7 Relationship: The Correlation Coefficient

In the preceding chapter the regression coefficient was presented as a measure of association between two variables. The regression coefficient as a measure of association is asymmetric and is expressed in the units of measurement of the variables. It is asymmetric in the sense that its value depends on which variable is considered as the criterion and which is the predictor. It is expressed in the units of the variables in that it measures the amount of change in the criterion as a function of a one-unit change in the predictor.

Sometimes it is not possible to specify which variable is the predictor and which variable is the criterion. For instance, in measuring the degree of relationship between reading comprehension and vocabulary skill in schoolchildren, one variable is not clearly the predictor and the other the criterion. Also, because the regression coefficient is expressed in the units of measurement of the variables, the strength of association is not very clear. It would be desirable to obtain a measure of association that was symmetric and expressed the degree of association between the variables. The correlation coefficient meets both requirements.

The *correlation coefficient* is symbolized by the letter r. Because it is a symmetric measure of association, it follows that $r_{xy} = r_{yx}$. The correlation coefficient is by far the most common measure of association used in the social and behavioral sciences. Only economists use the regression coefficient more frequently than the correlation coefficient. This is no doubt due to the fact that the unit of measurement in economics (the dollar) is readily interpretable.

As an example for this chapter, the variable of laughter in conversations will be considered. Duncan and Fiske (1977) coded the nonverbal behavior of 22 pairs of men and women for five minutes. The couples were instructed to get acquainted with each other. During these conversations, there was occa-

sional laughter. Table 7.1 lists the number of laughs of each person for the 22 couples over the five-minute period.

Rationale for the Correlation Coefficient

The correlation coefficient is a special regression coefficient. Consider the case in which there are two variables, X and Y. First, the scores for the X and Y variables are separately standardized. Thus, Z scores are created for each variable; that is, the mean for the variable is subtracted from each score and then this difference is divided by the variable's standard deviation. To compute the regression coefficient, one Z-scored variable is the predictor and the other is the criterion. The *correlation coefficient*, symbolized by the letter r, is the regression coefficient between two variables whose scores have been standardized.

The correlation coefficient is a symmetric measure of association and so r_{XY} equals r_{YX}. Unlike the regression coefficient, a correlation coefficient has

TABLE 7.1 **Number of Laughs in 22 Conversations**

Couple	Number of Laughs (Women)	Number of Laughs (Men)
1	0	0
2	4	1
3	17	9
4	4	4
5	2	0
6	1	0
7	3	1
8	9	5
9	5	1
10	1	0
11	4	5
12	8	2
13	4	2
14	0	2
15	6	0
16	12	3
17	8	1
18	3	2
19	5	2
20	7	0
21	5	3
22	8	3

Data were taken from Duncan and Fiske (1977).

an upper limit of +1 and a lower limit of −1. A +1 correlation indicates a perfect positive correlation and a −1 correlation indicates a perfect negative correlation. In a perfect correlation, all the points fall on the regression line. The line is ascending if the correlation is +1 and descending if it is −1. Like the regression coefficient, a zero value indicates no *linear* association between the variables. Any nonlinear association may not be reflected by the correlation coefficient.

Computation

As mentioned above, unlike the regression coefficient, the correlation coefficient is a symmetric measure of association: $r_{XY} = r_{YX}$. The relation of r_{XY} to b_{XY} and b_{YX} is straight forward. (Recall from the previous chapter that for b_{XY} the variable Y is the predictor and X the criterion and for b_{YX} the variable X is the predictor and Y the criterion.) The formulas for turning b into r are

$$r_{XY} = b_{XY}\left(\frac{s_Y}{s_X}\right)$$

$$r_{XY} = b_{YX}\left(\frac{s_X}{s_Y}\right)$$

In words, the correlation coefficient equals the regression coefficient times the standard deviation of the predictor divided by the standard deviation of the criterion. It is also true that

$$r_{XY}^2 = b_{XY}b_{YX}$$

To convert from r to b the formulas are

$$b_{XY} = r_{XY}\left(\frac{s_X}{s_Y}\right)$$

$$b_{YX} = r_{XY}\left(\frac{s_Y}{s_X}\right)$$

In words, a regression coefficient equals the correlation times the standard deviation of the criterion divided by the standard deviation of the predictor.

More typically, the correlation is computed directly without computing the regression coefficient. There is also no need to standardize or compute Z scores for each person. The correlation coefficient can be computed by the following formula.

$$r_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

The top term of the above formula is, as referred to in the previous chapter, the sum of cross-products. The denominator is the square root of the product of the sums of squares of both X and Y. A simpler and more practical computational formula is

$$r_{XY} = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sqrt{[\sum X^2 - (\sum X)^2/n][\sum Y^2 - (\sum Y)^2/n]}}$$

The computational formula for correlation has the following ingredients: the sum of the scores for all the subjects on both variables, ΣX and ΣY; the squared sum of the scores, $(\Sigma X)^2$ and $(\Sigma Y)^2$; the sum of each squared score for each variable; ΣX^2 and ΣY^2; and the sum of the product of scores, ΣXY. One common computational error in computing a correlation coefficient is to forget to take the square root of the denominator.

Conventionally correlations are computed to two digits. This is a sensible strategy in that the third digit is not ordinarily interpretable. So if a correlation is to be computed and interpreted, rounding to the second digit should suffice. However, correlation coefficients are often used to compute other statistics, some of which are presented in Chapter 15. If a correlation is to be used to compute other statistics, it should be computed to three or possibly four digits. In this chapter, correlations will be given to three digits.

Table 7.2 displays the computations for the laughing in male-female conversations. The female laughs are denoted as X and male laughs as Y. The sum of cross-products of X and Y is as follows:

$$369 - (116)(46)/22 = 126.4545$$

The sum of squares for X is

$$954 - (116)^2/22 = 342.3636$$

and the sum of squares for Y is

$$198 - (46)^2/22 = 101.8181$$

The correlation then equals

$$r_{XY} = \frac{126.4545}{\sqrt{(342.3636)(101.8181)}} = .677$$

Not surprisingly, there is a very large correlation in the amount of laughter between two persons in a conversation. Laughter is indeed contagious.

Interpretation of r

One way to understand what a correlation of a given size means is to examine various correlations between variables. In Table 7.3 are a set of correlations taken from research. It contains correlations that are small (.1), moderate (.3),

TABLE 7.2 **Computations for Laughing Example**

	X	Y	X^2	Y^2	XY
	0	0	0	0	0
	4	1	16	1	4
	17	9	289	81	153
	4	4	16	16	16
	2	0	4	0	0
	1	0	1	0	0
	3	1	9	1	3
	9	5	81	25	45
	5	1	25	1	5
	1	0	1	0	0
	4	5	16	25	20
	8	2	64	4	16
	4	2	16	4	8
	0	2	0	4	0
	6	0	36	0	0
	12	3	144	9	36
	8	1	64	1	8
	3	2	9	4	6
	5	2	25	4	10
	7	0	49	0	0
	5	3	25	9	15
	8	3	64	9	24
Total	116	46	954	198	369

and large (.5). Small correlations are the most common correlations in the social and behavioral sciences. The reason for so many small correlations is that most variables are caused by numerous factors, and so any one factor's correlation with a variable that it causes must be small. The relation between stress and physical disease, such as heart trouble, and the relation between intelligence and a grade in a course are in the .10 range. A moderate correlation is large enough for laypersons to recognize. An example of moderate correlation is general sense of self-worth and grade point average. Large correlations represent very strong correlations, such as the correlation between intelligence and overall GPA.

It is important to note that a large correlation is not a correlation of .90. Correlations of this size are often between two different measures of the same variable. For instance, the correlation of two measures of intelligence taken a year apart is about .90 once persons are age six or more. Also such large correlations often indicate not a meaningful relationship between variables, but an artificial one. For instance, the .677 correlation between male laughter and female laughter will be seen to be artificially high.

The differences between small, moderate, and large correlations can also be seen in their scatterplots. As explained in Chapter 6, a scatterplot is a graph

TABLE 7.3 **Illustration of Correlations of Various Sizes**

Small: .10

Viewing television violence — Aggressive behavior
Stress — Physical illness
Intelligence — Grade in a particular course

Moderate: .30

Psychotherapy — Adjustment
Self-esteem — Grades in school
Value similarity — Interpersonal attraction

Large: .50

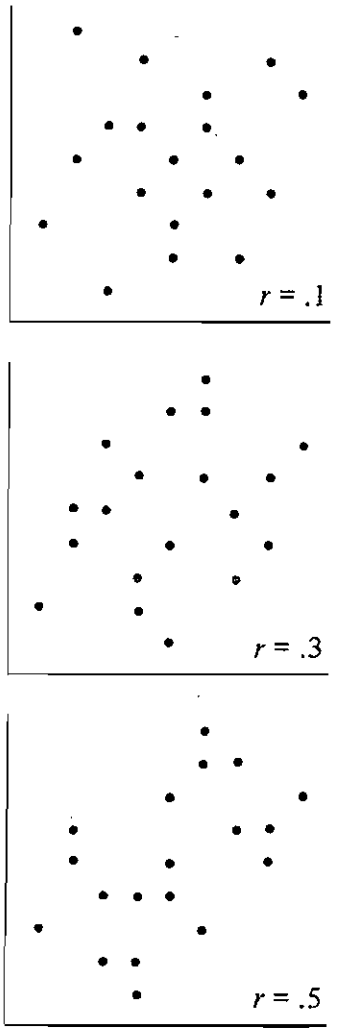
Intelligence — Grade point average
Wife's satisfaction — Husband's satisfaction
Father's occupation — Son's occupation
Belief in God — Church attendance

in which the two variables form the X and Y axes. The pairs of scores for each person are plotted in a scatterplot. The scatterplots for .1, .3, and .5 correlations are presented in Figure 7.1. For a .1 correlation, the correlation is not even visible to the naked eye. For .3 to the trained eye there is the hint of association. For the .5 correlation the linear relationship is clearly visible.

A correlation coefficient is a regression coefficient between standardized scores. It can be directly interpreted then as a regression coefficient between standard scores. If r_{XY} equals .5, then someone who is one standard deviation above the mean on X would tend to be .5 standard deviation units above the mean on Y. So a correlation between X and Y measures how many standard deviation units above or below the mean a person's score is on Y when the person is one standard deviation above the mean on X. Because it is a symmetric measure, it can also be interpreted as the predicted value for X for someone who is one standard deviation above the mean on Y.

The most common way to interpret a correlation coefficient is by squaring the correlation and interpreting the result as the proportion of variance that the two variables share in common. The proportion can be multiplied by 100 to obtain the percent of shared variance. So, for instance, if high school grades and college grades correlate .6, then $.6^2$ or .36 of their variance is shared in common. Besides shared variance, the squared correlation can also be interpreted as the proportion of variance explained. So a .6 correlation between high school grades and college grades implies that high school grades can explain .36 of the total variance in college grades. The squared correlation for shared or explained variance is often used to trivialize small correlations. For instance, a .1 correlation represents only .01 shared variance. It should be noted that the squared correlation represents shared or explained *variance* and

FIGURE 7.1 **Scatterplots of .1, .3, and .5 correlation coefficients.**



not standard deviation. Because variance is in squared units, the meaning of explained variance may be difficult to appreciate. For instance, if intelligence explains 25% of the variance in high school grades, it means that intelligence explains 25% of squared grade points.

A correlation can be viewed in terms of a probability. Consider two persons, one, called A, who is one standard deviation *above* the mean on X and the other, called B, who is one standard deviation *below* the mean on X.

Person A has a two-standard-deviation advantage on X over person B. If the correlation between X and Y is known, then the probability that person A scores higher than B on Y can be determined. For instance, assume that variable X is education and variable Y is income. Let one standard deviation above the mean on education be a master's degree and one standard deviation below the mean be a high school education. The issue is the probability of someone who has a master's degree earning more money than someone who has only graduated from high school. This probability will be referred to as the *two-standard-deviation advantage* and abbreviated as the 2sd advantage.

To determine the probability, it is assumed that both variables are normally distributed. The normal distribution is discussed in Chapter 10. Rosenthal and Rubin (1979) make radically different distributional assumptions, yet for r between 0 and .5 they obtain virtually the same result. (They assume that X and Y are dichotomies as opposed to normally distributed variables measured at the interval level of measurement.)

In Table 7.4 are the 2sd advantage probabilities for correlations of various sizes.[1] So for instance, if r is .45, then the probability that someone who is one standard deviation above the mean on X will score on Y above the person who is one standard deviation below the mean on X is .762. If the correlation is negative, the probabilities in the table can be read as the probability of someone one standard deviation above the mean on X scoring *below* someone one standard deviation below the mean on X.

The table is read as follows. First, find the correlation to be interpreted in the r column. Second, the value in the probability column states the probability that a person who is one standard deviation above the mean will outscore

TABLE 7.4 **Correlation in Terms of the Two-Standard-Deviation Advantage**

r	Probability[a]	r	Probability
.00	.500	.50	.793
.05	.528	.55	.824
.10	.557	.60	.856
.15	.585	.65	.887
.20	.614	.70	.917
.25	.642	.75	.945
.30	.672	.80	.970
.35	.701	.85	.989
.40	.731	.90	.998
.45	.762	.95	1.000

[a] The probability of a person who is one standard deviation above the mean on X scoring higher on Y than someone who is one standard deviation below the mean on X.

[1]The 2sd advantage can be shown to equal the probability that Z is less than $\sqrt{2}\rho/\sqrt{1-\rho^2}$ where ρ is the population correlation and Z is a standard normal variable.

someone else on Y who is one standard deviation below the mean on X. If the correlation is zero, the 2sd (two-standard-deviation) advantage is .5. That is, a person with a 2sd advantage on X over another person has only a 50/50 chance of outscoring the other person. Even a seemingly low correlation like .2 carries with it an impressive probability of .614. Thus, for a correlation of .2, over 60% of the time person A (who is one standard deviation above the mean on X) will outscore person B (who is one standard deviation below the mean of X) on Y.

The 2sd advantages for small, medium, and large correlations are:

Small (r = .1): .557
Medium (r = .3): .672
Large (r = .5): .793

Factors Affecting the Size of r

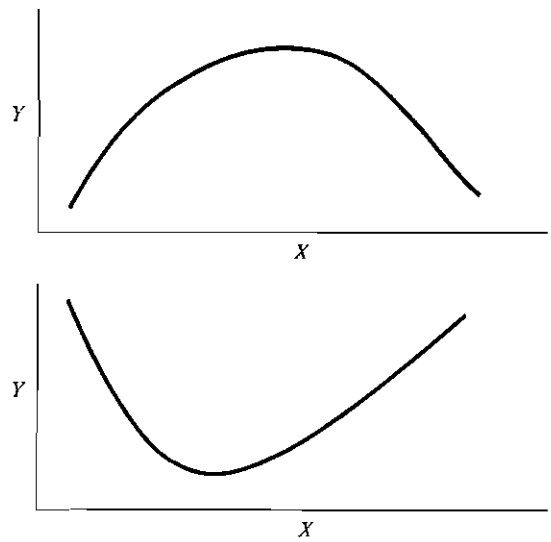
Special care must be taken in interpreting correlation and regression coefficients. At times, a coefficient can be artificially too small or too large. Various factors are discussed below that must be considered when interpreting measures of association, especially correlation coefficients.

Nonlinearity

The fundamental definition of a regression coefficient is that of a slope of the straight line fitted to a set of points. A correlation coefficient is the slope of the line when the two samples have been converted into Z scores. Both measures of association assume that the line to be fitted is *straight* and not curved. The association between variables may be systematic, but it need not be linear. There are two major types of nonlinear associations. They are nonlinear association in which the function changes direction and nonlinear association in which the function does not change direction.

In Figure 7.2 are examples of changes of direction. In the top diagram of the figure, the relationship starts as positive and then turns negative. In the bottom diagram, the relationship starts negative and then turns positive. Both of these patterns are called *curvilinear* association. More precisely, a relationship that begins as positive and turns negative (the upper half of the figure) is called a *convex curvilinear* or an *inverted U* relationship. And a relationship that begins as negative and turns positive (the bottom half of the figure) is called a *concave curvilinear* or *U-shaped* relationship. For either type of curvilinear association both the correlation and regression coefficient can be quite misleading measures of association. These measures may well be zero even when there is a strong curvilinear association. As an example of a

FIGURE 7.2 **Examples of curvilinear relationships.**

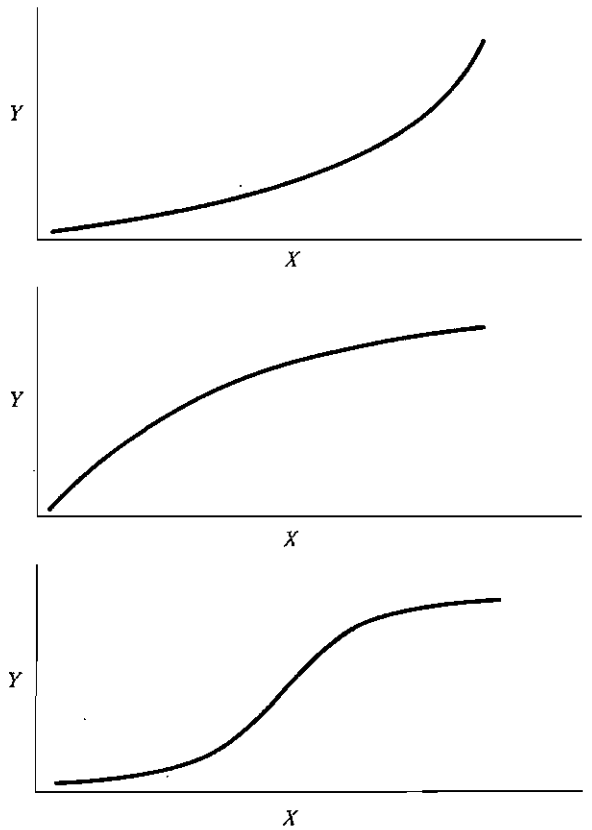


concave curvilinear association, amount of leisure time is curvilinearly related to age, with older and younger persons having more leisure time than middle-aged persons.

If the researcher expects that two variables are curvilinearly associated, the scatterplot should be carefully examined. If the point at which the relationship changes direction can be determined, a linear measure of association can be computed before and after that point. If the relationship is truly curvilinear, then one relationship should be positive and the other negative. For instance, if a researcher expects that the amount of leisure time begins to increase at age 45, then the correlation between age and leisure time should be negative for those under the age of 45 and positive for those who are 45 and older.

For the second type of nonlinear association, the relationship does not change in direction. This pattern is illustrated in Figure 7.3. For these relationships the direction of the relationship does not change but the strength does. For the three examples in Figure 7.3 the relationship is positive; that is, as X increases, Y increases. For the top diagram in the figure as X increases, the relationship between X and Y increases. This is an accelerating function. For the middle diagram in the figure, as X increases, the relationship between X and Y decreases. This is a decelerating function. For the bottom diagram in the figure, as X increases, the relationship first increases and then it decreases. This type of relationship is called *S-shaped*. Correlation and regression coefficients are less affected by this form of nonlinearity than the form in

FIGURE 7.3 **Examples of nonlinear relationships that do not change direction.**



which the relationship changes direction. Nonetheless, it is important to attempt to straighten out the relationship.

Nonlinear relationships that do not change direction can be turned into linear relationships by transformation. For instance, for the pattern in the top of Figure 7.3, the relationship can be made more linear by applying a one-stretch transformation (square root, logarithm, or reciprocal) to the *Y* variable. For the pattern in the middle of Figure 7.3, the relationship can be made more linear by applying a one-stretch transformation (square root, logarithm, or reciprocal) to the *X* variable. For the pattern in the bottom of the figure, the relationship can be made more linear by applying a two-stretch transformation (arcsin, logit, or probit) to the *Y* variable. If the researcher cannot specify the exact type of transformation, then Spearman's rank-order correlation (discussed in the next chapter) may be a more appropriate measure of association for any nonlinear relationship that does not change direction.

Unreliability

Measurement in the social and the behavioral sciences is imperfect. Although every effort is made to measure persons as accurately as possible, unintentional errors of measurement are inevitable. Measurement involves not only what the researcher hopes to be measuring but noise or error as well. The first component is called the *true score* and the second is called *error of measurement*. The percent of variance of a measure that is due to the true score is called *reliability*. Constructs that social and behavioral scientists measure hardly ever have perfect reliability. Even a variable such as age has error due to distortion (people lying) and rounding. It is not at all unusual for a personality test to have a reliability of .80. A reliability of .80 means that 20% of the variance in the test is attributable to error.

Less than perfect reliability in a measure affects the size of the correlation and regression coefficients. The effect is one of attenuation. That is, the estimated size of the coefficient is nearer to zero than it ought to be. A regression coefficient is lowered only when the predictor variable is unreliable. Correlations are attenuated when either variable has less than perfect reliability.

Aggregation

Sometimes researchers average the scores of a group of persons and use these averages as the basic data. For instance, students in the classroom are averaged and the basic analysis is on the classroom averages. When scores are averaged across persons, the data are called *aggregate data*. Generally, correlations computed from aggregate data are larger than what they would be if the individual scores were used. This increase is in part due to increased reliability, because aggregate data are generally more reliable than individual data. Though less likely, aggregation can reduce the size of a correlation.

Because a correlation computed from scores aggregated across persons can be quite different from a correlation of individual scores, one should never interpret the aggregated correlation as if it were the correlation from individuals. To do so would be what is called the *ecological fallacy*. An example of the ecological fallacy would be to correlate precinct voting data to make inferences about individual voting patterns. Correlations computed using aggregates (precincts) may not resemble correlations based on individuals (voters).

Part-Whole Correlation

A correlation involves two variables. Sometimes one of the variables is derived from the other variable. When one variable is derived from a second

variable, there can be a built-in correlation between the two. The variables must share variance because one is part of the other. In Table 7.5 the variable X is used to derive a second variable, and the direction of bias is indicated. If the direction of bias is indicated in the table as positive, it does not mean that the correlation is necessarily positive, but that the correlation is larger than it should be.

In the first case in Table 7.5, the variable X is used to derive the measure $X + Y$. Because X is present in both measures, there is a built-in positive correlation. In the second case X is subtracted from Y. In this case the correlation is negative. One should avoid computing correlations between variables that have common components.

Restriction in Range

Correlations computed from scores that have low variability generally tend to be small. This phenomenon is called *restriction of range*. It can be illustrated graphically as in Figure 7.4. The data in the figure show that the X variable has been split at the mean and the correlation has been recomputed for those scoring above and below the mean. Overall the correlation is .533, but for those who score below the mean (as is indicated by the dashed line in Figure 7.4) the correlation is .341 and for those who score above the mean, the correlation is also .341. When a variable has a narrow range of scores, correlations tend to be small.

Interestingly, restriction in range does not influence the regression coefficient nearly as much as it does the correlation coefficient. So if the range of a variable may be restricted, the regression coefficient is the preferred measure of association.

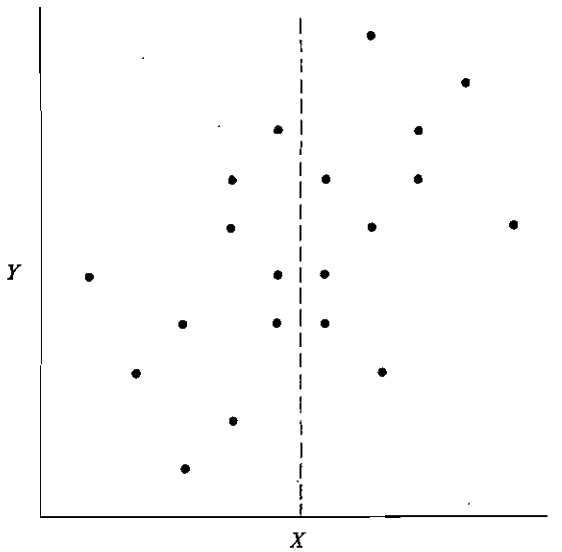
Outliers

Extreme values or outliers in the sample can distort the size of a correlation. For instance, for the following set of data

TABLE 7.5 **Part-Whole Correlation**

Variable 1	Variable 2	Bias in r
X	$X + Y$	Positive
X	$Y - X$	Negative
$X + Z$	$X + Y$	Positive
$X + Z$	$Y - X$	Negative
Y/X	W/X	Positive

FIGURE 7.4 **Effect of restriction of range.**



Person	*X*	*Y*
1	3	3
2	2	2
3	1	1
4	4	4
5	10	1

the correlation is negative even though four of the five persons have the same score on X and Y. The extreme value of 10 for person 5 on variable X distorts the size of the correlation. Outliers can also cause a correlation that is truly zero to appear to be very large. A careful analysis of each variable should be done to identify outliers. In Chapter 4 an outlier is defined as a value that is away from the median by more than twice the interquartile range.

It is an outlier that brings about the very large correlation of .677 for the laughing data in Table 7.1. Note in Table 7.1 that couple 3 has the largest number of laughs for both males and females. In fact, each can be considered an outlier given the definition given in Chapter 4. What happens to the correlation coefficient when the data from this one couple is discarded? The resulting correlation is

$$r_{XY} = \frac{216 - (99)(37)/21}{\sqrt{(665 - 99^2/21)(117 - 37^2/21)}} = .410$$

Dropping this one observation changes what was an unreasonably large correlation into a moderate-to-large correlation. Besides dropping the one observation, an alternative would be to transform the observations. An examination of the histograms for the observations reveals that both variables are positively skewed. If the observations are square rooted, the scores for couple 3 are no longer outliers. The resulting correlation of the square rooted data is .551.

Correlation and Causality

Correlations by their very nature seem to give rise to causal statements. If a newspaper publishes a report that persons who eat carrots live longer, it is a certainty that more carrots will be sold the following day. Finding out that carrot eating and longevity are associated inclines persons to jump to the conclusion that carrot consumption causes longer life. But correlation does not imply a particular causal relation. Just knowing that carrot eating and long life are associated does not mean that carrot eating causes longer life. There are other equally plausible explanations of the relationship. For instance, it may be that persons with more income tend both to live longer and also to eat carrots. And so the relationship between carrot eating and longevity may be due to the third variable of income.

Most of the time correlation *does* imply causality, but the exact form of the causality is uncertain. Consider another example. There is a small-to-moderate positive correlation between preference for violent television programs and the tendency to be physically and verbally aggressive among preadolescent males. Thus, boys who get into fights prefer to watch Kojak and the Three Stooges. The reason for this correlation is not clear. It could be that the violence on television makes the children more aggressive. Or it could be that being aggressive makes boys seek out more violent television shows. Or it may be that neither causes the other but both are caused by some other variable. For instance, parental socialization may affect both television viewing and aggressive behavior. It might be that authoritarian parental rearing leads to aggressive boys who watch violent television shows. Thus, knowing that there is a correlation between two variables does not tell us what brought about the correlation. As is often stated, "correlation does not imply causality." It is better to restate the maxim as "correlation does not imply one particular form of causality."

Sometimes the source of correlation is not a causal process but is just an accident. For instance, there is some indication that the economic climate is negatively correlated with the length of women's skirts. Good economic times have been associated with shorter skirts and bad times with longer skirts. Surely skirt length does not cause the financial climate. Nor is it likely that the

financial climate causes the length of skirts. Most likely this correlation is an accident, a statistical freak. One way to determine whether the correlation is just an accident is to check its continuation into the future. If it disappears, then it is likely an accident.

One should not take all that has been said to mean that correlations tell us nothing about causation. It is true that from correlations it is not possible to determine the particular causal connections. But if there is reason to believe that one variable causes the other, then the two variables should be correlated. Thus, a correlation can be used to *verify* a causal linkage, but it is indeed perilous to infer a particular causal linkage from a correlation. Thus, causation implies a correlation but correlation does not specify the exact form of causality.

Summary

The correlation between two variables is defined as the regression coefficient computed from two variables' Z scores. The correlation coefficient, symbolized by r, is a directionless measure of association that varies between -1 and $+1$.

The formula for a correlation coefficient is

$$r_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(\bar{X} - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

The computational formula for a correlation coefficient is

$$r_{XY} = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sqrt{[\sum X^2 - (\sum X)^2/n][\sum Y^2 - (\sum Y)^2/n]}}$$

A small correlation is .1, medium is .3, and large is .5. A correlaton can be interpreted as a regression coefficient. A squared correlation indicates the proportion of variance explained or variance shared. A correlation can also be interpreted as a probability of someone with a two-standard-deviation advantage over another person on one variable outscoring that person on a second variable.

Correlations are affected by nonlinearity, unreliability, aggregation, restriction in range, the part-whole problem, and outliers.

Just because two variables are correlated does not mean that one causes the other. It may be that the two variables are both caused by a third variable. A correlation indicates some type of causal connection but does not identify the particular type.

Problems

1. For the data

X	Y
1	7
2	9
4	13
3	11

compute r_{XY}.

2. Smith finds that the correlation between motivation and performance equals .391. How would you help her interpret her result?

3. If

X	Y
−3	1
9	9
1	15

compute r_{XY}.

4. Compute the following.

   a. b_{XY} given that $r_{XY} = .4$, $s_X = 2$, $s_Y = 4$
   b. b_{XY} given that $r_{XY} = -.3$, $s_X = 10$, $s_Y = 3$
   c. r_{XY} given that $b_{XY} = .1$, $s_X = s_Y$

5. Baxter (1972) used an adaptation of traditional methods to teach clerical skills to mildly retarded adults. At the end of training, the skill of these adults in each task was rated by the same standards. The scale ranged from zero to ten, with zero the lowest possible rating and ten the highest. Some of Baxter's results are given below.

Subject	*Typing*	*Stencils*
1	3	4
2	8	7
3	6	7
4	5	3
5	6	7
6	2	7
7	6	9
8	6	5
9	9	9
10	6	9

Compute the correlation between typing and stencil preparation. Interpret the result.

6. Compute r if $n = 10$, $\Sigma X = 20$, $\Sigma Y = 40$, $\Sigma X^2 = 60$, $\Sigma Y^2 = 180$, and $\Sigma XY = 90$.

7. For the age and memory data presented in Table 6.3 of the preceding chapter, compute and interpret the correlation coefficient.

8. Below are the life expectancies at birth for males (X) and females (Y) in six of the less developed countries of the world. Also given are ΣXY, ΣX, ΣY, ΣX^2, and ΣY^2. Compute the correlation between the life expectancies of males and that of females.

	Life Expectancy (years)	
Country	*Males (X)*	*Females (Y)*
India	51	50
Indonesia	45	48
Brazil	58	63
Bangladesh	50	47
Pakistan	49	47
Nigeria	40	43
Total	293	298

$\Sigma XY = 14{,}737$; $\Sigma X^2 = 14{,}491$; $\Sigma Y^2 = 15{,}040$

9. Below are the life expectancies at birth for males (X) and females (Y) in six of the more developed countries of the world. Also given are ΣXY, ΣX, ΣY, ΣX^2, and ΣY^2. Compute the correlation of the life expectancies of males and females.

	Life Expectancy (years)	
Country	*Males (X)*	*Females (Y)*
France	70	78
U.S.A.	70	78
Japan	73	79
W. Germany	70	76
Italy	70	76
United Kingdom	70	76
Total	423	463

$\Sigma XY = 32{,}647$; $\Sigma X^2 = 29{,}829$; $\Sigma Y^2 = 35{,}737$

10. Below are the life expectancies for males and females in the twelve countries given in problems 8 and 9.

	Life Expectancy (years)	
Country	*Males (X)*	*Females (Y)*
India	51	50
Indonesia	45	48
Brazil	58	63
Bangladesh	50	47
Pakistan	49	47
Nigeria	40	43
France	70	78
U.S.A.	70	78
Japan	73	79
W. Germany	70	76
Italy	70	76
United Kingdom	70	76
Total	716	761

$\Sigma XY = 47{,}384$; $\Sigma X^2 = 44{,}320$; $\Sigma Y^2 = 50{,}777$

a. Compute the correlation of life expectancies for males and females.
b. Compare this correlation to the correlation obtained in problems 8 and 9. What caused the correlation to change?

11. Draw a scatterplot for the data in problem 12 in Chapter 6. Describe the nonlinearity in the relationship and suggest a transformation to remove it.

12. For the following studies, state what might affect the size of the correlation, and explain how the correlation would be affected.

a. using the average score of children in 500 schools, the correlation between vocabulary and reading comprehension
b. the correlation between a child's height at birth and growth in the first year of life
c. the correlation of stress in the workplace with physical ailments among air traffic controllers
d. the correlation between number of calories ingested during the day and happiness
e. the correlation between intelligence, as measured by one item of an IQ test, and a student's grade-point average