with their partners.

These three possible directions of causality are shown in Figure 13a.

direction of causality Path of causal effect; if X is thought to cause Y, then the direction of causality is from X to Y.

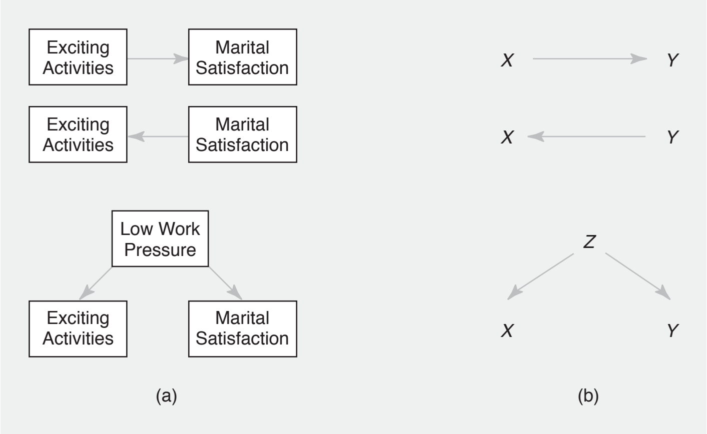


Figure 13 Three possible directions of causality (shown with arrows) for a correlation for (a) the exciting activities and marital satisfaction example and (b) the general principle for any two variables *X* and *Y*.

The principle is that for any correlation between variables *X* and *Y*, there are at least three possible directions of causality:

1. *X* could be causing *Y*.
2. *Y* could be causing *X*.
3. Some third factor could be causing both *X* and *Y*.

These three possible directions of causality are shown in Figure 13b.

It is also possible (and often likely) that there is more than one direction of causality making two variables correlated.

Ruling Out Some Possible Directions of Causality

Sometimes you can rule out one or more of these possible directions of causality based on additional knowledge of the situation. For example, the correlation between high school grades and college grades cannot be due to college grades causing high school grades—causality doesn't go backward in time. But we still do not know whether the high school grades somehow caused the college grades (e.g., by giving the students greater confidence), or some third factor, such as a tendency to study hard, makes for good grades in both high school and college.

In the behavioral and social sciences, one major strategy to rule out at least one direction of causality is to do studies where people are measured at two different points in time. This is called a **longitudinal study.** For example, we might measure a couple's level of exciting activities at one time and then examine the quality of their marriage several years later. Tsapelas, Aron, and Orbuch (2009) did exactly that and found that married couples who reported doing more exciting activities together were more satisfied with their relationship when they were surveyed again 9 years later.

Another major way we can rule out alternative directions of causality is by conducting a **true experiment.** In a true experiment, participants are randomly assigned (say, by flipping a coin) to a particular level of a variable and then measured on another variable.

longitudinal study A study where people are measured at two or more points in time.

true experiment A study in which participants are randomly assigned (say, by flipping a coin) to a particular level of a variable and then measured on another variable.

kristian sekulic/Getty

Figure 14 Picture of a couple engaging in the exciting activity task used by Aron et al. (2000).

For example, Aron et al. (2000) followed up their survey studies of married couples with a series of true experiments. Married couples came to their laboratory, spent 10 minutes doing a structured activity together, and then filled out a marital satisfaction questionnaire. What made this a true experiment is that half of the couples were randomly assigned (by flipping a coin) to do an activity that was exciting and the other half did an activity that was pleasant but not particularly exciting. As shown in Figure 14, the exciting activity involved the couple moving a cylindrical pillow back and forth over a barrier that was placed on large gym mats, without using their hands or arms, while being attached at the wrist and ankle with Velcro straps! The finding was that when those who had done the exciting activity filled out the marital satisfaction questionnaires, they reported substantially higher levels of marital satisfaction than did those who had done the pleasant but not exciting activity. As a result of this experiment, we can be confident that at least under these kinds of conditions, there is a direction of causality from the activities to marital satisfaction.

The main point to remember from all this is that just knowing that two variables are correlated, by itself, does not tell you anything about the direction of causality between them. *Understanding this principle is perhaps the single most important indication of sophistication in understanding behavioral and social science research!*

The Statistical Significance of a Correlation Coefficient

The correlation coefficient by itself is a descriptive statistic. It describes the direction (positive or negative) and strength of linear correlation in the particular group of people studied. However, when doing research, you often are more interested in a particular group of scores as representing some larger group that you have not studied directly. For example, the researcher studying sleep and mood tested only six individuals. But the researcher's intention in such a study is that the scores from these six people would tell us something about sleep and mood for people more generally. (In practice, you would want a much larger group than six for this purpose. We used small numbers of people in our examples to make them easier to learn from.)

There is a problem, however, in studying only some of the people in the larger group you want to know about. It is possible that, by chance, the ones you pick to study happen to be just those people for whom highs happen to go with highs and lows with lows—even though, had you studied all the people in the larger population, there might really be no correlation.

We say that a correlation is **statistically significant** if it is unlikely that you could have gotten a correlation as big as you did if in fact the overall group had no correlation. Specifically, you figure out whether that likelihood is less than some small degree of probability (p), such as .05 (5%) or .01 (1%). If the probability is that small, we say that the correlation is "statistically significant" with "$p < .05$" or "$p < .01$" (spoken as "p less than point oh five" or "p less than point oh one").

We bring up the topic of *statistical significance* now only to give you a general idea of what is being talked about if you see mentions of statistical significance—$p < .05$ or some such phrase—when reading a research article that reports correlation coefficients.

statistically significant Conclusion that the results of a study would be unlikely if in fact there were no association in the larger group you want to know about.

How are you doing?

1. If anxiety and depression are correlated, what are three possible directions of causality that might explain this correlation?
2. A researcher randomly assigns participants to eat zero, two, or four cookies and then asks them how full they feel. The number of cookies eaten and feeling full are highly correlated. What directions of causality can and cannot be ruled out?
3. What does it mean to say that a particular correlation coefficient is statistically significant?

Answers

1. Three possible directions of causality are: (a) being depressed can cause a person to be anxious; (b) being anxious can cause a person to be depressed; and (c) some third variable (such as some aspect of heredity or childhood trauma) could be causing both anxiety and depression.
2. Eating more cookies can cause participants to feel full. Feeling full cannot have caused participants to have eaten more cookies, because how many cookies were eaten was determined randomly. Third variables can't cause both, because how many cookies were eaten was determined randomly.
3. If a particular correlation coefficient is statistically significant, it means that the probability is very low of getting a correlation this big between these two variables in the group of people studied, if in fact there is no correlation between these two variables for people in general.

Prediction

TIP FOR SUCCESS

Be sure you have fully mastered the material on correlation before reading this part of the chapter.

Building on what you have already learned about correlations in this chapter, we now consider one of their major practical applications: making predictions. Behavioral and social scientists of various kinds are called on to make informed (and precise) guesses about such things as how well a particular job applicant is likely to perform if hired, how much a reading program is likely to help a particular third-grader, how likely a particular patient is to attempt to commit suicide, or how likely a potential parolee is to commit a violent crime if released.

Statistical prediction also plays a major part in helping behavioral and social scientists understand how various factors affect outcomes of interest. For example, what are the factors in people who marry that predict whether they will be happy and together 10 years later; what are the factors in childhood that predict depression and anxiety in adulthood; what are the circumstances of learning something that predict good or poor memory for it years later; or what are the various kinds of support from friends and family that predict how quickly or poorly someone recovers from a serious accident. Also, learning the details of statistical prediction prepares you for central themes in more advanced statistics courses.

We first consider procedures for making predictions about one variable, such as college GPA, based on information about another variable, such as SAT scores. Then, in an “Advanced Topic” section, we introduce situations in which predictions about one variable, such as college GPA, are made based on the combined information from two or more other variables, such as using both SAT scores and high school GPA.

Predictor (*X*) and Criterion (*Y*) Variables

With correlation it does not matter much which variable is which. But with prediction you have to decide which variable is being *predicted from* and which variable is being *predicted to*. The variable being predicted from is called the **predictor variable.** The variable being predicted to is called the **criterion variable.** In formulas the predictor variable is usually labeled *X*, the criterion variable, *Y*. That is, *X* predicts *Y*. In the example we just considered, SAT scores would be the predictor variable or *X* and college GPA would be the criterion variable or *Y* (see Table 5).

predictor variable (usually *X*) In prediction, variable that is used to predict scores of individuals on another variable.

criterion variable (usually *Y*) In prediction, a variable that is predicted.

Prediction Using *Z* Scores

It is easier to learn about prediction if we first consider prediction using *Z* scores. (We will get to prediction using ordinary scores shortly.)

Table 5 Predictor and Criterion Variables

	Variable Predicted From	Variable Predicted To
Name	Predictor variable	Criterion variable
Symbol	*X*	*Y*
Example	SAT scores	College GPA

The Prediction Model

The **prediction model,** or *prediction rule*, to make predictions with *Z* scores is as follows: A person's predicted *Z* score on the criterion variable is found by multiplying a particular number, called a **standardized regression coefficient,** by that person's *Z* score on the predictor variable. The standardized regression coefficient is symbolized by the Greek letter **beta (β)**.

β is called a standardized regression *coefficient* because a coefficient is a number you multiply by another number. It is called a standardized *regression* coefficient because the statistical method for prediction is sometimes called regression (for reasons we discuss later in the chapter). Finally, it is called a *standardized* regression coefficient because you are working with *Z* scores, which are also called *standard scores*.

Formula for the Prediction Model Using Z Scores. Here is the formula for the prediction model using *Z* scores (also known as the *Z-score prediction model*):

$$\text{Predicted } Z_Y = (\beta)(Z_X) \qquad (3)$$

A person's predicted *Z* score on the criterion variable is the standardized regression coefficient multiplied by that person's *Z* score on the predictor variable.

In this formula, Predicted Z_Y is the predicted value of the particular person's *Z* score on the criterion variable *Y*. (The predicted value of a score often is written with a hat symbol. Thus, $\hat{Z}_Y$ means Predicted Z_Y.) β is the standardized regression coefficient. Z_X is the particular person's *Z* score on the predictor variable *X*. Thus, $(\beta)(Z_X)$ means multiply the standardized regression coefficient by the person's *Z* score on the predictor variable.

For example, suppose that at your university the standardized regression coefficient (β) is .30 for predicting college GPA at graduation from SAT score at admission. So, the *Z*-score prediction model for predicting college GPA from high school SAT score is:

$$\text{Predicted } Z_Y = (.30)(Z_X)$$

A person applying to your school has an SAT score that is 2 standard deviations above the mean (i.e., a *Z* score of +2). The predicted *Z* score for this person's GPA would be .30 multiplied by 2, which is .60. That is, this person's predicted *Z* score for his or her college GPA is .60 standard deviations above the mean. In terms of the prediction model (Formula 3),

$$\text{Predicted } Z_Y = (\beta)(Z_X) = (.30)(2) = .60$$

Steps for the Prediction Model Using Z Scores. Here are the steps for the prediction model using *Z* scores.

❶ **Determine the standardized regression coefficient (β).**
❷ **Multiply the standardized regression coefficient (β) by the person's *Z* score on the predictor variable.**

We can illustrate the steps using the same example as above for predicting college GPA of a person at your school with an entering SAT 2 standard deviations above the mean.

❶ **Determine the standardized regression coefficient (β).** In the example, it was .30.
❷ **Multiply the standardized regression coefficient (β) by the person's *Z* score on the predictor variable.** In the example, the person's *Z* score on the predictor

prediction model Formula or rule for making predictions; that is, formula for predicting a person's score on a criterion variable based on the person's score on one or more predictor variables.

standardized regression coefficient (β) Regression coefficient in a prediction model using *Z* scores.

β standardized regression coefficient.

variable is 2. Multiplying .30 by 2 gives .60. Thus, .60 is the person's predicted *Z* score on the criterion variable (college GPA).

The Standardized Regression Coefficient (β)

It can be proved mathematically that the best number to use for the standardized regression coefficient when predicting one variable from another is the correlation coefficient. That is, when predicting one variable from another using *Z* scores, $\beta = r$.

An Example. Consider again the sleep and mood example from earlier in the chapter. In this example, six students had a correlation of .85 between number of hours slept the night before and happy mood that day. Because the correlation is .85, the standardized regression coefficient (β) is also .85. That is, $r = .85$, thus $\beta = .85$. This means that the model for predicting a person's *Z* score for happy mood is to multiply .85 by the person's *Z* score for the number of hours slept the night before.

Suppose you were thinking about staying up so late one night you would get only 4 hours' sleep. This would be a *Z* score of -1.84 on numbers of hours slept—that is, nearly 2 standard deviations less sleep than the mean. (We changed 4 hours to a *Z* score using the mean and standard deviation for the scores in this example and applying a procedure for changing raw scores to *Z* scores: $Z = (X - M)/SD$.) We could then predict your *Z* score on happy mood the next day by multiplying .85 by -1.84. The result comes out to -1.56. This means that based on the results of our little study, if you sleep only 4 hours tonight, tomorrow we would expect you to have a happy mood that is more than 1½ standard deviations below the mean (i.e., you would be very unhappy). In terms of the formula,

$$\text{Predicted } Z_Y = (\beta)(Z_X) = (.85)(-1.84) = -1.56$$

In terms of the steps,

❶ **Determine the standardized regression coefficient (β).** Because the correlation coefficient is .85, the standardized regression coefficient (β) is also .85.

❷ **Multiply the standardized regression coefficient (β) by the person's *Z* score on the predictor variable.** Your *Z* score on the predictor variable is -1.84. Multiplying .85 by -1.84 gives a predicted *Z* score on happy mood of -1.56.

By contrast, if you planned to get 9 hours' sleep, the prediction model would predict that tomorrow you would have a *Z* score for happy mood of .85 multiplied by 1.23 (the *Z* score when the number of hours slept is 9), which is $+1.05$. You would be somewhat happier than the average. In terms of the formula,

$$\text{Predicted } Z_Y = (\beta)(Z_X) = (.85)(1.23) = 1.05$$

(Incidentally, it is not a good idea to make predictions that involve values of the predictor variable very far from those in the original study. For example, you should not conclude that sleeping 20 hours would make you extremely happy the next day!)

Why Prediction Is Also Called Regression

Behavioral and social scientists often call this kind of prediction *regression*. Regression means going back or returning. We use the term *regression* here because in the usual situation in which there is less than a perfect correlation between two variables, the criterion variable *Z* score is some fraction of the predictor variable *Z* score. This fraction is β (the standardized regression coefficient). In our sleep and mood example, $\beta = .85$, thus the fraction is 85/100. This means that a person's predicted *Z* score on the criterion variable is 85/100 of the person's *Z* score on the predictor variable. As a result, the predicted *Z* score on the criterion variable is closer to the mean of 0 than is the *Z* score on the predictor variable. (In our sleep and mood example, when the *Z* score on the predictor variable was 1.23, the *Z* score predicted for the criterion variable was 1.05, a number closer to 0 than 1.23.) That is, the *Z* score on the criterion variable *regresses*, or goes back, toward a *Z* of 0.

How are you doing?

1. In words, what is the prediction model using *Z* scores?
2. Why does the standardized regression coefficient have this name? That is, explain the meaning of each of the three words that make up the term: standardized, regression, and coefficient.
3. Write the formula for the prediction model using *Z* scores, and define each of the symbols.
4. Figure the predicted *Z* score on the criterion variable (*Y*) in each of the following situations:

Situation	***r***	Z_X
a	.20	1.20
b	.50	2.00
c	.80	1.20

Answers

1. A person's predicted *Z* score on the variable being predicted about (the criterion variable) is the standardized regression coefficient (β) multiplied by the person's *Z* score on the variable being predicted from (the predictor variable).
2. The standardized regression coefficient is called *standardized* because you are predicting with *Z* scores, which are also called standard scores. It is called a *regression* coefficient because it is used in prediction, which is also called regression. (Prediction is called regression because the result of the prediction process is a predicted score on the criterion variable that is closer to the mean—goes back toward the mean—than is the score on the predictor variable.) It is called a *coefficient* because it is a number you multiply by another number.
3. The formula for the prediction model using *Z* scores is: Predicted $Z_Y = (\beta)(Z_X)$. Predicted Z_Y is the predicted *Z* score on the criterion variable. β is the standardized regression coefficient. Z_X is the *Z* score on the predictor variable.
4. (a) (.20) (1.20) = .24.
(b) (.50) (2.00) = 1.00.
(c) (.80) (1.20) = .96.

Prediction Using Raw Scores

Based on what you have learned, you can now also make predictions involving raw scores. To do this, change the raw score on the predictor variable to a *Z* score, make the prediction using the prediction model with *Z* scores, and then change the predicted *Z* score on the criterion variable to a raw score.

Steps of Raw-Score Prediction

❶ **Change the person's raw score on the predictor variable to a *Z* score.** That is, change *X* to Z_X. Based on the Formula $M = \frac{\sum X}{N}$, $Z_X = (X - M_X)/SD_X$.

❷ **Multiply the standardized regression coefficient (β) by the person's *Z* score on the predictor variable.** That is, multiply β (which is the same as *r*) by Z_X. This gives the predicted *Z* score on the criterion variable. This is Formula (3):

$$\text{Predicted } Z_Y = (\beta)(Z_X)$$

❸ **Change the person's predicted *Z* score on the criterion variable to a raw score.**[2] That is, change the Predicted Z_Y to Predicted *Y*. Based on the Formula $X = (Z)(SD) + M$,

$$\text{Predicted } Y = (SD_Y)(\text{Predicted } Z_Y) + M_Y$$

An Example. Recall our example from the sleep and mood study in which we wanted to predict your mood the next day if you sleep 4 hours the night before. In this example, the mean for sleep was 7 and the standard deviation was 1.63; for happy mood, the mean was 4 and the standard deviation was 1.92. The correlation between sleep and mood was .85.

❶ **Change the person's raw score on the predictor variable to a *Z* score.** $Z_X = (X - M_X)/SD_X = (4 - 7)/1.63 = -3/1.63 = -1.84$. That is, as we saw earlier, the *Z* score for 4 hours' sleep is -1.84.

❷ **Multiply the standardized regression coefficient (β) by the person's *Z* score on the predictor variable.** Predicted $Z_Y = (\beta)(Z_X) = (.85)(-1.84) = -1.56$. That is, as we also saw earlier, your predicted *Z* score for mood if you sleep only 4 hours is -1.56.

❸ **Change the person's predicted *Z* score on the criterion variable to a raw score.** Predicted $Y = (SD_Y)(\text{Predicted } Z_Y) + M_Y = (1.92)(-1.56) + 4 = -3.00 + 4 = 1.00$. In other words, using the prediction model based on the study of six students, we would predict that if you sleep only 4 hours tonight, tomorrow, on a happy mood rating scale from 0 to 8, you will have a rating of just 1—not happy at all!

Table 6 shows these steps worked out for sleeping 9 hours tonight.

Table 6 Steps, Formulas, and Example of Raw-Score Prediction

Step	Formula	Example
❶	$Z_X = (X - M_X)/SD_X$	$Z_X = (9 - 7)/1.63 = 1.23$
❷	Predicted $Z_Y = (\beta)(Z_X)$	Predicted $Z_Y = (.85)(1.23) = 1.05$
❸	Predicted $Y = (SD_Y)(\text{Predicted } Z_Y) + M_Y$	Predicted $Y = (1.92)(1.05) + 4 = 6.02$

[2] In practice, if you are going to make predictions for many different people, you would use a *raw-score prediction formula* that allows you to just plug in a particular person's raw score on the predictor variable and then solve directly to get the person's predicted raw score on the criterion variable. What the raw score prediction formula amounts to is taking the usual *Z*-score prediction formula, but substituting for the *Z* scores the formula for getting a *Z* score from a raw score. If you know the mean and standard deviation for both variables and the correlation coefficient, this whole thing can then be reduced algebraically to give the raw-score prediction formula. This raw-score prediction formula is of the form Predicted $Y = a + (b)(X)$ where *a* is called the *regression constant* (because this number that is added into the prediction does not change regardless of the value of *X*) and *b* is called the *raw-score regression coefficient* (because it is the number multiplied by the raw-score value of *X* and then added to *a* to get the predicted value of *Y*). You will sometimes see these terms referred to in research reports. However, since the logic of regression and its relation to correlation is most directly understood from the *Z*-score prediction formula, that is our emphasis in this introductory text.

How are you doing?

1. Explain the principle behind prediction using raw scores.
2. List the steps of making predictions using raw scores.
3. For a variable X, the mean is 10 and the standard deviation is 3. For a variable Y, the mean is 100 and the standard deviation is 10. The correlation of X and Y is .60. (a) Predict the score on Y for a person who has a score on X of 16. (b) Predict the score on Y for a person who has a score on X of 7. (c) Give the proportion of variance accounted for (r^2).
4. For a variable X, the mean is 20 and the standard deviation is 5. For a variable Y, the mean is 6 and the standard deviation is 2. The correlation of X and Y is .80. (a) Predict the score on Y for a person who has a score on X of 20. (b) Predict the score on Y for a person who has a score on X of 25. (c) Give the proportion of variance accounted for (r^2).

Answers

1. The principle behind prediction using raw scores is that you first change the raw score you are predicting from to a Z score, then make the prediction using the prediction model with Z scores, then change the predicted Z score to a predicted raw score.
2. ❶ **Change the person's raw score on the predictor variable to a *Z* score.**
   ❷ **Multiply the standardized regression coefficient (β) by the person's *Z* score on the predictor variable.**
   ❸ **Change the person's predicted *Z* score on the criterion variable to a raw score.**
3. (a) $Z_X = (X - M_X)/SD_X = (16 - 10)/3 = 6/3 = 2.$
   Predicted $Z_Y = (\beta)(Z_X) = (.60)(2) = 1.20.$
   Predicted $Y = (SD_Y)(\text{Predicted } Z_Y) + M_Y = (10)(1.20) + 100 = 12 + 100 = 112.$
   (b) $Z_X = (X - M_X)/SD_X = (7 - 10)/3 = -3/3 = -1.$
   Predicted $Z_Y = (\beta)(Z_X) = (.60)(-1) = -.60.$
   Predicted $Y = (SD_Y)(\text{Predicted } Z_Y) + M_Y = (10)(-.60) + 100 = -6 + 100 = 94.$
   (c) $r^2 = .60^2 = .36.$
4. (a) $Z_X = (X - M_X)/SD_X = (20 - 20)/5 = 0/5 = 0.$
   Predicted $Z_Y = (\beta)(Z_X) = (.80)(0) = 0.$
   Predicted $Y = (SD_Y)(\text{Predicted } Z_Y) + M_Y = (2)(0) + 6 = 0 + 6 = 6.$
   (b) $Z_X = (X - M_X)/SD_X = (25 - 20)/5 = 5/5 = 1.$
   Predicted $Z_Y = (\beta)(Z_X) = (.80)(1) = .80.$
   Predicted $Y = (SD_Y)(\text{Predicted } Z_Y) + M_Y = (2)(.80) + 6 = 1.60 + 6 = 7.60.$
   (c) $r^2 = .80^2 = .64.$

The Correlation Coefficient and the Proportion of Variance Accounted for

A correlation coefficient tells you the strength of a linear relationship. As you learned earlier in the chapter, bigger rs (values farther from 0) mean a stronger correlation. So, an r of .40 is a stronger correlation than an r of .20. However, it turns out that an r of .40 is *more than* twice as strong as an r of .20. To compare correlations with each other, you have to square each correlation (that is, you use r^2 instead of r). For example, a correlation of .20 is equivalent to an r^2 of .04, and a correlation of .40 is equivalent to an r^2 of .16. Therefore, a correlation of .20

actually means a relationship between *X* and *Y* that is only one-quarter as strong as a correlation of .40.

The correlation squared is called the **proportion of variance accounted for (r^2).** It is given this name because in the context of correlation it represents the proportion of the total variance in one variable (i.e., its total variability) that can be explained by the other variable.[3]

Correlation and Prediction in Research Articles

Scatter diagrams are occasionally included in research articles. For example, Gump and colleagues (2007) conducted a study of the level of lead in children's blood and the socioeconomic status (SES) of their families. The participants were 122 children who were taking part in an ongoing study of the developmental effects of environmental toxicants (chemicals that are poisonous). For each child in the study, the researchers determined the SES of the child's family based on the parents' occupation and education level. Also, between the ages of 2 and 3 years, the researchers took a blood sample from each child (with parental permission) and measured the amount of lead in it. As shown in Figure 15, Gump et al. used a scatter diagram to describe the relationship between family SES and children's blood levels of lead. There was a clear linear negative trend, with the researchers noting "... increasing family SES was significantly associated with declining blood levels" (p. 300). The scatter diagram shows that children from families with a higher SES had lower levels of lead in their blood. Of course, this is a correlational result. Thus, it does not necessarily mean that family SES directly influences the amount of lead in children's blood. It is possible that some other factor may explain this association or even that the amount of lead in the blood influenced SES. The researchers acknowledged this latter notion in the discussion section of their paper: "In addition, perhaps heightened blood lead in children (and their parents) affects cognitive functioning and thereby social and economic selection (failure to reach or keep expected social position) or drift (movement from higher to lower social class) occurs" (p. 302).

proportion of variance accounted for (r^2) Measure of association between variables used when comparing associations found in different studies or with different variables; correlation coefficient squared; the proportion of the total variance in one variable that can be explained by the other variable.

r^2 Proportion of variance accounted for.

Correlation coefficients are very commonly reported in research articles, both in the text of articles and in tables. The result with which we started the chapter would be described as follows: There was a positive correlation ($r = .51$) between excitement of activities done with partner and marital satisfaction. (Usually the "significance level" of the correlation will also be reported—in this example it would be $r = .51, p < .05$.) Sometimes a correlation coefficient is used to describe the consistency of a measure or test. One way to assess the consistency of a measure is to use it with the same group of people twice. The correlation between the two testings is

[3] The reason r^2 is called proportion of variance accounted for can be understood as follows: Suppose you used the prediction formula to predict each person's score on *Y*. Unless the correlation between *X* and *Y* was perfect (1.0), the variance of those predicted *Y* scores would be smaller than the variance of the original *Y* scores. There is less variance in the predicted *Y* scores because these predicted scores are on average closer to the mean than are the original scores. (We discussed this in the section on why prediction is called regression.) However, the more accurate the prediction, the more the predicted scores are like the actual scores. Thus, the more accurate the prediction, the closer the variance of the predicted scores is to the variance of the actual scores. Now suppose you divide the variance of the predicted *Y* scores by the variance of the original *Y* scores. The result of this division is the proportion of variance in the actual scores "accounted for" by the variance in the predicted scores. This proportion turns out (for reasons beyond what we can cover in this book) to be exactly r^2.

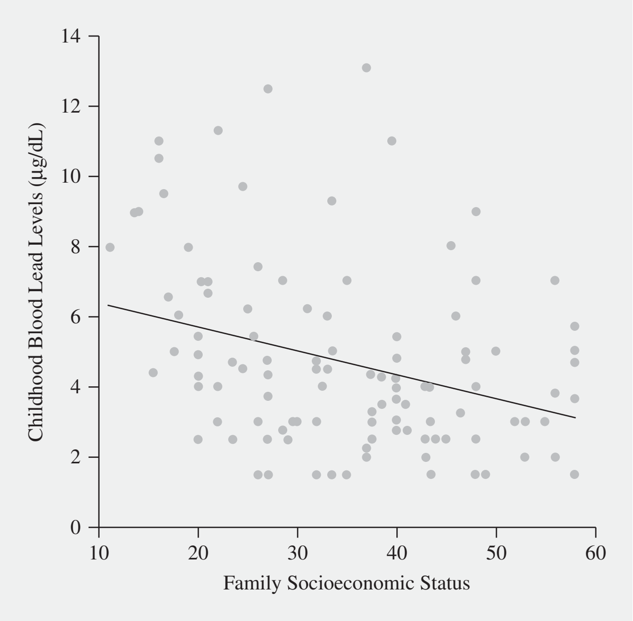


Figure 15 Children's family socioeconomic status (Hollingshead Index) as a function of childhood lead levels.

Source: Gump, B. B., Reihman, J., Stewart, P., Lonky, E., Darvill, T., & Matthews, K. A. (2007). Blood lead (Pb) levels: A potential environmental mechanism explaining the relation between socioeconomic status and cardiovascular reactivity in children. *Health Psychology, 26*, 296–304. Copyright © 2007 by the American Psychological Association. Reproduced with permission. The use of APA information does not imply endorsement by APA.

called *test–retest reliability*. For example, Morris, Gullekson, Morse, and Popovich (2009) conducted a study to update a test of attitudes toward computer usage. As part of the study, 48 undergraduates completed the test on two occasions about two weeks apart. The researchers noted that the measure had high test–retest reliability, with a correlation of $r = .93$ between the two testings.

Tables of correlations are common when several variables are involved. Usually, the table is set up so that each variable is listed down the left and also across the top. The correlation of each pair of variables is shown inside the table. This is called a **correlation matrix.**

Table 7 is a correlation matrix from a study of 114 expert Scrabble players (Halpern & Wai, 2007). The researchers asked the expert Scrabble players a series of questions about their Scrabble playing. The questions included the age at which they started playing, the age at which they started competing, the number of days a year and the number of hours per day they play Scrabble, and the number of years they had been practicing. The expert Scrabble players also provided their official Scrabble rating to the researchers. Table 7 shows the correlations among all the study measures.

This example shows several features that are typical of the way a correlation matrix is laid out. First, notice that the matrix does not include the correlation of a variable with itself. In this example, they put in a short line instead. Sometimes the correlation of a variable with itself is just left blank. Also notice that only the upper right triangle is filled in. This is because the lower left triangle would contain exactly

correlation matrix Common way of reporting the correlation coefficients among several variables in a research article; table in which the variables are named on the top and along the side and the correlations among them are all shown (only half of the resulting square, above or below the diagonal, is usually filled in, the other half being redundant).

Table 7 Correlations with Official Scrabble Ratings (Experts Only)

Variable	1	2	3	4	5	6	7	8	9
1. Official Scrabble rating	—	−.178	.116	−.173	−.202*	.021	−.128	.227*	.224*
2. Gender		—	.318*	.094	.265*	.104	−.181	.220*	.242*
3. Current age			—	.167	.727**	.088	−.094	.769**	.515**
4. Age started playing Scrabble				—	.355*	.233*	.094	−.501**	.058
5. Age started competing					—	.096	.112	.386*	.121
6. Days of year playing Scrabble						—	.050	−.093	−.196
7. Hours per day playing Scrabble							—	−.134	.377*
8. Years of practice								—	.492**
9. Total hours playing (Years × Hours)									—

*$p < .05$. **$p < .01$.

Source: Halpern, D. F., & Wal, J. (2007). The world of competitive Scrabble: Novice and expert differences in visiospatial and verbal abilities. *Journal of Experimental Psychology: Applied, 13,* 79–94. Copyright © 2007 by the American Psychological Association. Reproduced with permission. The use of APA information does not imply endorsement by APA.

the same information. For example, the correlation of official Scrabble rating with current age (which is .116) has to be the same as the correlation of current age with official Scrabble rating. Another shortcut saves space across the page: the names of the variables are listed only on the side of the table, with the numbers for them put across the top.

Looking at this example, among other results, you can see that there is a small to moderate negative correlation $(r = -.202)$ between official Scrabble rating and the age at which a person started competing in Scrabble. Also, there is a small to moderate correlation $(r = .227)$ between official Scrabble rating and the years of practice. The asterisks—* and **—after some of the correlation coefficients tell you that those correlations are statistically significant. The note at the bottom of the table tells you the significance levels associated with the asterisks.

It is rare for prediction models in research articles to focus on predicting a criterion variable from a single predictor variable. Instead, when prediction models are given in research articles, they are usually for *multiple regression* in which scores on a criterion variable are predicted from *two or more* predictor variables (see the "Advanced Topic" section below).

Advanced Topic: Multiple Regression

So far, we have predicted a person's score on a criterion variable using the person's score on a single predictor variable. Suppose you could use more than one predictor variable? For example, in predicting a happy mood, all you had to work with was the number of hours slept the night before. Suppose you also knew how well the person slept or how many dreams the person had. With this added information, you might be able to make a much more accurate prediction of mood.

The association between a criterion variable and two or more predictor variables is called **multiple correlation.** Making predictions in this situation is called **multiple regression.**

multiple correlation Correlation of a criterion variable with two or more predictor variables.

multiple regression Procedure for predicting scores on a criterion variable from scores on two or more predictor variables.

We explore these topics only briefly. However, multiple regression and correlation are frequently used in research articles in the behavioral and social sciences, so it is valuable for you to have a general understanding of them.

Multiple Regression Prediction Models

In multiple regression, each predictor variable has its own regression coefficient. The predicted *Z* score of the criterion variable is found by multiplying the *Z* score for each predictor variable by its standardized regression coefficient and then adding up the results. For example, here is the *Z*-score multiple regression formula with three predictor variables:

$$\text{Predicted } Z_Y = (\beta_1)(Z_{X_1}) + (\beta_2)(Z_{X_2}) + (\beta_3)(Z_{X_3}) \qquad (4)$$

The predicted *Z* score for the criterion variable is the standardized regression coefficient for the first predictor variable multiplied by the person's *Z* score on the first predictor variable, plus the standardized regression coefficient for the second predictor variable multiplied by the person's *Z* score on the second predictor variable, plus the standardized regression coefficient for the third predictor variable multiplied by the person's *Z* score on the third predictor variable.

Predicted Z_Y is the person's predicted score on the criterion variable. $\beta_1$is the standardized regression coefficient for the first predictor variable; β_2 and β_3 are the standardized regression coefficients for the second and third predictor variables. Z_{X_1} is the person's *Z* score for the first predictor variable; Z_{X_2} and Z_{X_3} are the person's *Z* scores for the second and third predictor variables. $(\beta_1)(Z_{X_1})$ means multiplying β_1 by Z_{X_1} and so forth.

For example, in the sleep and mood study, a multiple regression model for predicting happy mood (*Y*) using the predictor variables of number of hours slept, which we could now call X_1, and also a rating of how well you slept (X_2) and number of dreams during the night (X_3) might turn out to be as follows:

$$\text{Predicted } Z_Y = (.53)(Z_{X_1}) + (.28)(Z_{X_2}) + (.03)(Z_{X_3})$$

Suppose you were asked to predict the mood of a student who had a *Z* score of −1.82 for number of hours slept, a *Z* score of 2.34 for how well she slept, and a *Z* score of .94 for number of dreams during the night. That is, the student did not sleep very long, but did sleep very well, and had a few more dreams than average. You would figure the predicted *Z* score for happy mood by multiplying .53 by the number-of-hours slept *Z* score, multiplying .28 by the how-well-slept *Z* score, and multiplying .03 by the number-of-dreams *Z* score, then adding up the results:

$$\begin{aligned}\text{Predicted } Z_Y &= (.53)(-1.82) + (.28)(2.34) + (.03)(.94)\\ &= -.96 + .66 + .03 = -.27\end{aligned}$$

Thus, under these conditions, you would predict a happy mood *Z* score of −.27. This means a happy mood about one-quarter of a standard deviation below the mean. You can see that how well the student slept partially offset getting fewer hours' sleep. Given the very low standardized regression coefficient (β_3) for dreams in this model, once you have taken into account the number of hours slept and how well you slept, number of dreams (no matter how many or how few) would in general make very little difference in mood the next day.

In general terms, the size of the standardized regression coefficient for a predictor variable shows the amount of influence that variable has when predicting a score on the criterion variable. The larger the standardized regression coefficient for a predictor variable, the more influence that variable has when predicting a score on the criterion variable.

An Important Difference between Multiple Regression and Prediction Using One Predictor Variable

There is one particularly important difference between multiple regression and prediction when using only one predictor variable. In prediction when using one predictor variable, $\beta = r$. That is, the standardized regression coefficient is the same as the correlation coefficient. But in multiple regression, the standardized regression coefficient (β) for a predictor variable is *not* the same as the ordinary correlation coefficient (*r*) of that predictor with the criterion variable.

In multiple regression, a β will usually be closer to 0 than *r*. The reason is that part of what makes any one predictor successful in predicting the criterion variable will usually overlap with what makes the other predictors successful in predicting the criterion variable. In multiple regression, the standardized regression coefficient is about the unique, distinctive contribution of the predictor variable, excluding any overlap with other predictor variables.

Consider the sleep and mood example. When we were predicting mood using just the number of hours slept, β was the same as the correlation coefficient of .85. Now, with multiple regression, the β for number of hours slept is only .53. It is less because part of what makes number of hours slept predict mood overlaps with what makes sleeping well predict mood (in this fictional example, people who sleep more hours usually sleep well).

In multiple regression, the overall correlation between the criterion variable and all the predictor variables is called the **multiple correlation coefficient** and is symbolized as ***R***. However, because of the usual overlap among the predictor variables, the multiple correlation (*R*) is usually smaller than the sum of the individual *r*s of each predictor variable with the criterion variable. In multiple regression, the proportion of variance in the criterion variable accounted for by all the predictor variables taken together is the multiple correlation coefficient squared, R^2.

multiple correlation coefficient (*R*) Measure of degree of multiple correlation; positive square root of the proportion of variance accounted for in a multiple regression analysis.

R Multiple correlation coefficient.

How are you doing?

1. What is multiple regression?
2. Write the multiple regression prediction model with two predictors, and define each of the symbols.
3. In a multiple regression model, the standardized regression coefficient for the first predictor variable is .40 and for the second predictor variable is .70. What is the predicted criterion variable *Z* score for (a) a person with a *Z* score of +1 on the first predictor variable and a *Z* score of +2 on the second predictor variable, and (b) a person with a *Z* score of +2 on the first predictor variable and a *Z* score of +1 on the second predictor variable?
4. In multiple regression, why are the standardized regression coefficients for each predictor variable often smaller than the ordinary correlation coefficient of that predictor variable with the criterion variable?

Answers

1. Multiple regression is the procedure for predicting a criterion variable from a prediction rule that includes more than one predictor variable.
2. Predicted $Z_Y = (\beta_1)(Z_{X_1}) + (\beta_2)(Z_{X_2})$. Predicted Z_Y is the person's predicted score on the criterion variable. β_1 is the standardized regression coefficient for the first predictor variable, β_2 is the

standardized regression coefficient for the second predictor variable. Z_{X_1} is the person's Z score for the first predictor variable, and Z_{X_2} is the person's Z score for the second predictor variable.

3. (a) Predicted $Z_Y = (.40)(Z_{X_1}) + (.70)(Z_{X_2}) = (.40)(1) + (.70)(2) = .40 + 1.40$
$= 1.80.$

(b) Predicted $Z_Y = (.40)(Z_{X_1}) + (.70)(Z_{X_2}) = (.40)(2) + (.70)(1) = .80 + .70$
$= 1.50.$

4. In multiple regression, a predictor variable's association with the criterion variable usually overlaps with the other predictor variables' association with the criterion variable. Thus, the unique association of a predictor variable with the criterion variable (the standardized regression coefficient) is usually smaller than the ordinary correlation of the predictor variable with the criterion variable.

Advanced Topic: Multiple Regression in Research Articles

Multiple regression results are quite common and are often reported in tables. Buboltz, Johnson, and Woller (2003) conducted a study of the relationship between various aspects of college students' family relationships and students' level of "psychological reactance." "Reactance" in this study referred to a tendency to have an extreme reaction when your behavior is restricted in some way. Buboltz et al. used a multiple regression model to predict psychological reactance from three family characteristics: conflict, cohesion, and expressiveness. Each of these three characteristics represents a different dimension of the relationship among family members. The standardized regression coefficients for this multiple regression model are shown in Table 8 (in the β column). You can see that family conflict had a negative standardized regression coefficient ($\beta = -0.23$) and the standardized regression coefficients were both positive for family cohesion ($\beta = 0.22$) and family expressiveness ($\beta = 0.10$). You can also see that the standardized regression coefficients for conflict and cohesion were larger than the standardized regression coefficient for expressiveness. This tells you that family conflict and family cohesion were more important unique predictors of psychological reactance than family expressiveness.

Table 8 Summary of Regression Analysis Assessing the Unique Effects of Each Relationship Dimension Predicting Psychological Reactance

Relationship Dimension	*B*	*SE B*	β
Conflict	−1.71	0.49	-0.23^{***}
Cohesion	1.46	0.45	0.22^{***}
Expressiveness	0.90	0.52	0.10

Note: Beta coefficients for each predictor are over and above the other predictors.
$*p < .05$, $**p < .01$, $***p < .001$.
Source: Buboltz, W. C., Jr., Johnson, P., & Woller, K. M. P. (2003). Psychological reactance in college students: Family-of-origin predictors. *Journal of Counseling and Development, 81,* 311–317. Copyright © 2003 by ACA. Reprinted by permission. No further reproduction authorized without written permission of the American Counseling Association.

The table also includes the raw-score (unstandardized) regression coefficients (labeled with a capital *B* here). (We describe the raw score regression coefficient in footnote 2.) Finally, for each *B*, it gives what is called its standard error (*SE B*). These have to do with how accurately you can apply to the general population the coefficients they found in the particular sample used in this study.

Multiple correlations are rarely reported in research articles. When they are, it is usually in passing as part of a focus on multiple regression.

Learning Aids

Summary

1. Two variables are correlated when the two variables are associated in a clear pattern, for example, when high scores on one consistently go with high scores on the other, and lows on one go with lows on the other.
2. A scatter diagram shows the relationship between two variables. The lowest to highest possible values of one variable (the one you are predicting from, if they are distinguishable) are marked on the horizontal axis. The lowest to highest possible values of the other variable are marked on the vertical axis. Each individual's pair of scores is shown as a dot.
3. When the dots in the scatter diagram generally follow a straight line, this is called a linear correlation. In a curvilinear correlation, the dots follow a line pattern other than a simple straight line. No correlation exists when the dots do not follow any kind of line. In a positive linear correlation, the line goes upward to the right (so that low scores tend to go with lows and highs with highs). In a negative linear correlation, the line goes downward to the right (so that low scores generally go with highs and highs with lows).
4. The correlation coefficient (r) gives the direction and strength of linear correlation. It is the average of the cross-products of the Z scores. The correlation coefficient is highly positive when there is a strong positive linear correlation. This is because positive Z scores are multiplied by positive, and negative Z scores by negative. The correlation coefficient is highly negative when there is a strong negative linear correlation. This is because positive Z scores are multiplied by negative and negative Z scores by positive. The coefficient is 0 when there is no linear correlation. This is because positive Z scores are sometimes multiplied by positive and sometimes by negative Z scores and negative Z scores are sometimes multiplied by negative and sometimes by positive. Thus, positive and negative cross-products cancel each other out.
5. The maximum positive value of r is $+1$. $r = +1$ when there is a perfect positive linear correlation. The maximum negative value of r is -1. $r = -1$ when there is a perfect negative linear correlation.
6. The actual value of the correlation coefficient—from a low of 0 to a high of 1, ignoring the sign of the correlation coefficient—tells you the strength of the linear correlation. The closer a correlation coefficient is to 1 or to -1, the stronger the linear correlation.
7. Correlation does not tell you the direction of causation. If two variables, X and Y, are correlated, this could be because X is causing Y, Y is causing X, or a third factor is causing both X and Y.

8. A correlation figured using scores from a particular group of people is often intended to apply to people in general. A correlation is statistically significant when statistical procedures (taught later in this book) tell you that it is highly unlikely that you would get a correlation as big as the one found with the group of people studied, if in fact there were no correlation between these two variables among people in general.
9. Prediction (or regression) makes predictions about scores on a criterion variable based on scores on a predictor variable. The prediction model for predicting a person's *Z* score on the criterion variable is to multiply the standardized regression coefficient (β) by the person's *Z* score on the predictor variable. The best number to use for the standardized regression coefficient in this situation is the correlation coefficient (*r*).
10. Predictions with raw scores can be made by changing a person's score on the predictor variable to a *Z* score, multiplying it by the standardized regression coefficient (β), and then changing the predicted criterion variable *Z* score to a raw score.
11. Comparisons of the strength of linear correlation are considered most accurate in terms of the correlation coefficient squared (r^2), the proportion of variance accounted for.
12. Correlational results are usually presented in research articles either in the text with the value of *r* (and sometimes the significance level) or in a table (a correlation matrix) showing the correlations among several variables. Results of prediction in which a criterion variable is predicted from a single variable are rarely described directly in research articles.
13. ADVANCED TOPIC: In multiple regression, a criterion variable is predicted from two or more predictor variables. In a multiple regression model, each predictor variable is multiplied by its own standardized regression coefficient, and the results are added up to make the prediction. However, because the predictor variables overlap in their influence on the criterion variable, each of the regression coefficients generally is smaller than the variable's correlation coefficient with the criterion variable. The multiple correlation coefficient (*R*) is the overall degree of association between the criterion variable and the predictor variables taken together. The multiple correlation coefficient squared (R^2) is the proportion of variance in the criterion variable accounted for by all the predictor variables taken together. Multiple regressions are commonly reported in articles, often in a table that includes the regression coefficients.

correlation
scatter diagram
linear correlation
curvilinear correlation
no correlation
positive correlation
negative correlation
cross-product of *Z* scores
perfect correlation

correlation coefficient (*r*)
r
Z_X
Z_Y
direction of causality
longitudinal study
true experiment
statistically significant
predictor variable
criterion variable

prediction model
standardized regression coefficient (β)
proportion of variance accounted for (r^2)
correlation matrix
multiple correlation
multiple regression
multiple correlation coefficient (*R*)

Example Worked-Out Problems

Making a Scatter Diagram and Describing the General Pattern of Association

Based on the number of hours studied and the test score for the five students in the following table, make a scatter diagram and describe in words the general pattern of association.

Student	Hours Studied	Test Score
A	0	52
B	10	92
C	6	75
D	8	71
E	6	64

Answer

The steps in solving the problem follow; Figure 16 shows the scatter diagram with markers for each step.

❶ **Draw the axes and decide which variable goes on which axis.** Studying comes before the test score. Thus, we will put number of hours studied along the bottom.

❷ **Determine the range of values to use for each variable and mark them on the axes.** We will assume that the test scores go from 0 to 100. We do not know the maximum possible for the number of hours studied, but let us assume a maximum of 10 in this example.

❸ **Mark a dot for each pair of scores.** For example, to mark the dot for student D, you go across to 8 and up to 71.

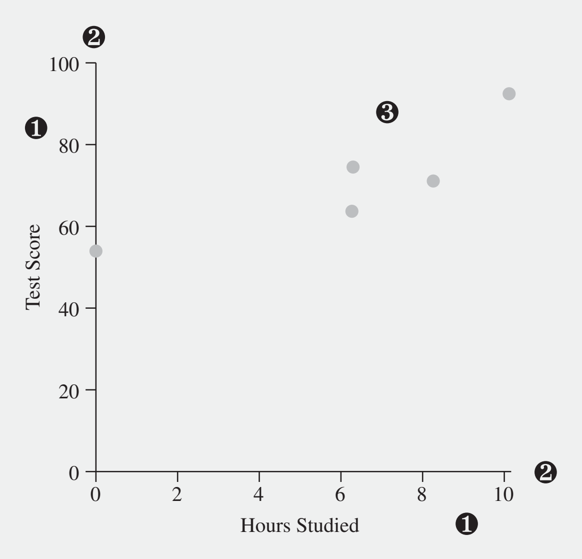


Figure 16 Scatter diagram for scores in Example Worked-Out Problem. ❶ Draw the axes and decide which variable goes on which axis. ❷ Determine the range of values to use for each variable and mark them on the axes. ❸ Mark a dot for each pair of scores.

The general pattern is roughly linear. Its direction is positive (it goes up and to the right, with more hours studied going with higher test scores and vice versa). It is a quite strong correlation, since the dots all fall fairly close to a straight line—it should be fairly close to +1. In words, it is a strong, linear, positive correlation.

Figuring the Correlation Coefficient

Figure the correlation coefficient for hours studied and test score in the preceding example.

Answer

You can figure the correlation using either the formula or the steps. The basic figuring is shown in Table 9 with markers for each of the steps.

Using the formula,

$$r = (\sum Z_X Z_Y)/N = 4.49/5 = .90.$$

Using the steps,

❶ **Change all scores to *Z* scores.** For example, the mean for hours studied is 6, and the standard deviation is 3.35. Thus, the Z score for student A, who studied for 0 hours, is $(0 - 6)/3.35 = -1.79$.
❷ **Figure the cross-product of the *Z* scores for each person.** For example, for student A, the cross-product is −1.79 multiplied by −1.43, which is 2.56. For student B, it is 1.19 multiplied by 1.61, which equals 1.92.
❸ **Add up the cross-products of the *Z* scores.** The total is 4.49.
❹ **Divide by the number of people in the study.** The sum (4.49) divided by 5 is .90; that is, $r = .90$.

Outline for Writing Essays on the Logic and Figuring of a Correlation Coefficient

1. If the question involves creating a scatter diagram, explain how and why you created the diagram to show the pattern of relationship between the two variables. Explain the meaning of the term *correlation*. Mention the type of correlation (e.g., linear; positive or negative; small, moderate, or strong) shown by the scatter diagram.

Table 9 Figuring Correlation Coefficient for Answer to Example Worked-Out Problem

	Hours Studied		Test Score		Cross-Products
	X	Z_X ❶	Y	Z_Y ❶	$Z_X Z_Y$ ❷
	0	−1.79	52	−1.43	2.56
	10	1.19	92	1.61	1.92
	6	.00	75	.32	.00
	8	.60	71	.02	.01
	6	.00	64	−.52	.00
Σ:	30		354		**4.49** ❸
M:	6		70.8		$r = 4.49/5 = .90$ ❹
SD:	$\sqrt{56/5} = 3.35$		$\sqrt{866.8/5} = 13.17$		

2. Explain the idea that a correlation coefficient provides an indication of the direction and strength of linear correlation between two variables.
3. Outline and explain the steps for figuring the correlation coefficient. Be sure to mention that the first step involves changing all of the scores to *Z* scores. (If required by the question, explain the meaning of *Z* scores, mean, and standard deviation.) Describe how to figure the cross-products of the *Z* scores. Explain why the cross-products of the *Z* scores will tend to be positive if the correlation is positive and will tend to be negative if the correlation is negative. Mention that the correlation coefficient is figured by taking the mean of the cross-products of the *Z* scores so that it does not get higher just because there are more cases. Explain what the value of the correlation coefficient means in terms of the direction and strength of linear correlation.
4. Be sure to discuss the direction and strength of correlation of your particular result.

Prediction Using *Z* Scores

Based on the data shown in the following table (these are the same data used for an example earlier in the chapter), predict the *Z* scores for achievement for schools that have class sizes with *Z* scores of −2, −1, 0, +1, +2.

Elementary School	Class Size	Achievement Test Score
Main Street	25	80
Casat	14	98
Lakeland	33	50
Shady Grove	28	82
Jefferson	20	90
M	24	80
SD	6.54	16.30
$r = -.90$		

Answer

This can be done using either the formula or the steps. Using the formula, Predicted $Z_Y = (\beta)(Z_X)$, the *Z*-score prediction model for this problem is,

$$\begin{aligned}\text{Predicted } Z_Y &= (-.90)(Z_X) = (-.90)(-2) = 1.80.\\ &= (-.90)(-1) = .90.\\ &= (-.90)(0) = 0.\\ &= (-.90)(+1) = -.90.\\ &= (-.90)(+2) = -1.80.\end{aligned}$$

Using the steps,

❶ **Determine the standardized regression coefficient (β).** The correlation coefficient is −.90. Thus, $\beta = -.90$.

❷ **Multiply the standardized regression coefficient by the person's *Z* score on the predictor variable.** $-.90 \times -2 = 1.80$; $-.90 \times -1 = .90$; $-.90 \times 0 = 0$; $-.90 \times 1 = -.90$; $-.90 \times 2 = -1.80$.

Prediction Using Raw Scores

Using the data from the example above on class size and achievement test score, predict the raw scores for achievement for a school that has a class size of 27.

Answer

Using the steps,

❶ **Change the person's raw score on the predictor variable to a Z score (note that in this example, we have a school's raw score).** $Z_X = (X - M_X)/SD_X = (27 - 24)/6.54 = 3/6.54 = .46$.

❷ **Multiply the standardized regression coefficient (β) by the person's predictor variable *Z* score.** Predicted $Z_Y = (\beta)(Z_X) = (-.90)(.46) = -.41$. That is, the predicted *Z* score for this school is $-.41$.

❸ **Change the person's predicted *Z* score on the criterion variable to a raw score.** Predicted $Y = (SD_Y)(\text{Predicted } Z_Y) + M_Y = (16.30)(-.41) + 80 = -6.68 + 80 = 73.32$. In other words, using the prediction model based on the study of five schools, we would predict that a school with an average class size of 27 students will have an average achievement test score of 73.32.

Advanced Topic: Multiple Regression Predictions

A (fictional) researcher studied the talkativeness of children in families with a mother, a father, and one grandparent. The researcher found that the child's talkativeness score depended on the quality of the child's relationship with each of these people. The multiple regression prediction model using *Z* scores is as follows:

$$\text{Predicted talkativeness Z score of the child} = (.32)(Z \text{ mother}) + (.21)(Z \text{ father}) + (.11)(Z \text{ grandparent})$$

Predict a child's talkativeness *Z* score who had *Z* scores for relationship quality of .48 with mother, −.63 with father, and 1.25 with grandparent.

Answer

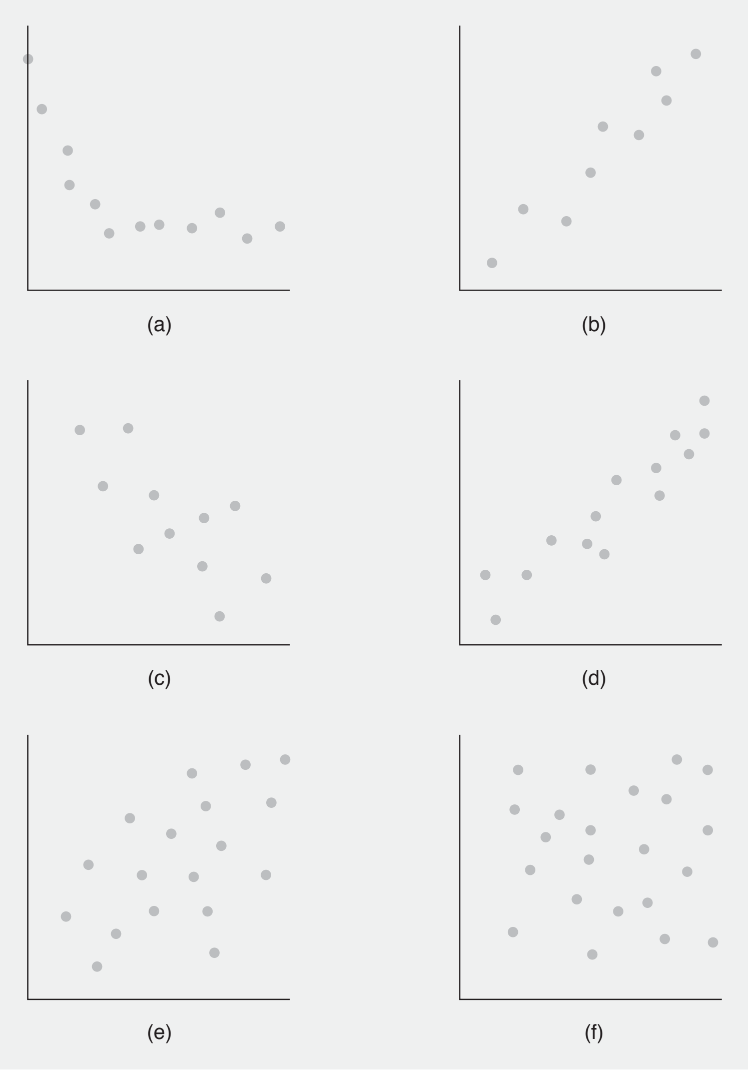
Predicted talkativeness *Z* score of the child $= (.32)(Z \text{ mother}) + (.21)(Z \text{ father}) + (.11)(Z \text{ grandparent}) = (.32)\,(.48) + (.21)(-.63) + (.11)(1.25) = .15 + -.13 + .14 = .16$.

These problems involve figuring. Most real-life statistics problems are done on a computer with special statistical software. Even if you have such software, do these problems by hand to ingrain the method in your mind. To learn how to use a computer to solve statistics problems like those in this chapter, refer to the "Using SPSS" section at the end of this chapter.

All data are fictional unless an actual citation is given.

Set I (for answers, see the end of this chapter)

1. For each of the following scatter diagrams, indicate whether the pattern is linear, curvilinear, or no correlation; if it is linear, indicate whether it is positive or negative and approximately how strong the correlation is (large, moderate, small).



2. (a) The following have been prepared so that data sets B through D are slightly modified versions of data set A. Make scatter diagrams and figure the correlation coefficients for each data set. (b) Discuss how and why the correlations change.

Data Set A		**Data Set B**		**Data Set C**		**Data Set D**	
X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
1	1	1	1	1	5	1	1
2	2	2	2	2	2	2	4
3	3	3	3	3	3	3	3
4	4	4	5	4	4	4	2
5	5	5	4	5	1	5	5

3. A researcher is interested in whether a new drug affects the development of a cold. Eight people are tested: four take the drug and four do not. (Those who take it are rated 1; those who don't, 0.) Whether they get a cold (rated 1) or not (0) is recorded. Four possible results are shown. Figure the correlation coefficient for each possibility (A, B, C, and D).

Possibility A		**Possibility B**		**Possibility C**		**Possibility D**	
Take Drug	*Get Cold*	*Take Drug*	*Get Cold*	*Take Drug*	*Get Cold*	*Take Drug*	*Get Cold*
0	1	0	1	0	1	0	1
0	1	0	1	0	1	0	1
0	1	0	1	0	0	0	1
0	1	0	0	0	0	0	0
1	0	1	1	1	1	1	0
1	0	1	0	1	1	1	0
1	0	1	0	1	0	1	0
1	0	1	0	1	0	1	0

For problems 4 and 5, (a) make a scatter diagram of the raw scores; (b) describe in words the general pattern of correlation, if any; and (c) figure the correlation coefficient. In these problems, the mean, standard deviation, and *Z* scores for each variable are given to save you some figuring.

4. The Louvre Museum in Paris is interested in the relation of the age of a painting to public interest in it. The number of people stopping to look at each of 10 randomly selected paintings is observed over a week. The results are as shown:

Painting Title	**Approximate Age (Years)**		**Number of People Stopping to Look**	
	(M = 268.40, SD = 152.74)		**(M = 88.20, SD = 29.13)**	
	X	Z_X	Y	Z_Y
The Entombment	480	1.39	68	−.69
The Mystic Marriage of St. Catherine	530	1.71	71	−.59
The Bathers	255	−.09	123	1.19
The Toilette	122	−.96	112	.82
Portrait of Castiglione	391	.80	48	−1.38
Charles I of England	370	.67	84	−.14
Crispin and Scapin	155	−.75	66	−.76
Nude in the Sun	130	−.91	148	2.05
The Balcony	137	−.86	71	−.59
The Circus	114	−1.01	91	.10

5. A schoolteacher thought that he had observed that students who dressed more neatly were generally better students. To test this idea, the teacher had a friend rate each of the students for neatness of dress. Following are the ratings for neatness, along with each student's score on a standardized school achievement test.

Child	Neatness Rating (M = 19.60, SD = 3.07)		Achievement Test (M = 63.10, SD = 4.70)	
	X	Z_X	Y	Z_Y
Janet	18	−.52	60	−.66
Michael	24	1.43	58	−1.09
Grove	14	−1.82	70	1.47
Kevin	19	−.20	58	−1.09
Joshua	20	.13	66	.62
Emily	23	1.11	68	1.04
Susan	20	.13	65	.40
Tyler	22	.78	68	1.04
Sarah	15	−1.50	56	−1.51
Chad	21	.46	62	−.23

For problems 6 and 7, do the following: (a) Make a scatter diagram of the raw scores; (b) describe in words the general pattern of correlation, if any; (c) figure the correlation coefficient; (d) explain the logic of what you have done, writing as if you are speaking to someone who has never had a statistics course (but who does understand the mean, standard deviation, and *Z* scores); (e) give three logically possible directions of causality, saying for each whether it is a reasonable direction in light of the variables involved (and why); (f) make raw score predictions on the criterion variable for persons with *Z* scores on the predictor variable of −2, −1, 0, +1, +2; and (g) give the proportion of variance accounted for (r^2).

6. Four young children were monitored closely over a period of several weeks to measure how much they watched violent television programs and their amount of violent behavior toward their playmates. (For part (f), assume that hours watching violent television is the predictor variable.) The results were as follows:

Child's Code Number	Weekly Viewing of Violent TV (hours)	Number of Violent or Aggressive Acts toward Playmates
G3368	14	9
R8904	8	6
C9890	6	1
L8722	12	8

7. A political scientist studied the relation between the number of town-hall meetings held by the four candidates for mayor in a small town and the percentage of people in the town who could name each candidate. (For part (f), assume that the number of town-hall meetings is the predictor variable.) Here are the results:

Mayor Candidate	Number of Town-Hall Meetings	Percentage of People Who Can Name Candidate
A	4	70
B	5	94
C	2	36
D	1	48

Table 10 Zero-Order Correlations for Study Variables

Variable	1	2	3	4	5	6	7	8	9	10
1. Women's report of stress	—									
2. Men's report of women's stress	.17	—								
3. Partner Support 1	−.28*	−.18	—							
4. Partner Support 2	−.27*	−.18	.44***	—						
5. Depressed Mood 1	.23*	.10	−.34**	−.17	—					
6. Depressed Mood 2	.50***	.14	−.42***	−.41***	.55***	—				
7. Women's age	.06	.16	.04	−.24*	−.35*	−.09	—			
8. Women's ethnicity	−.19	−.09	−.16	−.14	.11	.13	−.02	—		
9. Women's marital status	−.18	.01	.12	.24*	−.04	−.20	.05	−.34**	—	
10. Parity	.19	.13	−.11	−.17	.10	.16	.26*	.31*	−.12	—

*$p < .05$, **$p < .01$, ***$p < .001$.

Source: Chapman, H. A., Hobfoll, S. E., & Ritter, C. (1997). Partners' stress underestimations lead to women's distress: A study of pregnant inner-city women. *Journal of Personality and Social Psychology, 73,* 418–425. Copyright © 1997 by the American Psychological Association. Reproduced with permission. The use of APA information does not imply endorsement by APA.

8. Chapman, Hobfoll, and Ritter (1997) interviewed 68 pregnant inner-city women and their husbands (or boyfriends) twice during their pregnancy, once between 3 and 6 months into the pregnancy and again between 6 and 9 months into the pregnancy. Table 10 shows the correlations among several of their measures. ("Zero-Order Correlations" means the same thing as ordinary correlations.) Most important in this table are the correlations among women's reports of their own stress, men's reports of their partners' stress, women's perception of their partners' support at the first and at the second interviews, and women's depression at the first and at the second interviews.

   Explain the results for these measures as if you were writing to a person who has never had a course in statistics. Specifically, (a) explain what is meant by a correlation coefficient using one of the correlations as an example; (b) study the table and then comment on the patterns of results in terms of which variables are relatively strongly correlated and which are not very strongly correlated; and (c) comment on the limitations of making conclusions about direction of causality based on these data, using a specific correlation as an example (noting at least one plausible alternative causal direction and why that alternative is plausible).
9. A researcher working with hockey players found that knowledge of fitness training principles correlates .40 with number of injuries received over the subsequent year. The researcher now plans to test all new athletes for their knowledge of fitness training principles and use this information to predict the number of injuries they are likely to receive. Indicate the (a) predictor variable, (b) criterion variable, and (c) standardized regression coefficient. (d) Write the Z-score prediction model. Indicate the predicted Z scores for number of injuries for athletes whose Z scores on the principles of fitness training test are (e) -2, (f) -1, (g) 0, (h) $+1$, and (i) $+2$.
10. ADVANCED TOPIC: Gunn, Biglan, Smolkowski, and Ary (2000) studied reading in a group of Hispanic and non-Hispanic third-graders. As part of this study, they did an analysis predicting reading comprehension (called "passage comprehension") from three more specific measures of reading ability: "Letter–Word Identification" (ability to read irregular words), "Word Attack" (ability to use phonic and structural analysis), and "Oral Reading Fluency" (correct words per

Table 11 Multiple Regression Predicting Comprehension from Decoding Skill and Oral Reading Fluency

Variable	Beta	t	p	R^2	F	p
Passage Comprehension raw score						
Letter-Word Identification	−.227	−2.78	.006			
Word Attack	.299	3.75	.001			
Oral Reading Fluency—	.671	9.97	.001			
Correct words per minute				.534	57.70	.001

Source: Gunn, B., Biglan, A., Smolkowski, K., & Ary, D. (2000). The efficacy of supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school. *Journal of Special Education, 34,* 90–103. Copyright © 2000, Hammill Institute on Disabilities. Reprinted by permission of Sage Publications.

minute). The results are shown in Table 11. Explain the results as if you were writing to a person who understands correlation but has never learned anything about regression or multiple regression analysis. (Ignore the columns for *t*, *p*, *F*, and *p*. These have to do with statistical significance.)

11. ADVANCED TOPIC: Based on Table 11 (from Gunn et al., 2000), (a) determine the *Z* score multiple regression formula, and (b) calculate the predicted passage comprehension *Z* score for each of the following third-graders (figures are *Z* scores):

Third-Grader	Letter–Word Identification	Word Attack	Oral Reading Fluency
A	1	1	1
B	0	0	0
C	−1	−1	−1
D	1	0	0
E	0	1	0
F	0	0	1
G	3	1	1
H	1	3	1
I	3	1	3

Set II

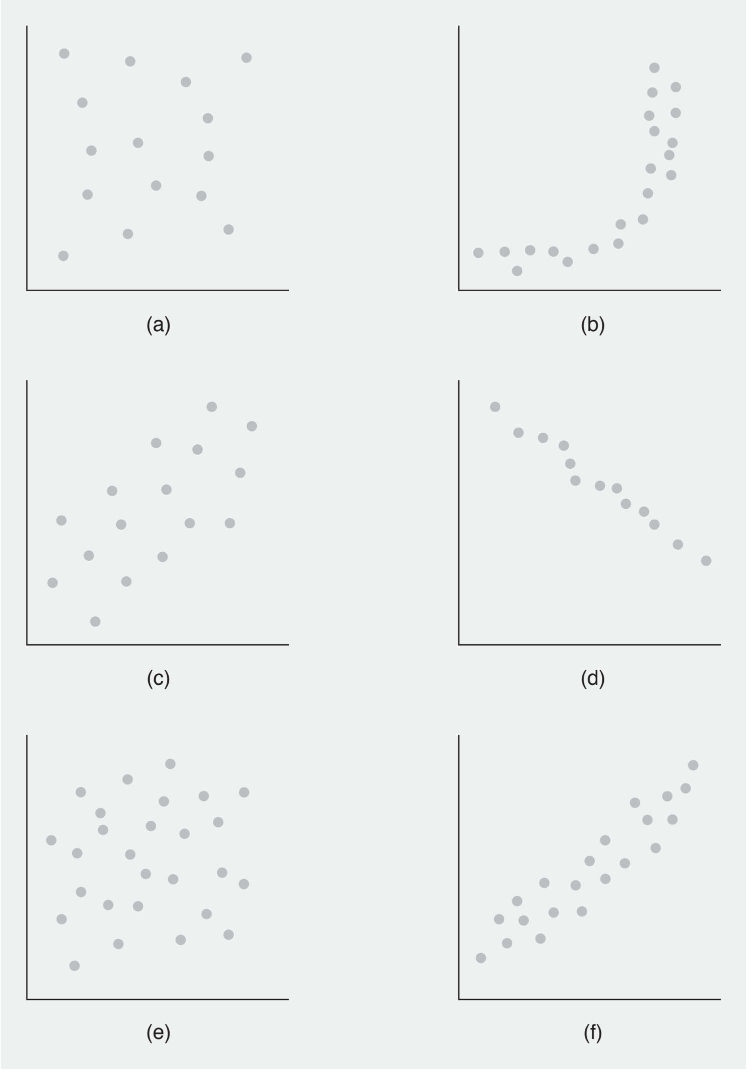
12. For each of the scatter diagrams on the following page, indicate whether the pattern is linear, curvilinear, or no correlation; if it is linear, indicate whether it is positive or negative and approximately how strong the correlation is (strong, moderate, small).

13. Make up a scatter diagram with 10 dots for each of the following situations: (a) perfect positive linear correlation, (b) strong but not perfect positive linear correlation, (c) weak positive linear correlation, (d) strong but not perfect negative linear correlation, (e) no correlation, (f) clear curvilinear correlation.

For problems 14 and 15, (a) make a scatter diagram of the raw scores; (b) describe in words the general pattern of correlation, if any; and (c) figure the correlation coefficient.

14. A researcher studying people in their 80s was interested in the relation between number of very close friends and overall health (on a scale from 0 to 100). The scores for six research participants are shown on the following page.

Research Participant	Number of Friends	Overall Health
A	2	41
B	4	72
C	0	37
D	3	84
E	2	52
F	1	49



(a)

(b)

(c)

(d)

(e)

(f)

15. In a study of people first getting acquainted with each other, researchers reasoned the amount of self-disclosure of one's partner (on a scale from 1 to 30) and one's liking for one's partner (on a scale from 1 to 10). The *Z* scores for the variables are given to save you some figuring. Here are the results:

Partner's Self-Disclosure		Liking for Partner	
Actual Score	*Z Score*	*Actual Score*	*Z Score*
18	.37	8	1.10
17	.17	9	1.47
20	.80	6	.37
8	−1.72	1	−1.47
13	−.67	7	.74
24	1.63	1	−1.47
11	−1.09	3	−.74
12	−.88	5	.0
18	.38	7	.74
21	1.00	3	−.74

For problems 16 and 17, (a) make a scatter diagram of the raw scores; (b) describe in words the general pattern of correlation, if any; (c) figure the correlation coefficient; (d) explain the logic of what you have done, writing as if you are speaking to someone who has never had a statistics course (but who does understand the mean, standard deviation, and *Z* scores); (e) give three logically possible directions of causality, saying for each whether it is a reasonable direction in light of the variables involved (and why); (f) make raw score predictions on the criterion variable for persons with *Z* scores on the predictor variable of $-2, -1, 0, +1, +2$, and (g) give the proportion of variance accounted for (r^2).

16. Four research participants take a test of manual dexterity (high scores mean better dexterity) and an anxiety test (high scores mean more anxiety). (For part (f), assume that dexterity is the predictor variable.) The scores are as follows:

Person	Dexterity	Anxiety
A	1	10
B	1	8
C	2	4
D	4	−2

17. Five university students were asked about how important a goal it is to them to have a family and about how important a goal it is for them to be highly successful in their work. Each variable was measured on a scale from 0 "Not at all important goal" to 10 "Very important goal." (For part (f), assume that the family goal is the predictor variable.) The scores are as follows:

Student	Family Goal	Work Goal
A	7	5
B	6	4
C	8	2
D	3	9
E	4	1

18. As part of a larger study, Speed and Gangestad (1997) colleced ratings and nominations on a number of characteristics for 66 fraternity men from their

fellow fraternity members. The following paragraph is taken from their "Results" section:

> . . . men's romantic popularity significantly correlated with several characteristics: best dressed ($r = .48$), most physically attractive ($r = .47$), most outgoing ($r = .47$), most self-confident ($r = .44$), best trendsetters ($r = .38$), funniest ($r = .37$), most satisfied ($r = .32$), and most independent ($r = .28$). Unexpectedly, however, men's potential for financial success did not significantly correlate with romantic popularity ($r = .10$). (p. 931)

Explain these results as if you were writing to a person who has never had a course in statistics. Specifically, (a) explain what is meant by a correlation coefficient using one of the correlations as an example; (b) explain in a general way what is meant by "significantly" and "not significantly," referring to at least one specific example; and (c) speculate on the meaning of the pattern of results, taking into account the issue of direction of causality.

19. Gable and Lutz (2000) studied 65 children, 3 to 10 years old, and their parents. One of their results was: "Parental control of child eating showed a negative association with children's participation in extracurricular activities ($r = .34$; $p < .01$)" (p. 296). Another result was: "Parents who held less appropriate beliefs about children's nutrition reported that their children watched more hours of television per day ($r = .36$; $p < .01$)" (p. 296).

    Explain these results as if you were writing to a person who has never had a course in statistics. Be sure to comment on possible directions of causality for each result.

20. Arbitrarily select eight people, each from a different page of a newspaper. Do each of the following: (a) make a scatter diagram for the relation between the number of letters in each person's first and last name, (b) figure the correlation coefficient for the relation between the number of letters in each person's first and last name, (c) describe the result in words, and (d) suggest a possible interpretation for your results.

21. A researcher studying adjustment to the job of new employees found a correlation of .30 between amount of employees' education and rating by job supervisors 2 months later. The researcher now plans to use amount of education to predict supervisors' later ratings of employees. Indicate the (a) predictor variable, (b) criterion variable, and (c) standardized regression coefficient (β). (d) Write the *Z*-score prediction model. Give the predicted *Z* scores for supervisor ratings for employees with amount of education *Z* scores of (e) -1.5, (f) -1, (g) $-.5$, (h) 0, (i) $+.5$, (j) $+1$, and (k) $+1.5$.

22. Ask five other students of the same gender as yourself (each from different families) to give you their own height and also their mother's height. Based on the numbers these five people give you, (a) figure the correlation coefficient, and (b) determine the *Z*-score prediction model for predicting a person's height from his or her mother's height. Finally, based on your prediction model, predict the height of a person of your gender whose mother's height is (c) 5 feet, (d) 5 feet 6 inches, and (e) 6 feet. (Note: Either convert inches to decimals of feet or do the whole problem using inches.)

23. ADVANCED TOPIC: Hahlweg, Fiegenbaum, Frank, Schroeder, and von Witzleben (2001) carried out a study of a treatment method for agoraphobia, a condition that affects about 4% of the population and involves unpredictable panic attacks in public spaces such as shopping malls, buses, or movie theaters. Table 12 shows the correlation coefficient (*r*s) and standardized regression coefficient (βs) for four variables predicting the effectiveness of the treatment. The actual criterion variable

Table 12 Multiple Regression Analysis Predicting Average Intragroup Effect Size at Postassessment

Independent Variable	*r*	β
BDI	.30***	.30***
Age	−.21***	−.20**
No. of sessions	.12*	.08
Duration of disorder	−.13*	−.02

Note: $R = .36$; $R^2 = .13$. BDI = Beck Depression Inventory.
*$p < .05$. **$p < .01$. ***$p < .000$.
Source: Hahlweg. K., Fiegenbaum, W., Frank, M., Schroeder, B., & von Witzleben, I. (2001). Short- and long-term effectiveness of an empirically supported treatment of agoraphobia. *Journal of Consulting and Clinical Psychology, 69,* 375–382. Copyright © 2001 by the American Psychological Association. Reproduced with permission. The use of APA information does not imply endorsement by APA.

is labeled "Average Intragroup Effect Size at Postassessment." The article explains that this is each patient's change from before to after treatment, averaged across several measures of mental health. A higher intragroup effect size score indicates a greater improvement in mental health. Explain the results of this study as if you were writing to a person who understands correlation but has never had any exposure to prediction or multiple regression analysis.

24. ADVANCED TOPIC: Based on Table 12 from problem 23, (a) write out the *Z*-score multiple regression formula for predicting average intragroup effect size at postassessment, and (b) figure the predicted *Z* score for average intragroup effect size at postassessment for persons A through J, whose *Z* scores on each predictor variable are shown below.

	Predictor Variable			
Person	*BDI*	*Age*	*No. of sessions*	*Duration of disorder*
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	0	0	0	1
E	0	0	0	0
F	0	0	0	0
G	1	1	1	0
H	0	0	1	1
I	1	1	1	1
J	−1	−1	−1	−1

The 🖱 in the steps below indicates a mouse click. (We used SPSS version 17.0 for Windows to carry out these analyses. The steps and output may be slightly different for other versions of SPSS.)

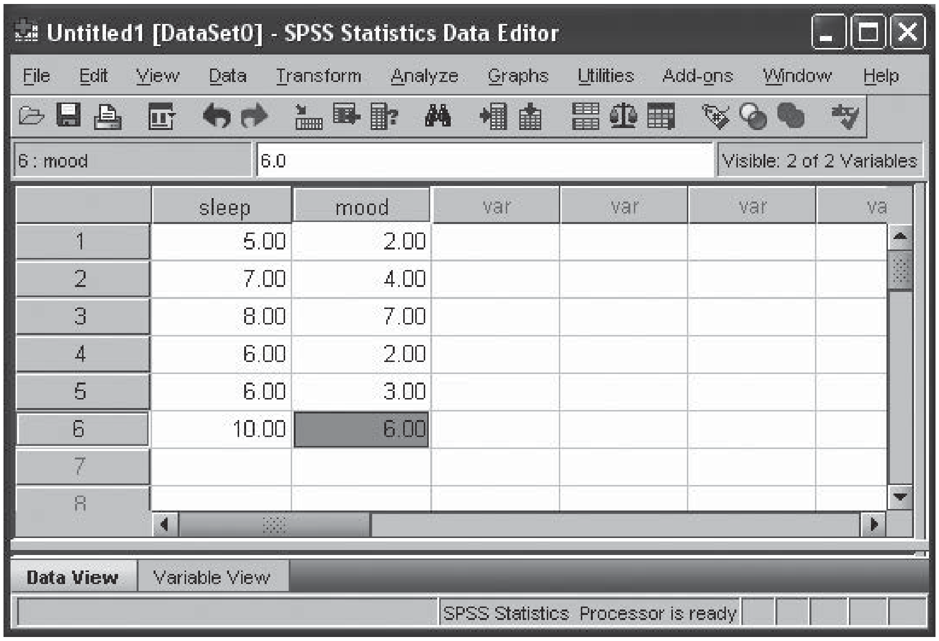
In the steps below for the scatter diagram, correlation coefficient, and prediction, we will use the example of the sleep and happy mood study. The scores for that study are shown in Table 1, the scatter diagram is shown in Figure 2, and the figuring for the correlation coefficient is shown in Table 2.

Creating a Scatter Diagram

❶ Enter the scores into SPSS. Enter the scores as shown in Figure 17.
❷ ✐ *Graphs*, ✐ *Legacy Dialogs*.
❸ ✐ *Scatter/Dot.* A box will appear that allows you to select different types of scatter diagrams (or *scatterplots*, as SPSS calls them). You want the "Simple scatter" diagram. This is selected as the default type of scatter diagram, so you just need to ✐ *Define.*
❹ ✐ the variable called "mood" and then ✐ the arrow next to the box labeled "Y axis." This tells SPSS that the scores for the "mood" variable should go on the vertical (or Y) axis of the scatter diagram. ✐ the variable called "sleep" and then ✐ the arrow next to the box labeled "X axis." This tells SPSS that the scores for the "sleep" variable should go on the horizontal (or X) axis of the scatter diagram.
❺ ✐ *OK.* Your SPSS output window should look like Figure 18.

Finding the Correlation Coefficient

❶ Enter the scores into SPSS. Enter the scores as shown in Figure 17.
❷ ✐ *Analyze.*
❸ ✐ *Correlate.*
❹ ✐ *Bivariate.*
❺ ✐ on the variable called "mood" and then ✐ the arrow next to the box labeled "Variables." ✐ on the variable called "sleep" and then ✐ the arrow next to the box labeled "Variables." This tells SPSS to figure the correlation between the "mood" and "sleep" variables. (If you wanted to find the correlation between each of several variables, you would put all of them into the "Variables" box.) Notice that by default SPSS will carry out a Pearson correlation (the type of correlation you have learned in this chapter), will automatically give the significance level (using a two-tailed test), and will flag statistically significant correlations using the .05 significance level.
❻ ✐ *OK.* Your SPSS output window should look like Figure 19.



6 : mood | 6.0 | Visible: 2 of 2 Variables

	sleep	mood	var	var	var	va
1	5.00	2.00				
2	7.00	4.00				
3	8.00	7.00				
4	6.00	2.00				
5	6.00	3.00				
6	10.00	6.00				
7						
8						

SPSS Statistics Processor is ready

Figure 17 SPSS data editor window for the fictional study of the relationship between hours slept last night and mood.

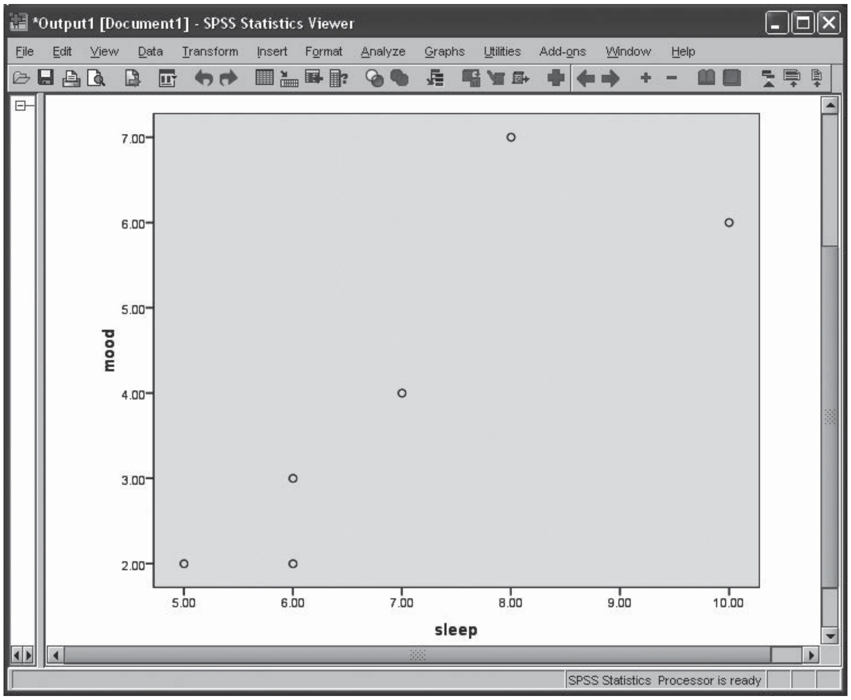
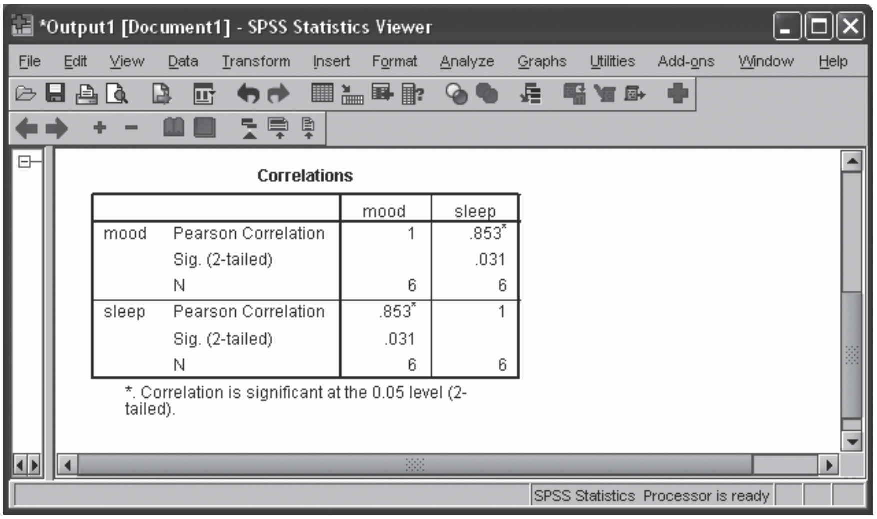


Figure 18 An SPSS scatter diagram showing the relationship between hours slept last night and mood (fictional data).


*Output1 [Document1] - SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help
SPSS Statistics Processor is ready

Correlations

		mood	sleep
mood	Pearson Correlation	1	.853*
	Sig. (2-tailed)		.031
	N	6	6
sleep	Pearson Correlation	.853*	1
	Sig. (2-tailed)	.031	
	N	6	6

*. Correlation is significant at the 0.05 level (2-tailed).

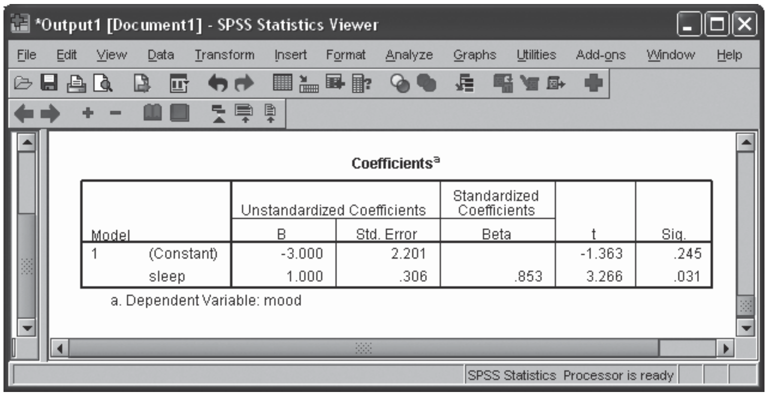
Figure 19 SPSS output window for the correlation between hours slept last night and mood (fictional data).

The table shown in Figure 19 is a small correlation matrix (there are only two variables). (If you were interested in the correlations among more than two variables—which is often the case in behavioral and social sciences research—SPSS would produce a larger correlation matrix.) The correlation matrix shows the correlation coefficient ("Pearson Correlation"), the exact significance level of the correlation coefficient [“Sig. (2-tailed)”], and the number of people in the correlation analysis (“N”). Note that two of the cells of the correlation matrix show a correlation coefficient of exactly 1. You can ignore these cells, as they simply show that each variable is perfectly correlated with itself. (In larger correlation matrixes all of the cells on the diagonal from the top left to the bottom right of the table will have a correlation coefficient of 1.) You will also notice that the remaining two cells provide identical information. This is because the table shows the correlation between sleep and mood and also between mood and sleep (which are, of course, identical correlations). So you can look at either one. (In a larger correlation matrix, you need only look either at all of the correlations above the diagonal that goes from top left to bottom right or at all of the correlations below that diagonal.) The correlation coefficient is .853 (which is usually rounded to two decimal places in research articles). The significance level of .031 is less than the usual .05 cutoff, which means that it is a statistically significant correlation. The asterisk (*) by the correlation of .853 also shows that it is statistically significant (at the .05 significance level, as shown by the note under the table).

Prediction with a Single Predictor Variable

In actual research, prediction is most often done using raw scores (as opposed to *Z* scores). However, as you will learn below, even if you do prediction with raw scores, the output provided by SPSS allows you to determine the *Z* score prediction model (or prediction rule).

❶ Enter the scores into SPSS. Enter the scores as shown in Figure 17.
❷ ✐ *Analyze.*
❸ ✐ *Regression.*
❹ ✐ *Linear.* This tells SPSS that you are figuring a *linear* prediction rule (as opposed to any one of a number of other prediction rules).
❺ ✐ the variable called “mood” and then ✐ the arrow next to the box labeled “Dependent.” This tells SPSS that the “mood” variable is the criterion variable (which is also called the *dependent variable* in prediction, because it “depends” on the predictor variable’s score). ✐ the variable called “sleep” and then ✐ the arrow next to the box labeled “Independent(s).” This tells SPSS that the “sleep” variable is the predictor variable (which is also called the *independent variable* in prediction).
❻ ✐ *OK.* The final table in your SPSS output window should look like Figure 20.


*Output1 [Document1] - SPSS Statistics Viewer
File
Edit
View
Data
Transform
Insert
Format
Analyze
Graphs
Utilities
Add-ons
Window
Help
SPSS Statistics Processor is ready

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-3.000	2.201		-1.363	.245
	sleep	1.000	.306	.853	3.266	.031

a. Dependent Variable: mood

Figure 20 SPSS output window for predicting mood from hours slept last night (fictional data).

SPSS provided four tables in the output. For our purposes here, we focus only on the final table (shown in Figure 20), which gives the information for the prediction model. Most important for us, the table gives the standardized regression coefficient (labeled "Beta"), which is .853. This tells us that the Z score prediction model for predicting mood from the number of hours slept is: $Z_{\text{mood}} = (.853)(Z_{\text{hours slept}})$. (Notice this is the same as the correlation coefficient.)

Advanced Topic: Multiple Regression

If you were conducting a multiple regression, you would put all of the predictor variables in the "Independent(s)" box in Step ❺ above. In the SPSS output, the standardized regression coefficient for each predictor would be listed in the "Beta" column. Examination of these standardized regression coefficients would tell you the unique influence that each predictor has for predicting the criterion variable. (As you learned earlier in the chapter, the larger the standardized regression coefficient for a predictor variable, the more influence that variable has when predicting a score on the criterion variable.)

Appendix: Hypothesis Tests and Power for the Correlation Coefficient

Significance of a Correlation Coefficient

Hypothesis testing of a correlation coefficient follows the usual steps of hypothesis testing. However, there are five important points to note.

1. Usually, the null hypothesis is that the correlation in a population like that studied is no different from a population in which the true correlation is 0.
2. If the data meet assumptions (explained below), the comparison distribution is a *t* distribution with degrees of freedom equal to the number of people minus 2 (that is, $df = N - 2$).
3. You figure the correlation coefficient's score on that *t* distribution using the formula

The *t* score is the correlation coefficient multiplied by the square root of 2 less than the number of people in the study, divided by the square root of 1 minus the correlation coefficient squared.

$$t = \frac{(r)(\sqrt{N - 2})}{\sqrt{1 - r^2}} \qquad (5)$$

4. Note that significance tests of a correlation, like a *t* test, can be either one-tailed or two-tailed. A one-tailed test means that the researcher has predicted the sign (positive or negative) of the correlation.
5. Assumptions for the significance test of a correlation coefficient are that (a) the populations for both variables are normally distributed, and (b) in the population, the distribution of each variable at each point of the other variable has about equal variance. However, as with the *t* test and analysis of variance, moderate violations of these assumptions are not fatal.

An Example

Here is an example using the sleep and mood study example. Let's suppose that the researchers predicted a positive correlation between sleep and mood, to be tested at the .05 level.

❶ **Restate the question as a research hypothesis and a null hypothesis about the populations.** There are two populations:

Population 1: People like those in this study.

Population 2: People for whom there is no correlation between number of hours slept the night before and mood the next day.

The null hypothesis is that the two populations have the same correlation. The research hypothesis is that Population 1 has a higher correlation than Population 2. (That is, the prediction is for a population correlation greater than 0.)

❷ **Determine the characteristics of the comparison distribution.** Assuming we meet the assumptions (in this example, it would be hard to tell with only six people in the study), the comparison distribution is a *t* distribution with $df = 4$. (That is, $df = N - 2 = 6 - 2 = 4$.)

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** The *t* table shows that for a one-tailed test at the .05 level, with 4 degrees of freedom, you need a *t* of at least 2.132.

❹ **Determine your sample's score on the comparison distribution.** We figured a correlation of $r = .85$. Applying the formula to find the equivalent *t*, we get

$$t = \frac{(r)\sqrt{N - 2}}{\sqrt{1 - r^2}} = \frac{(.85)\sqrt{6 - 2}}{\sqrt{1 - .85^2}} = \frac{(.85)(2)}{.53} = 3.21.$$

❺ **Decide whether to reject the null hypothesis.** The *t* score of 3.21 for our sample correlation is more extreme than the minimum needed *t* score of 2.132. Thus, you can reject the null hypothesis, and the research hypothesis is supported.

Effect Size and Power

The correlation coefficient itself is a measure of effect size. (Thus, in the example, effect size is $r = .85$.) Cohen's (1988) conventions for the correlation coefficient are .10 for a small effect size, .30 for a medium effect size, and .50 for a large effect size. You can find the power for a correlation using a power table, a power software package, or an Internet power calculator. Table 13 gives the approximate power, and Table 14 gives minimum sample size for 80% power at the .05 level of significance. (More complete tables are provided in Cohen, 1988, pp. 84–95, 101–102.) For

Table 13 Approximate Power of Studies Using the Correlation Coefficient (*r*) for Testing Hypotheses at the .05 Level of Significance

			Effect Size	
		Small ($r = .10$)	Medium ($r = .30$)	Large ($r = .50$)
Two-tailed Total *N*:	10	.06	.13	.33
	20	.07	.25	.64
	30	.08	.37	.83
	40	.09	.48	.92
	50	.11	.57	.97
	100	.17	.86	*
One-tailed Total *N*:	10	.08	.22	.46
	20	.11	.37	.75
	30	.13	.50	.90
	40	.15	.60	.96
	50	.17	.69	.98
	100	.26	.92	*

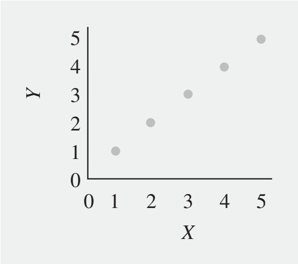
*Power is nearly 1.

Table 14 Approximate Number of Participants Needed for 80% Power for a Study Using the Correlation Coefficient (r) for Testing a Hypothesis at the .05 Significance Level

	Effect Size		
	Small (r = .10)	**Medium (r = .30)**	**Large (r = .50)**
Two-tailed	783	85	28
One-tailed	617	68	22

example, the power for a study with an expected medium effect size ($r = .30$), two-tailed, with 50 participants, is .57 (which is below the standard desired level of at least .80 power). This means that even if the research hypothesis is in fact true and has a medium effect size (that is, even if the two variables are correlated at $r = .30$ in the population), there is only a 57% chance that the study will produce a significant correlation.

1. (a) Curvilinear; (b) linear, positive, strong; (c) linear, negative, strong; (d) linear, positive, strong; (e) linear, positive, small to moderate; (f) no correlation.
2. (a) Data Set A:



X		Y		Cross-Product of Z Scores
Raw	Z	Raw	Z	
1	−1.41	1	−1.41	2.0
2	−.71	2	−.71	.5
3	.00	3	.00	0
4	.71	4	.71	.5
5	1.41	5	1.41	2.0
$M = 3$; $SD = 1.41$				$\overline{5.0}$
				$r = 5.0/5 = 1.00$

Data Set B: $r = 4.5/5 = .90$; Data Set C: $r = -3.0/5 = -.60$; Data Set D: $r = 3.0/5 = .60$.

(b) In Data Set A, all of the pairs of X and Y scores are identical and thus there is a perfect positive correlation between the scores. In Data Set B, the pairs of scores are not all identical, although they are very close to each other and low scores tend to go with low, medium with medium, and high with high. Thus, it makes sense that there is a strong positive correlation ($r = .90$) for Data Set B. For Data Set C, the general pattern is for low scores to go with high scores, medium scores to go with medium, and high scores to go with low scores. However, the pattern is not perfect (for example, the X score of 2 is paired with a Y score of 2), and thus the correlation of $r = -.60$ makes sense. For Data Set D, the general pattern is for low scores to go with low, medium scores with medium, and high scores with high. However, the pattern is not perfect (for example, the X score of 2 is paired with a Y score of 4), and thus the correlation of $r = .60$ makes sense.

3. Possibility A:

Take Drug		Get Cold		Cross-Product of Z Scores
Raw	**Z**	**Raw**	**Z**	
0	−1	1	1	−1
0	−1	1	1	−1
0	−1	1	1	−1
0	−1	1	1	−1
1	1	0	−1	−1
1	1	0	−1	−1
1	1	0	−1	−1
1	1	0	−1	$\underline{-1}$
				−8
				$r = -8/8 = -1.00$

Possibility B: $r = -4/8 = -.50$; Possibility C: $r = 0/8 = .00$; Possibility D: $r = -6.2/8 = -.78$.

4. (a) See answer to 2 above for an example; (b) linear, negative, moderate. (c)

Approximate Age (Years)		**Number of People Stopping to Look**		**Cross-Product of *Z* Scores**
X	Z_X	***Y***	Z_Y	Z_XZ_Y
480	1.39	68	−.69	−.96
530	1.71	71	−.59	−1.01
255	−.09	123	1.19	−.11
122	−.96	112	.82	−.79
391	.80	48	−1.38	−1.10
370	.67	84	−.14	−.09
155	−.74	66	−.76	.56
130	−.91	148	2.05	−1.87
137	−.86	71	−.59	.51
114	−1.01	91	.10	−.10
				−4.94
				r = −4.94/10 = −.49

5. (a) See answer to 2 above for an example; (b) linear, no correlation (or very small positive correlation).
(c)

Neatness Rating		**Achievement Test**		**Cross-Product of *Z* Scores**
X	Z_X	***Y***	Z_Y	Z_XZ_Y
18	−.52	60	−.66	.34
24	1.43	58	−1.09	−1.56
14	−1.82	70	1.47	−2.68
19	−.20	58	−1.09	.22
20	.13	66	.62	.08
23	1.11	68	1.04	1.15
20	.13	65	.40	.05
22	.78	68	1.04	.81
15	−1.50	56	−1.51	2.27
21	.46	62	−1.23	−.11
				.57
				r = .57/10 = .06

6. (a) See answer to 2 above for an example; (b) linear, positive, strong.
(c)

Hours of Violent TV per Week		**Number of Violent or Aggressive Acts**		**Cross-Product of *Z* Scores**
X	Z_X	***Y***	Z_Y	Z_XZ_Y
14	1.27	9	.97	1.23
8	−.63	6	.00	.00
6	−1.27	1	−1.62	2.06
12	.63	8	.65	.41
Σ = 40		Σ = 24		Σ = 3.70
M = 10		M = 6		
SD = 3.16		SD = 3.08		SD = 3.70/4 = .93

(d) The first thing I did was make a graph, called a scatter diagram, putting one variable on each axis, then putting a dot where each person's pair of scores goes on that graph. This gives a picture of the pattern of relationship between the two variables. In this example, high scores generally go with high scores and lows with lows. The scores going together in a systematic pattern makes this a *correlation;* that highs go with highs and lows with lows makes this correlation *positive;* that dots fall in a roughly straight line pattern makes this positive correlation *linear;* the dots fall very close to a straight line, which makes this a *strong* positive linear correlation.

Next, I figured the *correlation coefficient,* a number describing the degree of linear correlation between weekly viewing of violent TV and the number of violent or aggressive acts toward playmates (in a positive correlation, how consistently highs go with highs and lows with lows). To do this, I changed all the scores to *Z* scores because *Z* scores tell you how much a score is low or high relative to the other scores in its distribution. You figure the correlation coefficient by multiplying each person's two *Z* scores by each other, totaling up these products, then averaging this total over the number of people. This will be a high number if highs go with highs and lows with lows, because with *Z* scores, highs are always positive and positive times positive is positive, and with *Z* scores, lows are always negative and negative times negative becomes positive too. Following this procedure, the highest number you can get, if the scores for the two variables are perfectly correlated, is +1. If there were no linear correlation between the two variables, the results would be 0 (because highs would sometimes be multiplied by highs and sometimes by lows, giving a mixture of positive and negative products that would cancel out).

In this example, the products of the *Z* scores add up to 3.70, which when divided by the number of children is .93. This is called a *Pearson correlation coefficient* (r) of .93 and indicates a strong, positive linear correlation between the hours of violent TV watched each week and the number of violent or aggressive acts toward playmates.

(e) Three logically possible directions of causality: (i) Watching violent TV makes children act more aggressively toward playmates; (ii) being aggressive makes children more interested in watching violent TV; (iii) a third factor—such as living in a violent family environment—makes children more interested in watching violent TV and also makes them act aggressively toward playmates.

(f) Formulas: Predicted $Z_Y = (\beta)(Z_X)$; Predicted $Y = (SD_Y)(\text{Predicted } Z_Y) + M_Y$;

$Z_X = -2$: Predicted $Z_Y = (.93)(-2) = -1.86$; Predicted $Y = (3.08)(-1.86) + 6 = .27$.

$Z_X = -1$: Predicted $Z_Y = (.93)(-1) = -.93$; Predicted $Y = (3.08)(-.93) + 6 = 3.14$.

$Z_X = 0$: Predicted $Z_Y = (.93)(0) = 0$; Predicted $Y = (3.08)(0) + 6 = 6.00$.

$Z_X = +1$: Predicted $Z_Y = (.93)(1) = .93$; Predicted $Y = (3.08)(.93) + 6 = 8.86$.

$Z_X = +2$: Predicted $Z_Y = (.93)(2) = 1.86$; Predicted $Y = (3.08)(1.86) + 6 = 11.73$.

(g) $r^2 = .93^2 = .86$.

7. (a) See answer to 2 above for an example; (b) linear, positive, strong.
   (c)

Number of Town-Hall Meetings		Percentage of People Who Can Name Candidate		Cross-Product of *Z* Scores
X	Z_X	Y	Z_Y	Z_XZ_Y
4	.63	70	.36	.23
5	1.26	94	1.45	1.83
2	−.63	36	−1.17	.74
1	−1.26	48	−.63	.79
$\Sigma = 12$		$\Sigma = 248$		$\Sigma = 3.59$
$M = 3$		$M = 62$		
$SD = 1.58$		$SD = 22.14$		$r = 3.59/4 = .90$

   (d) See 6d above.
   (e) Three logically possible directions of causality: (i) If a candidate has more town-hall meetings, this makes more people become aware of the candidate; (ii) If a candidate is well known, this causes the candidate to have more town-hall meetings; (iii) A third factor—such as the amount of campaign money the candidate has—means the candidate can afford to hold more town-hall meetings and also causes more people to be aware of the candidate (perhaps because the candidate has been able to afford a large advertising campaign).
   (f) Formulas: Predicted $Z_Y = (\beta)(Z_X)$; Predicted $Y = (SD_Y)(\text{Predicted } Z_Y) + M_Y$;
   $Z_X = -2$: Predicted $Z_Y = (.90)(-2) = -1.80$; Predicted $Y = (22.14)(-1.80) + 62 = 22.15$.
   $Z_X = -1$: Predicted $Z_Y = (.90)(-1) = -.90$; Predicted $Y = (22.14)(-.90) + 62 = 42.07$.
   $Z_X = 0$: Predicted $Z_Y = (.90)(0) = 0$; Predicted $Y = (22.14)(0) + 62 = 62.00$.
   $Z_X = +1$: Predicted $Z_Y = (.90)(1) = .90$; Predicted $Y = (22.14)(.90) + 62 = 81.93$.
   $Z_X = +2$: Predicted $Z_Y = (.90)(2) = 1.80$; Predicted $Y = (22.14)(1.80) + 62 = 101.85$.
   (g) $r^2 = .90^2 = .81$.
8. (a) This table shows the degree of association among scores on several measures given to pregnant women and their partners. (Here continue with an explanation of the correlation coefficient like that in 6d above except in this problem you also need to explain the mean, standard deviation, and *Z* scores.) For example, the correlation of .17 between women's reports of stress and men's reports of stress indicates that the association between these two measures is quite weak. That is, how much stress a woman is under is not highly related to how much stress her partner believes she is under. On the other hand, the correlation of .50 (near the middle of the first column of correlations) tells you that there is a much stronger association between a woman's report of stress and her depressed mood in the second interview. That is, women who report being under stress are also likely to report being depressed, those reporting being under not much stress are likely to report not being very depressed.

   (b) In general, the correlations shown in this table are strongest among the stress, support, and mood items; correlations of these variables with demographics (age, ethnicity, etc.) were fairly weak. Partner support seemed to be strongly correlated with stress and mood, and depressed mood at the second testing was particularly related to the other variables.
   (c) Just because two variables are correlated, even strongly correlated, does not mean that you can know the particular direction of causality that creates that association. For example, there is a strong negative correlation between partner support at time 1 and depressed mood at time 2. There are three logically possible directions of causality here: (i) Support can be causing lower depression, (ii) lower depression can be causing support, or (iii) some third factor can be causing both. You can rule out the second possibility, since something in the future (low depression) can't cause the past (initial support). However, the other two possibilities remain. It is certainly plausible that having her partner's support helps reduce depression. But it is also possible that a third factor is causing both. For example, consider level of income. Perhaps when a couple has more income, the partner has more time and energy to provide support and the greater comfort of living keeps depression down.
9. (a) Score on knowledge of fitness training principles; (b) number of injuries over subsequent year; (c) .4; (d) Predicted $Z_{\text{Injuries}} = (.4)(Z_{\text{Score}})$; (e) $(.4)(-2) = -.8$; (f) −.4; (g) 0; (h) .4; (i) .8.
10. This study used a statistical procedure called multiple regression. This procedure produces a formula for predicting a person's score on a criterion variable (in this example, third-graders' reading comprehension) from his or her scores on a set of predictor variables (in this example, the three specific measures of reading ability). The formula is of the form that you multiply the person's score on each of the predictor variables by some particular number, called a regression coefficient (or beta), and then add up the products. The procedure produces the most accurate prediction rule of this kind.

    In this example, the prediction rule for the *Z* score for Reading Comprehension is −.227 multiplied by the *Z* score for Letter-Word Identification, plus .299 multiplied by the *Z* score for Word Attack, plus .671 multiplied by the *Z* score for Oral Reading Fluency. (These are the numbers in the table next to each predictor variable in the Beta column.)

    These regression coefficients suggest that reading comprehension is most strongly related to Oral Reading Fluency. Reading comprehension is also somewhat positively related to Word Attack. However, in the context of this prediction equation, reading comprehension is somewhat negatively related to Letter-Word Identification. This means that for any given level of Oral Reading Fluency and Word Attack, the better the child is at Letter-Word Identification, the child will be somewhat *worse* at reading comprehension!

    It is important to note, however, that the regression coefficients for each of these predictors reflect what the scores on each predictor contribute to the prediction, over and above what the others contribute. If we were to consider ordinary correlations between each of the predictor variables with the criterion variable, their relative importance could be quite different. (Those correlations, however, were not provided.)

    Another important piece of information in this table is R^2. This number tells you the proportion of variance accounted for in the criterion variable by the three predictor variables taken together. That is, 53.4% of the variation in the third-graders' reading comprehension is accounted for by these three measures of specific reading abilities. This is equivalent to a correlation between reading comprehension and these three predictor variables of .73 (the square root of .534).

11. (a) Predicted $Z_{\text{Comprehension}} = (-.227)(Z_{\text{Identification}}) + (.299)\ (Z_{\text{Attack}}) + (.671)(Z_{\text{Fluency}})$; (b) A: Predicted $Z_{\text{Comprehension}} = (-.227)(1) + (.299)(1) + (.671)(1) = .743$; B: Predicted $Z_{\text{Comprehension}} = 0$; C: Predicted $Z_{\text{Comprehension}} = -.743$; D: Predicted $Z_{\text{Comprehension}} = -.227$; E: Predicted $Z_{\text{Comprehension}} = .299$; F: Predicted $Z_{\text{Comprehension}} = .671$; G: Predicted $Z_{\text{Comprehension}} = .289$; H: Predicted $Z_{\text{Comprehension}} = 1.341$; I: Predicted $Z_{\text{Comprehension}} = 1.631$.