7 The Correlation Coefficient

GETTING STARTED

To understand this chapter, recall the following:

- From Chapter 2, that in a relationship, particular *Y* scores tend to occur with a particular *X*, a more consistent relationship is "stronger," and we can use someone's *X* to predict what his/her *Y* will be.
- From Chapter 5, that greater variability indicates a greater variety of scores is present and so greater variability produces a weaker relationship. Also that the phrase "accounting for variance" refers to accurately predicting *Y* scores.

Your goals in this chapter are to learn

- The logic of correlational research and how it is interpreted.
- How to read and interpret a *scatterplot* and a *regression line.*
- How to identify the *type* and *strength* of a relationship.
- How to interpret a *correlation coefficient.*
- When to use the *Pearson r* and the *Spearman* r_S.
- The logic of inferring a population correlation based on a sample correlation.

Recall that in research we want to not only demonstrate a relationship but also describe and summarize the relationship. The one remaining type of descriptive statistic for us to discuss is used to summarize relationships, and it is called the *correlation coefficient.* In the following sections, we'll consider when these statistics are used and what they tell us. Then we'll see how to compute the two most common versions of the correlation coefficient. First, though, a few more symbols.

NEW STATISTICAL NOTATION

Correlational analysis requires scores from two variables. Then, *X* stands for the scores on one variable, and *Y* stands for the scores on the other variable. Usually each pair of *X*–*Y* scores is from the same participant. If not, there must be a rational system for pairing the scores (for example, pairing the scores of roommates). Obviously we must have the same number of *X* and *Y* scores.

We use the same conventions for *Y* that we've previously used for *X*. Thus, ΣY is the sum of the *Y* scores, ΣY^2 is the sum of the squared *Y* scores, and $(\Sigma Y)^2$ is the squared sum of the *Y* scores.

You will also encounter three other notations. First, $(\Sigma X)(\Sigma Y)$ indicates to first find the sum of the *X*s and the sum of the *Y*s and then multiply the two sums together. Second, ΣXY, called the sum of the cross products, says to first multiply each *X* score in a pair times its corresponding *Y* score and then sum all of the resulting products.

> ***REMEMBER*** $(\Sigma X)(\Sigma Y)$ says to multiply the sum of *X* times the sum of *Y*. ΣXY says to multiply each *X* times its paired *Y* and then sum the products.

Finally, *D* stands for the numerical *difference* between the *X* and *Y* scores in a pair, which you find by subtracting one from the other.

Now, on to the correlation coefficient.

WHY IS IT IMPORTANT TO KNOW ABOUT CORRELATION COEFFICIENTS?

Recall that a relationship is present when, as the *X* scores increase, the corresponding *Y* scores change in a consistent fashion. Whenever we find a relationship, we then want to know its characteristics: What pattern is formed, how consistently do the scores change together, and what direction do the scores change? The best—and easiest—way to answer these questions is to compute a correlation coefficient. The **correlation coefficient** is the descriptive statistic that, in a single number, summarizes and describes the important characteristics of a relationship. The correlation coefficient *quantifies* the pattern in a relationship, examining *all X–Y* pairs at once. No other statistic does this. Thus, the correlation coefficient is important because it simplifies a complex relationship involving many scores into one, easily interpreted statistic. Therefore, in any research where a relationship is found, always calculate the appropriate correlation coefficient.

As a starting point, the correlation coefficients discussed in this chapter are most commonly associated with correlational research.

UNDERSTANDING CORRELATIONAL RESEARCH

Recall that a common research design is the correlational study. The term *correlation* is synonymous with *relationship,* so in a correlational design we examine the relationship between variables. (Think of *correlation* as meaning the shared, or “co,” relationship between the variables.) The relationship can involve scores from virtually any variable, regardless of how we obtain them. Often we use a questionnaire or observe participants, but we may also measure scores using any of the methods used in experiments.

Recall that correlational studies differ from experiments in terms of *how* we demonstrate the relationship. For example, say that we hypothesize that as people drink more coffee they become more nervous. To demonstrate this in an experiment, we might assign some people to a condition in which they drink 1 cup of coffee, assign others to a 2-cup condition and assign still others to a 3-cup condition. Then we would measure participants’ nervousness and see if more nervousness is related to more coffee. Notice that, by creating the conditions, we (the researchers) determine each participant’s *X* score because we decide whether their “score” will be 1, 2, or 3 cups on the coffee variable.

In a correlational design, however, we do *not* manipulate any variables, so we do not determine participants’ *X* scores. Rather, the scores on both variables reflect an amount or category of a variable that a participant has *already* experienced. Therefore, we simply measure the two variables and describe the relationship that is present. Thus, we might ask participants the amount of coffee they have consumed today and measure how nervous they are.

Recognize that computing a correlation coefficient does not create a correlational *design:* It is the absence of manipulation that creates the design. In fact, in later chapters we will compute correlation coefficients in experiments. However, correlation coefficients are most often used as the primary descriptive statistic in correlational research, and you must be careful when interpreting the results of such a design.

Drawing Conclusions from Correlational Research

People often mistakenly think that a correlation automatically indicates causality. However, recall from Chapter 2 that the existence of a relationship does not necessarily indicate that changes in *X cause* the changes in *Y*. A relationship—a *correlation*—can exist, even though one variable does not cause or influence the other. Two requirements must be met to confidently conclude that *X* causes *Y*.

First, *X must occur before Y*. However, in correlational research, we do not always know which factor occurred first. For example, if we simply measure the coffee drinking and nervousness of some people after the fact, it may be that participants who were already more nervous *then* tended to drink more coffee. Therefore, maybe greater nervousness actually caused greater coffee consumption. In any correlational study, it is possible that *Y* causes *X*.

Second, *X must be the only variable that can influence Y*. But, in correlational research, we do little to control or eliminate other potentially causal variables. For example, in the coffee study, some participants may have had less sleep than others the night before testing. Perhaps the lack of sleep caused those people to be more nervous *and* to drink more coffee. In any correlational study, some other variable may cause both *X* and *Y* to change. (Researchers often refer to this as "the third variable problem.")

Thus, a correlation by itself does not indicate causality. You must also consider the research method used to demonstrate the relationship. In experiments we apply the independent variable *first,* and we control other potential causal variables, so experiments provide better evidence for identifying the causes of a behavior.

Unfortunately, this issue is often lost in the popular media, so be skeptical the next time some one uses *correlation* and *cause* together. The problem is that people often ignore that a relationship may be a meaningless coincidence. For example, here's a relationship: As the number of toilets in a neighborhood increases, the number of crimes committed in that neighborhood also increases. Should we conclude that indoor plumbing causes crime? Of course not! Crime tends to occur more frequently in the crowded neighborhoods of large cities. Coincidentally, there are more indoor toilets in such neighborhoods.

The problem is that it is easy to be trapped by more mysterious relationships. Here's a serious example: A particular neurological disease occurs more often in the colder, northern areas of the United States than in the warmer, southern areas. Do colder temperatures cause this disease? Maybe. But, for all the reasons given above, the mere existence of this relationship is not evidence of causality. The north also has fewer sunny days, burns more heating oil, and differs from the south in many other ways. One of these variables might be the cause, while coincidentally, colder temperatures are also present.

Thus, a correlational study is not used to infer a causal relationship. It is possible that changes in *X* might cause changes in *Y,* but we will have no convincing evidence of this. Instead, correlational research is used to simply describe how nature relates the variables, without identifying the cause.

> ***REMEMBER*** We should not infer causality from correlational designs, because *X* may cause *Y*, *Y* may cause *X*, or a third variable may cause both *X* and *Y*.

Distinguishing Characteristics of Correlational Analysis

There are four major differences between how we handle data in a correlational analysis versus in an experiment. First, back in our coffee experiment, we would examine the *mean* nervousness score (*Y*) for each condition of the amount of coffee consumed (*X*). With correlational data, however, we typically have a large range of different *X* scores: People would probably report many amounts of coffee beyond only 1, 2, or 3 cups. Comparing the mean nervousness scores for many groups would be very difficult. Therefore, in correlational procedures, we do not compute a mean *Y* score at each *X*. Instead, the correlation coefficient summarizes the *entire* relationship at once.

A second difference is that, because we examine all pairs of *X–Y* scores, correlational procedures involve *one* sample: *In correlational designs, N always stands for the number of pairs of scores in the data.*

Third, we will not use the terms *independent* and *dependent variable* with a correlational study (although some researchers argue that these terms are acceptable here). Part of our reason is that either variable may be called *X* or *Y*. How do we decide? Recall that in a relationship the *X* scores are the "given" scores. Thus, if we ask, "For a given amount of coffee, what are the nervousness scores?" then amount of coffee is *X*, and nervousness is *Y*. Conversely, if we ask, "For a given nervousness score, what is the amount of coffee consumed?" then nervousness is *X*, and amount of coffee is *Y*. Further, recall that, in a relationship, particular *Y* scores naturally occur at a particular *X*. Therefore, if we know someone's *X*, we can predict his or her corresponding *Y*. The procedures for doing this are described in the next chapter, where the *X* variable is called the *predictor variable,* and the *Y* variable is called the *criterion variable.* As you'll see, researchers used correlational techniques to identify *X* variables that are "good predictors" of *Y* scores.

Finally, as in the next section, we graph correlational data by creating a *scatterplot*.

Plotting Correlational Data: The Scatterplot

A **scatterplot** is a graph that shows the location of each data point formed by a pair of *X–Y* scores. Figure 7.1 contains the scores and resulting scatterplot showing the relationship between coffee consumption and nervousness. It shows that people drinking 1 cup have nervousness scores around 1 or 2, but those drinking 2 cups have higher nervousness scores around 2 or 3, and so on. Thus, we see that one batch of data points (and *Y* scores) tend to occur with one *X*, and a different batch of data points (and thus different *Y* scores) are at a different *X*.

Real research typically involves a larger *N* and the data points will not form such a clear pattern. In fact, notice the strange data point produced by $X = 3$ and $Y = 9$. A data point that is relatively far from the majority of data points in the scatterplot is referred to as an **outlier**—it *lies out* of the general pattern. Why an outlier occurs is usually a mystery to the researcher.

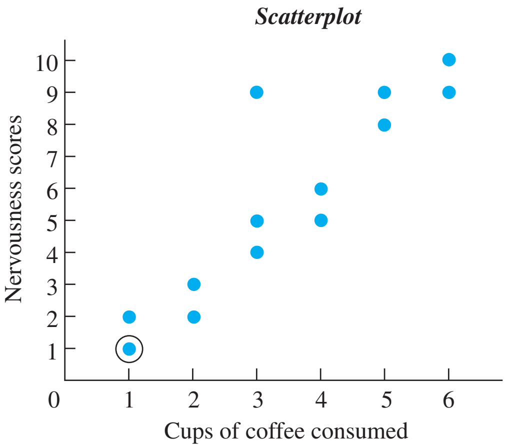
Notice that the scatterplot does summarize the data somewhat. In the table, two people had scores of 1 on coffee consumption and nervousness, but the scatterplot shows

FIGURE 7.1

Scatterplot showing nervousness as a function of coffee consumption

Each data point is created using a participant's coffee consumption as the X score and nervousness as the Y score.

Cups of Coffee: **X**	*Nervousness Scores:* **Y**
1	1
1	1
1	2
2	2
2	3
3	4
3	5
3	9
4	5
4	6
5	8
5	9
6	9
6	10



one data point for them. (As shown, some researchers circle such a data point to indicate that points are on top of each other.) In a larger study, many participants with a particular *X* score may obtain the same *Y,* so the number of data points may be considerably smaller than the number of pairs of raw scores.

When you conduct a correlational study, always begin the analysis by creating a scatterplot. The scatterplot allows you to see the relationship that is present and to map out the best way to summarize it. (Also, you can see whether the extreme scores from any outliers may be biasing your computations.) Published reports of correlational studies, however, often do *not* show the scatterplot. Instead, from the description provided, you should *envision* the scatterplot, and then you will understand the relationship formed by the data. You get the description of the scatterplot from the correlation coefficient. A correlation coefficient communicates two important characteristics of a relationship: the *type* of relationship that is present and the *strength* of the relationship.

> ***REMEMBER*** A correlation coefficient is a statistic that communicates the *type* and *strength* of relationship.

TYPES OF RELATIONSHIPS

The **type of relationship** that is present in a set of data is the overall direction in which the *Y* scores change as the *X* scores change. There are two general types of relationships: *linear* and *nonlinear* relationships.

Linear Relationships

The term *linear* means "straight line," and a linear relationship forms a pattern that follows *one* straight line. This is because in a **linear relationship**, as the *X* scores increase, the *Y* scores tend to change in only *one* direction. To understand this, first look at the data points in the scatterplot on the left in Figure 7.2. This shows the relationship between the hours that students study and their test performance. A scatterplot that slants in only one direction like this indicates a linear relationship: it indicates that as students study longer, their grades tend only to increase. The scatterplot on the right in Figure 7.2 shows the relationship between the hours that students watch television and their test scores. It too is a linear relationship, showing that, as students watch more television, their test scores tend only to decrease.

For our discussions, we will summarize a scatterplot by drawing a line around its outer edges. (Published research will *not* show this.) As in Figure 7.2, a scatterplot that forms an *ellipse* that slants in one direction indicates a linear relationship: by slanting, it indicates that the *Y* scores are changing as the *X* scores increase; slanting in one direction indicates it is linear relationship.

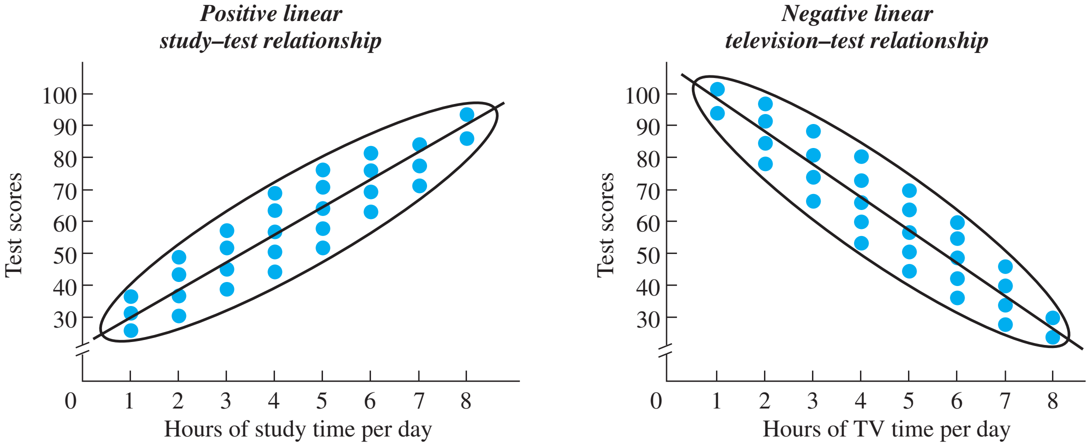
Further, as shown, we can also summarize a relationship by drawing a line through the scatterplot. (Published research *will* show this.) The line is called the *regression line*. While the correlation coefficient is the *statistic* that summarizes a relationship, the regression line is the *line* on a graph that summarizes the relationship. We will discuss the procedures for drawing the line in the next chapter, but for now, the **regression line** summarizes a relationship by passing through the center of the scatterplot. That is, although all data points are not *on* the line, the distance that some are above the line equals the distance that others are below it, so the regression line passes through the center of the scatterplot. Therefore, think of the regression line as showing the linear—straight line—relationship hidden in the data: It is how we visually summarize the general pattern in the relationship.

> ***REMEMBER*** The *regression line* summarizes a relationship by passing through the center of the scatterplot.

The difference between the scatterplots in Figure 7.2 illustrates the two subtypes of linear relationships that occur, depending on the *direction* in which the *Y* scores change. The study–test relationship is a positive relationship. In a **positive linear relationship,** as the *X* scores increase, the *Y* scores also tend to increase. Thus, low *X* scores are paired with low *Y* scores, and high *X* scores are paired with high *Y* scores. Any relationship that fits the pattern "the more *X,* the more *Y*" is a positive linear relationship.

FIGURE 7.2

Scatterplots showing positive and negative linear relationships



(Remember *positive* by remembering that as the *X* scores increase, the *Y* scores change in the direction away from zero, toward higher *positive* scores.)

On the other hand, the television–test relationship is a negative relationship. In a **negative linear relationship,** as the *X* scores increase, the *Y* scores tend to decrease. Low *X* scores are paired with high *Y* scores, and high *X* scores are paired with low *Y* scores. Any relationship that fits the pattern "the more *X,* the less *Y*" is a negative linear relationship. (Remember *negative* by remembering that as the *X* scores increase, the *Y* scores change toward zero, heading toward *negative* scores.)

Note: The term *negative* does not mean that there is something wrong with a relationship. It merely indicates the direction in which the *Y* scores change as the *X* scores increase.

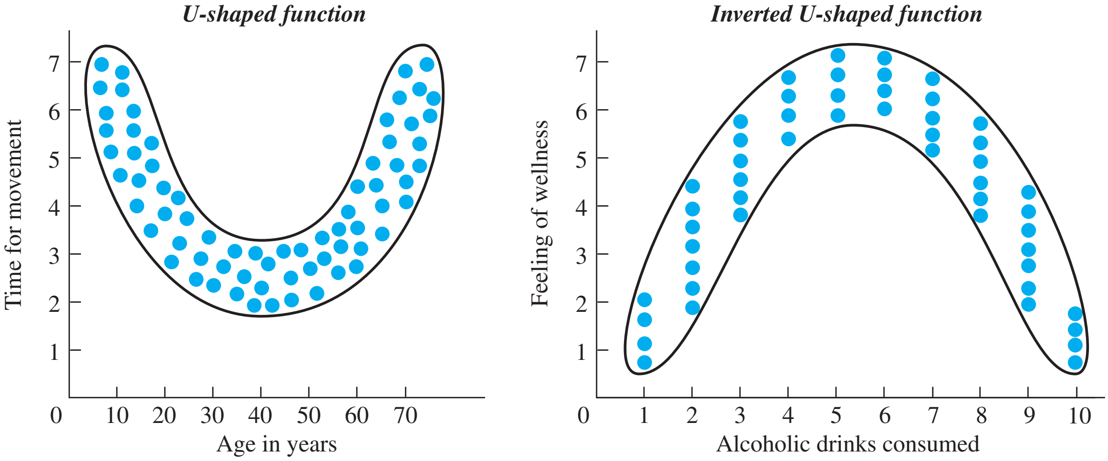
Nonlinear Relationships

If a relationship is not linear, then it is nonlinear. *Nonlinear* means that the data cannot be summarized by *one straight* line. Another name for a nonlinear relationship is a curvilinear relationship. In a **nonlinear,** or **curvilinear, relationship,** as the *X* scores change, the *Y* scores do not tend to *only* increase or *only* decrease: At some point, the *Y* scores change their direction of change.

Nonlinear relationships come in many different shapes, but Figure 7.3 shows two common ones. The scatterplot on the left shows the relationship between a person's age and the amount of time required to move from one place to another. Very young children move slowly, but as age increases, movement time decreases. Beyond a certain age, however, the time scores change direction and begin to increase. (Such a relationship is called *U-shaped.*) The scatterplot on the right shows the relationship between the number of alcoholic drinks consumed and feeling well. At first, people tend to feel better as they drink, but beyond a certain point, drinking more makes them feel progressively worse. (Such a scatterplot reflects an *inverted U-shaped relationship*.) Curvilinear relationships may be more complex than those above, producing a wavy pattern that repeatedly changes direction. To be nonlinear, however, a scatterplot does not need to be *curved.* A scatterplot might be best summarized by straight regression

FIGURE 7.3

Scatterplots showing nonlinear relationships



lines that form a V, an inverted V, or any other shape. It would still be nonlinear as long as it does not fit *one* straight line.

Notice that the terms *linear* and *nonlinear* are also used to describe relationships found in experiments. If, as the amount of the independent variable (X) increases, the dependent scores (Y) also increase, then it is a positive linear relationship. If the dependent scores decrease as the independent variable increases, it is a negative relationship. And if, as the independent variable increases, the dependent scores change their direction of change, it is a nonlinear relationship.

How the Correlation Coefficient Describes the Type of Relationship

Remember that the correlation coefficient is a number that we compute using our data. We communicate that the data form a linear relationship first because we compute a *linear* correlation coefficient—a coefficient whose formula is designed to summarize a linear relationship. (Behavioral research focuses primarily on linear relationships, so we'll discuss only them.) How do you know whether data form a linear relationship? If the scatterplot generally follows a straight line, then linear correlation is appropriate. Also, sometimes, researchers describe the extent to which a nonlinear relationship has a linear component and somewhat fits a straight line. Here, too, linear correlation is appropriate. However, do not try to summarize a nonlinear relationship by computing a linear correlation coefficient. This is like putting a round peg into a square hole: The data won't fit a straight line very well, and the correlation coefficient won't accurately describe the relationship.

The correlation coefficient communicates not only that we have a linear relationship but also whether it is positive or negative. Sometimes our computations will produce a negative number (with a minus sign), indicating that we have a negative relationship. Other data will produce a positive number (and we place a plus sign with it), indicating that we have a positive relationship. Then, with a positive correlation coefficient we envision a scatterplot that slants upward as the X scores increase. With a negative coefficient we envision a scatterplot that slants downward as the X scores increase.

The other characteristic of a relationship communicated by the correlation coefficient is the *strength* of the relationship.

STRENGTH OF THE RELATIONSHIP

Recall that the **strength of a relationship** is the extent to which one value of Y is consistently paired with one and only one value of X. The size of the coefficient that we compute (ignoring its sign) indicates the strength of the relationship. The largest value you can obtain is 1, indicating a perfectly consistent relationship. (*You cannot beat perfection so you can never have a coefficient greater than 1!*) The smallest possible value is 0, indicating that no relationship is present. Thus, when we include the positive or negative sign, the correlation coefficient may be any value between -1 and $+1$. The *larger* the absolute value of the coefficient, the *stronger* the relationship. In other words, the closer the coefficient is to ±1, the more consistently one value of Y is paired with one and only one value of X.

> ***REMEMBER*** A *correlation coefficient* has two components: The sign indicates either a positive or a negative linear relationship; the absolute value indicates the strength of the relationship.

Correlation coefficients do not, however, measure in units of "consistency." Thus, if one correlation coefficient is +.40 and another is +.80, we *cannot* conclude that one relationship is twice as consistent as the other. Instead, we evaluate any correlation coefficient by comparing it to the extreme values of 0 and ±1. The starting point is a perfect relationship.

Perfect Association

A correlation coefficient of +1 or −1 describes a perfectly consistent linear relationship. Figure 7.4 shows an example of each. (In this and the following figures, first look at the scores to see how they pair up. Then look at the scatterplot. Other data having the same correlation coefficient produce similar patterns, so we envision similar scatterplots.)

Here are four interrelated ways to think about what a correlation coefficient tells you about the relationship. First, it indicates *the relative degree of consistency.* A coefficient of ±1 indicates that *everyone* who obtains a particular *X* score obtains one and only one value of *Y*. Every time *X* changes, the *Y* scores all change to one new value.

Second, and conversely, *the coefficient communicates the variability in the Y scores paired with an X.* When the coefficient is ±1, only one *Y* is paired with an *X,* so there is no variability—no differences—among the *Y* scores paired with each *X.*

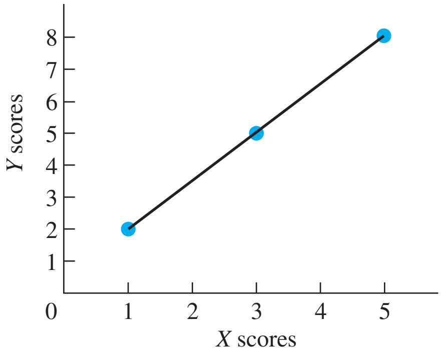
Third, *the coefficient communicates how closely the scatterplot fits the regression line.* Because a coefficient equal to ±1 indicates zero variability or *spread* in the *Y*

FIGURE 7.4

Data and scatterplots reflecting perfect positive and negative correlations

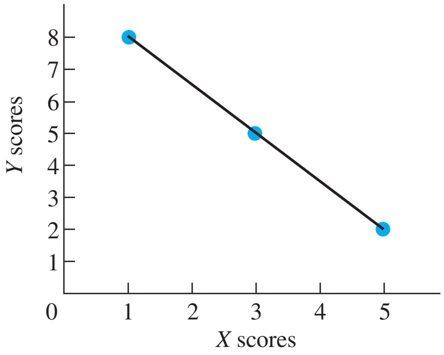
Perfect positive coefficient = +1.0

X	Y
1	2
1	2
1	2
3	5
3	5
3	5
5	8
5	8
5	8



Perfect negative coefficient = −1.0

X	Y
1	8
1	8
1	8
3	5
3	5
3	5
5	2
5	2
5	2



scores at each *X,* we know that their data points are on top of one another. And, because it is a perfect straight-line relationship, all data points will lie *on* the regression line.

Fourth, *the coefficient communicates the relative accuracy of our predictions* when we predict participants' *Y* scores by using their *X* scores. A coefficient of ±1 indicates perfect accuracy in predictions: because only one *Y* score occurs with each *X* we will *know* every participants' *Y* score every time. Look at the positive relationship back in Figure 7.4: We will always know when people have a *Y* score of 2 (when they have an *X* of 1), and we will know when they have a *different Y* of 5 or 8 (when they have an *X* of 3 or 5, respectively). The same accuracy is produced in the negative relationship.

Note: In statistical lingo, because we can perfectly predict the *Y* scores here, we would say that these *X* variables are perfect "predictors" of *Y*. Further, recall from Chapter 5 that the variance is a way to measure differences among scores. When we can accurately predict when *different Y* scores will occur, we say we are "accounting for the variance in *Y*." A better predictor (*X*) will account for more of the variance in *Y* . To communicate the perfect accuracy in predictions with correlations of ±1, we would say that "100% of the variance is accounted for."

> ***REMEMBER*** The correlation coefficient communicates the consistency of the relationship, the variability of the *Y* scores at each *X*, the shape of the scatterplot, and our accuracy when using *X* to predict *Y* scores.

Intermediate Association

A correlation coefficient that does not equal ±1 indicates that the data form a linear relationship to only some degree. The closer the coefficient is to ±1, however, the closer the data are to forming a perfect relationship, and the closer the scatterplot is to forming a straight line. Therefore, the way to interpret any other value of the correlation coefficient is to compare it to ±1.

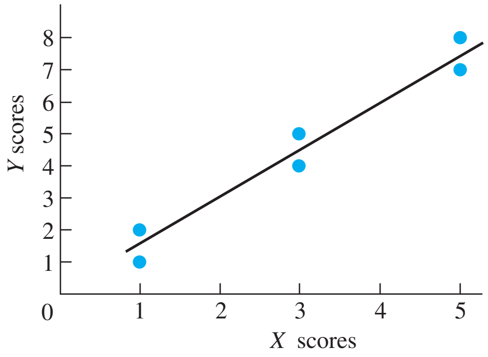
For example, Figure 7.5 shows data that produce a correlation coefficient of +.98. Again interpret the coefficient in four ways. First, *consistency:* A coefficient less than ±1 indicates that not every participant at a particular *X* had the same *Y*. However, a coefficient of +.98 is close to +1, so there is close to perfect consistency. That is, even though different values of *Y* occur with the same *X*, the *Y* scores are relatively close to each other.

Second, *variability:* By indicating reduced consistency, this coefficient indicates that there is now variability (differences) among the *Y* scores at each *X*. However, because

FIGURE 7.5

Data and scatterplot reflecting a correlation coefficient of +.98

X	Y
1	1
1	2
1	2
3	4
3	5
3	5
5	7
5	8
5	8



+.98 is close to +1 (close to the situation where there is zero variability) we know that the variability in our *Y* scores is close to zero and relatively small.

Third, *the scatterplot*: Because there is variability in the *Y*s at each *X*, not all data points fall *on* the regression line. Back in Figure 7.5, variability in *Y* scores results in a group of data points at each *X* that are vertically spread out above and below the regression line. However, a coefficient of +.98 is close to +1, so we know that the data points are close to, or hug, the regression line, resulting in a scatterplot that is a narrow, or skinny, ellipse.

Fourth, *predictions:* When the correlation coefficient is not ±1, knowing participants' *X* scores allows us to predict only *around* what their *Y* score will be. For example, in Figure 7.5, for an *X* of 1 we'd predict that a person has a *Y* around 1 or 2, but we won't know which. In other words, we will have some error in our predictions. However, a coefficient of +.98 is close to +1 (close to the situation where there is zero error). This indicates that our predicted *Y* scores will be close to the actual *Y* scores that participants obtained, and so our error will be small. With predictions that are close to participants' *Y* scores, we would describe this *X* variable as "a good predictor of *Y*." Further, because we will still know when *Y* scores around 1 or 2 occur and when *different Y*s around, say, 4 or 5 occur, this *X* variable still "accounts for" a sizable portion of the variance among all *Y* scores.

The key to understanding the strength of any relationship is this:

> **As the variability—differences—in the *Y* scores paired with an *X* becomes larger, the relationship becomes weaker.**

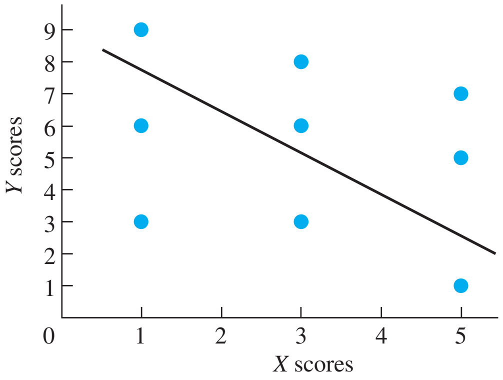
The correlation coefficient communicates this because, as the variability in the *Y*s at each *X* becomes larger, the value of the correlation coefficient approaches 0. Figure 7.6 shows data that produce a correlation coefficient of only −.28. (The fact that this is a negative relationship has nothing to do with its strength.) Here the spread in the *Y* scores (the variability) at each *X* is relatively large. This does two things that are contrary to a relationship. First, instead of seeing a different *Y* scores at *different X*s, we see very different *Y*s for individuals who have the *same X*. Second, instead of seeing one value of *Y* at only one *X*, the *Y* scores at different *X*s overlap, so we see one value of *Y* paired with *different* values of *X*. Thus, the weaker the relationship, the more the *Y* scores tend to change when *X* does not, and the more the *Y* scores tend to stay the same when *X* does change.

Thus, it is the variability in *Y* at each *X* that determines the consistency of a relationship, which in turn determines the characteristics we've examined. Thus, a coefficient

FIGURE 7.6

Data and scatterplot reflecting a correlation coefficient of −.28.

X	Y
1	9
1	6
1	3
3	8
3	6
3	3
5	7
5	5
5	1



of −.28 is not very close to ±1, so, as in Figure 7.6, we know that (1) only barely does one value or close to one value of *Y* tend to be associated with one value of *X*; (2) conversely, the variability among the *Y* scores at every *X* is relatively large; (3) the large differences among *Y* scores at each *X* produce data points on the scatterplot at each *X* that are vertically spread out, producing a "fat" scatterplot that does not hug the regression line; and (4) because each *X* is paired with a wide variety of *Y* scores, knowing participants' *X* will not allow us to accurately predict their *Y*. Instead, our prediction errors will be large because we have only a very general idea of when higher *Y* scores tend to occur and when lower *Y* scores occur. Thus, this *X* is a rather poor "predictor" because it "accounts" for little of the variance among *Y* scores.

> ***REMEMBER*** Greater *variability* in the *Y* scores at each *X* reduces the strength of a relationship and the size of the correlation coefficient.

Zero Association

The lowest possible value of the correlation coefficient is 0, indicating that no relationship is present. Figure 7.7 shows data that produce such a coefficient. When no relationship is present, the scatterplot is circular or forms an ellipse that is parallel to the *X* axis. Likewise, the regression line is a horizontal line.

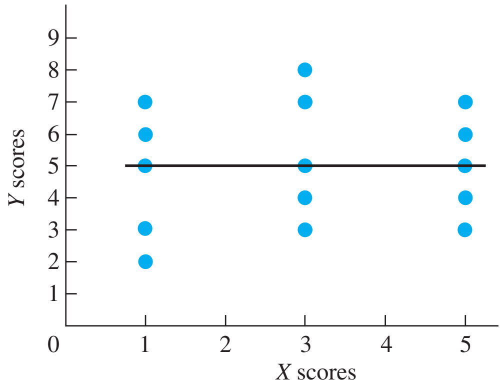
A scatterplot like this is as far from forming a slanted straight line as possible, and a correlation coefficient of 0 is as far from ±1 as possible. Therefore, this coefficient tells us that no *Y* score tends to be consistently associated with only one value of *X*. Instead, the *Y*s found at one *X* are virtually the same as those found at any other *X*. This also means that knowing someone's *X* score will not in any way help us to predict the corresponding *Y*. (We can account for none of the variance in *Y*.) Finally, this coefficient indicates that the spread in *Y* at any *X* equals the overall spread of *Y* in the data, producing a scatterplot that is a circle or horizontal ellipse that in no way hugs the regression line.

> ***REMEMBER*** The larger a correlation coefficient (whether positive or negative), the stronger the linear relationship, because the less the *Y*s are spread out at each *X*, and so the closer the data come to forming a straight line.

FIGURE 7.7

Data and scatterplot reflecting a correlation coefficient of 0.

X	Y
1	2
1	3
1	5
1	6
1	7
3	3
3	4
3	5
3	7
3	8
5	3
5	4
5	5
5	6
5	7



A QUICK REVIEW

- As X scores increase, in a positive linear relationship, the Y scores tend to increase, and in a negative linear relationship, the Y scores tend to decrease.
- The larger the correlation coefficient, the more consistently one Y occurs with one X, the smaller the variability in Ys at an X, the more accurate our predictions, and the narrower the scatterplot.

More Examples

A coefficient of $+.84$ indicates (1) as X increases, Y consistently increases; (2) everyone at a particular X shows little variability in Y scores; (3) by knowing an individual's X, we can closely predict his/her Y score; and (4) the scatterplot is a narrow ellipse, with the data points lying near the upward slanting regression line. However, a coefficient of $+.38$ indicates (1) as X increases, Y somewhat consistently increases; (2) a wide variety of Y scores paired with a particular X; (3) knowing an X score does not produce accurate predictions of the paired Y score; and (4) the scatterplot is a wide ellipse around the upward slanting regression line.

For Practice

1. In a ______ relationship, as the X scores increase, the Y scores increase or decrease only. This is not true in a ______ relationship.
2. The more that you smoke cigarettes, the lower is your healthiness. This is a ______ linear relationship, producing a scatterplot that slants ______ as X increases.
3. The more that you exercise, the better is your muscle tone. This is a ______ linear relationship, producing a scatterplot that slants ______ as X increases.
4. In a stronger relationship the variability among the Y scores at each X is ______, producing a scatterplot that forms a ______ ellipse.
5. The ______ line summarizes the scatterplot.

Answers

1. linear; nonlinear
2. negative; down
3. positive; up
4. smaller; narrower
5. regression

THE PEARSON CORRELATION COEFFICIENT

Now that you understand the correlation coefficient, we can discuss its computation. However, statisticians have developed a number of correlation coefficients having different names and formulas. Which one is used in a particular study depends on the nature of the variables and the scale of measurement used to measure them. By far the most common correlation coefficient in behavioral research is the Pearson correlation coefficient. The **Pearson correlation coefficient** describes the linear relationship between two interval variables, two ratio variables, or one interval and one ratio variable. (Technically, its name is the *Pearson Product Moment Correlation Coefficient*.) The symbol for the Pearson correlation coefficient is the lowercase r. (All of the example coefficients in the previous section were rs.)

Mathematically r compares how consistently each value of Y is paired with each value of X. In Chapter 6, you saw that we compare scores from different variables by transforming them into z-scores. Computing r involves transforming each Y score into a z-score (call it z_Y), transforming each X score into a z-score (call it z_X), and then determining the "average" amount of correspondence between all pairs of z-scores. The Pearson correlation coefficient is defined as

$$r = \frac{\sum(z_X z_Y)}{N}$$

Multiplying each z_X times the paired z_Y, summing the products, and then dividing by N produces the average correspondence between the scores.

Luckily, the computational formula for r does all of that at once. It is derived from the above formula by replacing the symbols z_X and z_Y with their formulas and then, in each, replacing the symbols for the mean and standard deviation with their formulas. This produces a monster of a formula. After reducing it, we have the smaller monster below.

The computational formula for the Pearson correlation coefficient is

$$r = \frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{[N(\Sigma X^2) - (\Sigma X)^2][N(\Sigma Y^2) - (\Sigma Y)^2]}}$$

In the numerator, N (the number of pairs) is multiplied times ΣXY. From this, subtract the quantity obtained by multiplying (ΣX) times (ΣX). In the denominator, in the left brackets, multiply N times ΣX^2 and from that subtract $(\Sigma X)^2$. In the right bracket, multiply N times ΣY^2 and from that subtract $(\Sigma Y)^2$. Multiply the answers in the two brackets together and find the square root. Then divide the denominator into the numerator and, voilà, the answer is the Pearson r.

As an example, say that we ask ten people the number of times they visited a doctor in the last year and the number of glasses of orange juice they drink daily. We obtain the data in Figure 7.8. To describe the linear relationship between juice drinking and doctor visits, (two ratio variables,) we compute r.

Table 7.1 shows a good way to organize your computations. In addition to the columns for X and Y, create columns containing X^2, Y^2, and XY. Sum all columns. Then square ΣX and ΣY.

FIGURE 7.8

The relationship between number of glasses of orange juice consumed daily and number of yearly doctor visits.

Participant	*Juice Scores:* **X**	*Doctor Visits:* **Y**
1	0	8
2	0	7
3	1	7
4	1	6
5	1	5
6	2	4
7	2	4
8	3	4
9	3	2
10	4	0

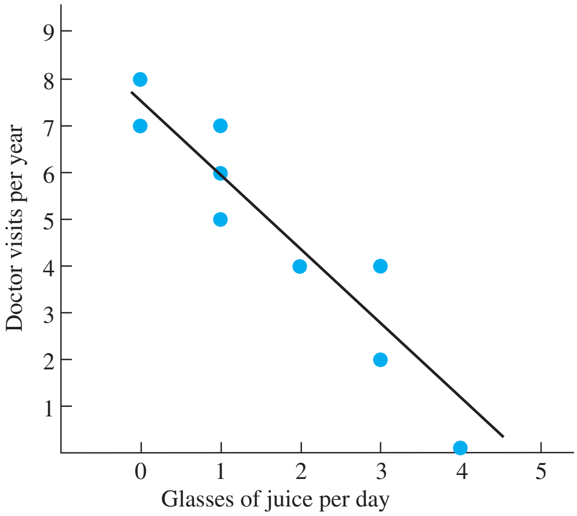


TABLE 7.1
Sample Data for Computing the r Between Orange Juice Consumed (the X variable) and Doctor Visits (the Y variable)

	Glasses of Juice per Day		*Doctor Visits per Year*		
Participant	X	X^2	Y	Y^2	XY
1	0	0	8	64	0
2	0	0	7	49	0
3	1	1	7	49	7
4	1	1	6	36	6
5	1	1	5	25	5
6	2	4	4	16	8
7	2	4	4	16	8
8	3	9	4	16	12
9	3	9	2	4	6
10	4	16	0	0	0
$N = 10$	$\Sigma X = 17$	$\Sigma X^2 = 45$	$\Sigma Y = 47$	$\Sigma Y^2 = 275$	$\Sigma XY = 52$
	$(\Sigma X)^2 = 289$		$(\Sigma Y)^2 = 2209$		

Putting these quantities in the formula for r we get

$$r = \frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{[N(\Sigma X^2) - (\Sigma X)^2][N(\Sigma Y^2) - (\Sigma Y)^2]}} = \frac{10(52) - (17)(47)}{\sqrt{[10(45) - 289][10(275) - 2209]}}$$

In the numerator, multiplying 10 times 52 is 520. Also, 17 times 47 is 799. Now we have

$$r = \frac{520 - 799}{\sqrt{[10(45) - 289][10(275) - 2209]}}$$

Complete the numerator: 799 *from* 520 is −279. (Note the negative sign.)

In the denominator, first perform the operations within each bracket. In the left bracket, 10 times 45 is 450. From that subtract 289, obtaining 161. In the right bracket, 10 times 275 is 2750. From that subtract 2209, obtaining 541. We have

$$r = \frac{-279}{\sqrt{[161][541]}}$$

Now multiply the quantities in the brackets together: 161 times 541 equals 87,101. After taking the square root we have

$$r = \frac{-279}{295.129}$$

Divide and there you have it: $r = -.95$.

Thus, the correlation coefficient between orange juice drinks and doctor visits is −.95. (*Note:* We usually round the coefficient to two decimals.) Had this been a positive relationship, r would not be negative and we would include the + sign. Instead, on a scale of 0 to ±1, a −.95 indicates that this is an extremely strong, negative linear relationship. Therefore, we envision a very narrow, downward slanting scatterplot like that back in Figure 7.8. We know that each amount of orange juice is associated with a very small range of doctor visits, and as juice scores increase, doctor visits consistently decrease. Further, based on participants' juice scores, we can very accurately predict their doctor visits. (Orange juice is an extremely good "predictor" of doctor visits, accounting for a substantial portion of the variance in these Y scores.) Of course if the correlation were this large in real life, we'd all be drinking a lot more orange juice, incorrectly thinking that this would *cause* fewer doctor visits.

> ***REMEMBER*** Compute the *Pearson correlation coefficient* to describe the linear relationship between interval and/or ratio variables.

Recognize that this correlation coefficient describes the relationship in our *sample*. Ultimately we will want to describe the laws of nature, inferring the correlation coefficient we would expect to find if we could measure everyone in the population. However, before we can do this, we must perform the appropriate *inferential* procedure (discussed in Chapter 11). Only if our sample correlation coefficient passes the inferential test will we then talk about how this relationship occurs in nature.

A QUICK REVIEW

- The Pearson correlation coefficient (r) describes the linear relationship between two interval and/or ratio variables.

More Examples

X	Y
1	3
1	2
2	4
2	5
3	5
3	6

To compute r for the above scores:

$\Sigma X = 12$, $(\Sigma X)^2 = 144$, $\Sigma X^2 - 28$, $\Sigma Y = 25$, $(\Sigma Y)^2 = 625$, $\Sigma Y^2 = 155$, $\Sigma XY = 56$ and $N = 6$

$$r = \frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{[N(\Sigma X^2) - (\Sigma X)^2][N(\Sigma Y^2) - (\Sigma Y)^2]}}$$

$$r = \frac{6(56) - (12)(25)}{\sqrt{[6(28) - 144][6(115) - 625]}}$$

In the numerator, 6 times 56 is 336, and 12 times 25 is 300, so

$$r = \frac{336 - 300}{\sqrt{[(6(28) - 144][6(115) - 625]}}$$

$$r = \frac{+36}{\sqrt{[6(28) - 144][6(115) - 625]}}$$

In the denominator, 6 times 28 is 168; 6 times 115 is 690, so

$$r = \frac{+36}{\sqrt{[168 - 144][690 - 625]}} = \frac{+36}{\sqrt{[24][65]}}$$

$$r = \frac{+36}{\sqrt{1560}} = \frac{+36}{39.497} = +.91$$

For Practice

Compute r for the following:

X	Y
1	1
1	3
2	2
2	4
3	4

Answer

$$r = \frac{5(28) - (9)(14)}{\sqrt{[5(19) - 81][5(46) - 196]}}$$

$$= \frac{+14}{\sqrt{[14][34]}} = +.64$$

THE SPEARMAN RANK-ORDER CORRELATION COEFFICIENT

Another very common correlation coefficient is used when we have *ordinal* scores (when we have the equivalent of 1st, 2nd, 3rd, etc., on each variable). The **Spearman rank-order correlation coefficient** describes the linear relationship between two variables when measured by ranked scores. The symbol for the Spearman correlation coefficient is r_S. (The subscript s stands for Spearman.)

Sometimes we compute r_S because we have initially assigned each participant a rank on each of two variables. Or, if we want to correlate one ranked variable with one interval or ratio variable, we transform the interval or ratio scores into ranked scores (we might give the participant with the highest score a 1, the next highest score is ranked 2, and so on). Either way that we obtain the ranks, r_S tells us the extent to which the ranks on one variable consistently match the ranks on the other variable to form a linear relationship. If every participant has the same rank on both variables, r_S will equal $+1$. If everyone's rank on one variable is the opposite of his or her rank on the other variable, r_S will equal -1. With only some degree of consistent pairing of the ranks, r_S will be between 0 and ± 1. If there is no consistent pairing, r_S will equal 0.

Ranked scores often occur in behavioral research because a variable is difficult to measure quantitatively. Instead we must evaluate participants by asking observers to make subjective judgments that are then used to rank order the participants. For example, say that we ask two observers to judge how aggressively children behave while playing. Each observer assigns the rank of 1 to the most aggressive child, 2 to the second-most aggressive child, and so on. Because r_S describes the consistency with which rankings match, one use of r_S is to determine the extent to which the two observers' rankings agree.

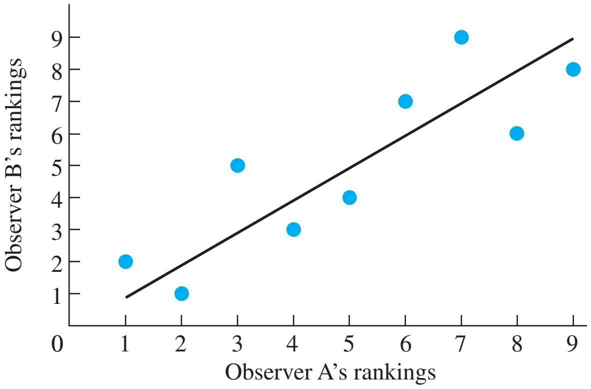
Figure 7.9 shows the ranked scores and the resulting scatterplot that the two observers might produce for nine children. Notice that we treat each observer as a variable. Judging from the scatterplot, it appears that they form a positive relationship. To describe this relationship, we compute r_S.

Note: If you have any "tied ranks" (when two or more participants receive the same score on the *same* variable) you must first adjust them as described in the section "Resolving Tied Ranks" in Chapter 15.

FIGURE 7.9

Sample data for computing r_S between rankings assigned to children by observer A and observer B

Participant	*Observer A:* X	*Observer B:* Y
1	4	3
2	1	2
3	9	8
4	8	6
5	3	5
6	5	4
7	6	7
8	2	1
9	7	9



Here is the formula for r_S.

The computational formula for the Spearman rank-order correlation coefficient is

$$r_S = 1 - \frac{6(\Sigma D^2)}{N(N^2 - 1)}$$

The logic of the formula here is similar to that in the previous Pearson formula, except that r_S accommodates the peculiarities of ranks (e.g., zero cannot occur). This is why the formula always contains the 6 in the numerator. The *D* in the numerator stands for the difference between the two ranks in each *X*–*Y* pair, and *N* is the number of pairs of ranks. Note that after dealing with the fraction, the final step is to subtract from 1.

A good way to organize your computations is shown in Table 7.2. For the column labeled *D,* either subtract every *X* from its paired *Y* or, as shown, every *Y* from its *X*. Then compute D^2 by squaring the difference in each pair. Finally, determine the sum of the squared differences, ΣD^2 (here ΣD^2 is 18). You will also need *N,* the number of *X*–*Y* pairs (here 9), and N^2 (here 81). Filling in the formula gives

$$r_S = 1 - \frac{6(\Sigma D^2)}{N(N^2 - 1)} = 1 - \frac{6(18)}{9(81 - 1)}$$

In the numerator, 6 times 18 is 108. In the denominator, 81 − 1 is 80, and 9 times 80 is 720. Now

$$r_S = 1 - \frac{108}{720}$$

After dividing

$$r_S = 1 - .15$$

Subtracting yields

$$r_S = +.85$$

TABLE 7.2

Data Arrangement for Computing r_S

Participant	*Observer A:* **X**	*Observer B:* **Y**	**D**	**D^2**
1	4	3	1	1
2	1	2	−1	1
3	9	8	1	1
4	8	6	2	4
5	3	5	−2	4
6	5	4	1	1
7	6	7	−1	1
8	2	1	1	1
9	7	9	−2	4
				$\Sigma D^2 = 18$

Thus, on a scale of 0 to ±1, these rankings form a consistent linear relationship to the extent that $r_S = +.85$. This tells us that a child receiving a particular rank from one observer tended to receive very close to the same rank from the other observer. Therefore, the data form a rather narrow scatterplot that tends to hug the regression line. (The r_S must also pass the inferential test in Chapter 11 before we can draw any inferences about it.)

> ***REMEMBER*** Compute the *Spearman correlation coefficient* to describe the linear relationship between two ordinal variables.

A QUICK REVIEW

- The Spearman correlation coefficient (r_S) describes the type and strength of the linear relationship between two sets of ranks.

More Examples

To determine r_S for the following ranks, find the D of each X–Y pair, and then D^2 and N.

X	Y		D	D²
1	1	=	0	0
2	2	=	0	0
4	3	=	1	1
3	6	=	−3	9
6	5	=	1	1
5	4	=	1	1
			$\Sigma D^2 =$	12
			$N =$	6

$$r_s = 1 - \frac{6(\Sigma D^2)}{N(N^2 - 1)} = 1 - \frac{6(12)}{6(36 - 1)}$$

$$= 1 - \frac{72}{210} = 1 - .343 = +.66$$

For Practice

1. When do we compute r_S ?
2. The first step in computing r_S is to compute each ____?

For the ranks:

X	Y
1	2
2	1
3	3
4	5
5	4

3. The $\Sigma D^2 =$ ____ and $N =$ _____?
4. The $r_S =$ ____?

Answers

1. When we have ordinal scores.
2. *D*
3. $\Sigma D^2 = 4$; $N = 5$
4. $r_S = +.80$

THE RESTRICTION OF RANGE PROBLEM

As you learn about conducting research, you'll learn of potential mistakes to avoid that otherwise can lead to problems with your statistical conclusions. One important mistake to avoid with all correlation coefficients is called the **restriction of range problem**. It occurs when we have data in which the range between the lowest and highest scores on one or both variables is limited. This will produce a correlation coefficient that is *smaller* than it would be if the range were not restricted. Here's why.

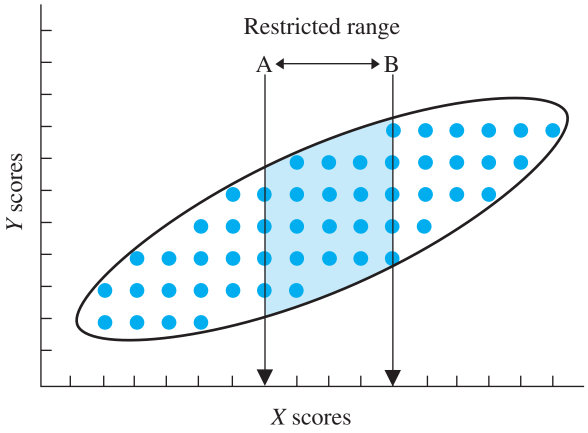


FIGURE 7.10

Scatterplot showing restriction of range in *X* scores

In Figure 7.10, first consider the entire scatterplot showing the full (unrestricted) range of *X* and *Y* scores. We see a different batch of similar *Y* scores occurring as *X* increases, producing an elongated, relatively narrow ellipse that clearly slants upwards. Therefore, the correlation coefficient will be relatively large, and we will correctly conclude that there is a strong linear relationship between these variables.

However, say that instead we restricted the range of *X* when measuring the data, giving us only the scatterplot located between the lines labeled A and B in Figure 7.10. Now, we are seeing virtually the same batch of *Y* scores as these few *X* scores increase. This produces a scatterplot that looks relatively fat and more horizontal. Therefore, the correlation coefficient from these data will be very close to 0, so we will conclude that there is a very weak—if any—linear relationship here. This would be wrong, however, because without us restricting the range, we would have seen that nature actually produces a much stronger relationship. (Because either variable can be the *X* or *Y* variable, restricting the range of *Y* has the same effect.)

REMEMBER *Restricting the range* of *X* or *Y* scores leads to an underestimate of the true strength of the relationship between the variables.

How do you avoid restricting the range? Generally, restriction of range occurs when researchers are too selective when obtaining participants. Thus, if you study the relationship between participants' high school grades and their subsequent salaries, don't restrict the range of grades by testing only honor students: Measure all students to get the entire range of grades. Or, if you're correlating personality types with degree of emotional problems, don't study only college students. People with severe emotional problems tend not to be in college, so you won't have their scores. Instead, include the full range of people from the general population. Likewise, any task you give participants should not be too easy (because then everyone scores in a narrow range of very high scores), nor should the task be too difficult (because then everyone obtains virtually the same low score). In all cases, the goal is to allow a wide range of scores to occur on both variables so that you have a complete description of the relationship.

STATISTICS IN PUBLISHED RESEARCH: CORRELATION COEFFICIENTS

In APA-style publications, the Pearson correlation coefficient is symbolized by *r*, and the Spearman coefficient is symbolized by r_S. Later we'll also see other coefficients that are designed for other types of scores, and you may find additional, advanced coefficients in published research. However, all coefficients are interpreted in the same ways that we have discussed: the coefficient will have an absolute value between 0 and 1, with 0 indicating no relationship and 1 indicating a perfectly consistent relationship.

In real research, however, a correlation coefficient near ±1 simply does not occur. Recall from Chapter 2 that individual differences and extraneous environmental variables produce inconsistency in behaviors, which results in inconsistent relationships.

Therefore, adjust your expectations: Most research produces coefficients with absolute values in the neighborhood of only .30 to .50. Thus, coefficients below .20 tend to be considered very weak and often negligible. Coefficients around .40 are described as strong and coefficients above .60 are uncommon and considered very strong. A correlation near 1 is most likely a computational error.

PUTTING IT ALL TOGETHER

It should be obvious why you should compute a correlation coefficient whenever you have a relationship to summarize. It is the one number that allows you to envision and summarize the important information in a scatterplot. For example, in our study on nervousness and the amount of coffee consumed, say that I tell you that the *r* in the study equals +.50. *Without even seeing the data,* you know this is a positive linear relationship such that as coffee consumption increases, nervousness also tends to increase. Also, you know that it is a rather consistent relationship so there are similar *Y* scores paired with an *X*, producing a narrow, elliptical scatterplot that hugs the regression line. And, you know that coffee consumption is a reasonably good predictor of nervousness so, given someone's coffee score, you'll have considerable accuracy in predicting his or her nervousness score. No other type of statistic so directly summarizes a relationship. Therefore, as you'll see in later chapters, even when you conduct an experiment, always think "correlation coefficient" to describe the strength and type of relationship you've observed.

Using the SPSS Appendix As shown in Appendix B.4, the SPSS program will calculate the Pearson *r*, as well as computing the mean and standard deviation of the *X* scores and of the *Y* scores. SPSS will also compute the Spearman r_S (even if your data contains tied ranks.) Also, you may enter interval or ratio scores and the program will first convert them to ranks and then compute r_S.

CHAPTER SUMMARY

1. A *scatterplot* is a graph that shows the location of each pair of *X–Y* scores in the data. An *outlier* is a data point that lies outside of the general pattern in the scatterplot. It is produced when a participant has an unusual *X* or *Y* score.

2. The *regression line* summarizes a relationship by passing through the center of the scatterplot.

3. In a *linear relationship*, as the *X* scores increase, the *Y* scores tend to change in only *one* direction. In a *positive linear relationship,* as the *X* scores increase, the *Y* scores tend to increase. In a *negative linear relationship,* as the *X* scores increase, the *Y* scores tend to decrease. In a *nonlinear,* or *curvilinear, relationship,* as the *X* scores increase, the *Y* scores do not only increase or only decrease.

4. Circular or elliptical scatterplots that produce horizontal regression lines indicate no relationship. Scatterplots with regression lines sloping up as *X* increases indicate a positive linear relationship. Scatterplots with regression lines sloping down as *X* increases indicate a negative linear relationship. Scatterplots producing wavy regression lines indicate curvilinear relationships.

5. A *correlation coefficient* describes the *type* of relationship (the direction *Y* scores change) and the *strength* of the relationship (the extent to which one value of *Y* is consistently paired with one value of *X*).

6. A smaller absolute value of the correlation coefficient indicates a weaker, less consistent relationship, with greater variability in Y scores at each X, greater vertical spread in the scatterplot, and less accuracy in predicting Y scores based on correlated X scores.

7. The *Pearson correlation coefficient* (r) describes the type (either positive or negative) and the strength of the linear relationship between two interval and/or ratio variables.

8. The *Spearman rank-order correlation coefficient* (r_S) describes the type and strength of the linear relationship between two ordinal variables.

9. The *restriction of range problem* occurs when the range of scores from one or both variables is limited. Then the correlation coefficient underestimates the strength of the relationship that would be found if the range were not restricted.

10. Because a stronger relationship allows for greater accuracy in predicting Y scores, researchers say the X variable is a better *predictor* of Y scores, allowing us to *account for more variance in Y.*

KEY TERMS

ΣXY r r_S
correlation coefficient *136*
curvilinear relationship *141*
linear relationship *139*
negative linear relationship *141*
nonlinear relationship *141*
outlier *138*
Pearson correlation coefficient *147*
positive linear relationship *140*
regression line *140*
restriction of range *153*
scatterplot *138*
Spearman rank-order correlation coefficient *152*
strength of a relationship *142*
type of relationship *139*

REVIEW QUESTIONS

(Answers for odd-numbered questions are in Appendix D.)

1. What is the difference between an experiment and a correlational study in terms of how the researcher (a) collects the data? (b) examines the relationship?
2. (a) You have collected data that you think show a relationship. What do you do next? (b) What is the advantage of computing a correlation coefficient? (c) What two characteristics of a linear relationship are described by a correlation coefficient?
3. What are the two reasons why you can't conclude you have demonstrated a causal relationship based on correlational research?
4. (a) When do you compute a Pearson correlation coefficient? (b) When do you compute a Spearman coefficient?
5. (a) What is a scatterplot? (b) What is an outlier? (c) What is a regression line?
6. Why can't you obtain a correlation coefficient greater than ±1?

7. (a) Define a positive linear relationship. (b) Define a negative linear relationship. (c) Define a curvilinear relationship.

8. As the value of r approaches ±1, what does it indicate about the following? (a) The consistency in the X–Y pairs; (b) the variability of the Y scores at each X; (c) the closeness of Y scores to the regression line; (d) the accuracy with which we can predict Y if X is known.

9. What does a correlation coefficient equal to 0 indicate about the four characteristics in question 8?

10. (a) What is the restriction of range problem? (b) What produces a restricted range? (c) How is it avoided?

11. (a) What does a researcher mean when he states that a particular variable is a "a good predictor?" (b) What does a researcher mean when she says an X variable accounts for little of the variance in Y?

APPLICATION QUESTIONS

12. For each of the following, indicate whether it is a positive linear, negative linear, or nonlinear relationship: (a) Quality of performance (Y) increases with increased arousal (X) up to an optimal level; then quality of performance decreases with increased arousal. (b) Overweight people (X) are less healthy (Y). (c) As number of minutes of exercise increases each week (X), dieting individuals lose more pounds (Y). (d) The number of bears in an area (Y) decreases as the area becomes increasingly populated by humans (X).

13. Poindexter sees the data in question 12d and concludes, "We should stop people from moving into bear country so that we can preserve our bear population." What is the problem with Poindexter's conclusion?

14. For each of the following, give the symbol for the correlation coefficient you should compute. You measure (a) SAT scores and IQ scores; (b) taste rankings of tea by experts and those by novices; (c) finishing position in a race and amount of liquid consumed during the race.

15. Poindexter finds that $r = -.40$ between the variables of number of hours studied (X) and number of errors on a statistics test (Y). He also finds that $r = +.36$ between the variables of time spent taking the statistics test and the number of errors on the test. He concludes that the time spent taking a test forms a stronger relationship with the number of errors than does the amount of study time. (a) Describe the relative shapes of the two scatterplots. (b) Describe the relative amount of variability in Y scores at each X in each study. (c) Describe the relative closeness of Y scores to the regression line in each study. (d) Is Poindexter correct in his conclusion? If not, what's his mistake?

16. In question 15, (a) which variable is a better predictor of test errors and how do you know this? (b) Which variable accounts for more of the variance in test errors and how do you know this?

17. Foofy and Poindexter study the relationship between IQ score and high school grade average, measuring a large sample of students from PEST (the Program for Exceptionally Smart Teenagers), and compute $r = +.03$. They conclude that there is virtually no relationship between IQ and grade average. Should you agree or disagree with this conclusion? Is there a problem with their study?

18. A researcher measures the following scores for a group of people. The *X* variable is the number of errors on a math test, and the *Y* variable is the person's level of satisfaction with his/her performance. (a) With such ratio scores, what should the researcher conclude about this relationship? (*Hint:* Compute something!) (b) How well will he be able to predict satisfaction scores using this relationship?

Participant	***Errors*** **X**	***Satisfaction*** **Y**
1	9	3
2	8	2
3	4	8
4	6	5
5	7	4
6	10	2
7	5	7

19. You want to know if a nurse's absences from work in one month (*Y*) can be predicted by knowing her score on a test of psychological "burnout" (*X*). What do you conclude from the following ratio data?

Participant	***Burnout*** **X**	***Absences*** **Y**
1	2	4
2	1	7
3	2	6
4	3	9
5	4	6
6	4	8
7	7	7
8	7	10
9	8	11

20. In the following data, the *X* scores reflect participants' rankings in a freshman class, and the *Y* scores reflect their rankings in a sophomore class. To what extent do these data form a linear relationship?

Participant	***Fresh.*** **X**	***Soph.*** **Y**
1	2	3
2	9	7
3	1	2
4	5	9
5	3	1
6	7	8
7	4	4
8	6	5
9	8	6

21. A researcher observes the behavior of a group of monkeys in the jungle. He determines each monkey's relative position in the dominance hierarchy of the group (1 being most dominant) and also notes each monkey's relative weight (1 being the lightest). What is the relationship between dominance rankings and weight rankings in these data?

Participant	*Dominance* **X**	*Weight* **Y**
1	1	10
2	2	8
3	5	6
4	4	7
5	9	5
6	7	3
7	3	9
8	6	4
9	8	1
10	10	2

INTEGRATION QUESTIONS

22. In an experiment, (a) which variable is assumed to be the causal variable? (b) Which variable is assumed to be caused? (c) Which variable does the researcher manipulate? (d) Which variable occurs first? (Ch. 2)

23. In a correlational study, we measure participants' creativity and their intelligence. (a) Which variable does the researcher manipulate? (b) Which variable is the causal variable? (c) Which variable occurred first? (d) Which variable is called the independent variable? (Chs. 2, 7)

24. In question 23, (a) How would you determine which variable to call *X*? (b) In a different study, my title is "Creativity as a function of Intelligence." Which variable is my *X* variable? Why? (Ch. 2)

25. Indicate which of the following is a correlational design and the correlation coefficient to compute. (a) We measure participants' age and their daily cell phone usage. (b) We separate participants into three age groups, and then observe their cell phone usage during a one hour period. (c) A teacher uses students' grades on their first exam to predict their final exam grades. (d) We ask whether a website rated as most attractive has more visitors than one rated as second most attractive, and so on, for the top ten websites. (e) We compare performance on an attention test of people who were and were not given an energy drink. (Chs. 2, 7)

SUMMARY OF FORMULAS

1. The formula for the Pearson r is

$$r = \frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{[N(\Sigma X^2) - (\Sigma X)^2][N(\Sigma Y^2) - (\Sigma Y)^2]}}$$

2. The formula for the Spearman r_S is

$$r_S = 1 - \frac{6(\Sigma D^2)}{N(N^2 - 1)}$$