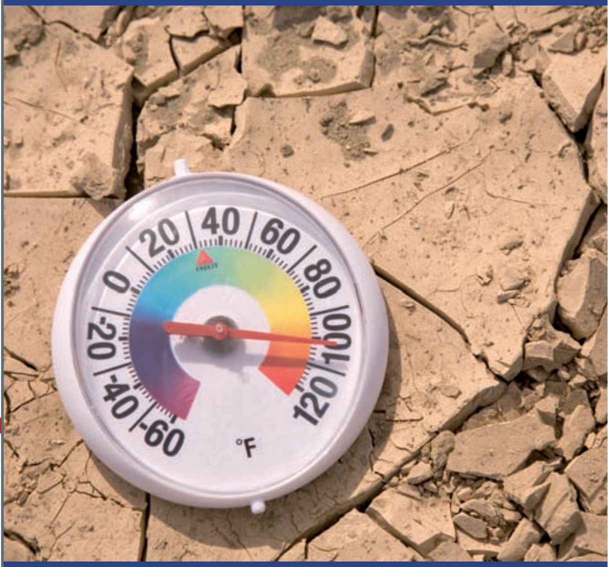
This page intentionally left blank

10

Correlation and Regression



STATISTICS TODAY

Can Temperature Predict Crime?

Over the last years, researchers have been interested in the relationship between increasing temperatures and increasing crime rates. To test this relationship, the author selected a city on the East Coast and obtained the average monthly temperatures for that city as well as the number of crimes committed each month for the year 2011. The data are shown.

Month	Average temperature	Total offenses
January	36	83
February	35	82
March	42	81
April	52	102
May	60.5	122
June	71.5	117
July	77	126
August	77.5	115
September	73	84
October	63	123
November	53	82
December	45	102

Source: City of Annapolis, Maryland, Police Department and www.average-temperature.com

Using the statistical methods described in this chapter, you will be able to answer these questions:

1. Is there a linear relationship between the monthly average temperatures and the number of crimes committed during the month?
2. If so, how strong is the relationship between the average monthly temperature and the number of crimes committed?
3. If a relationship exists, can it be said that an increase in temperatures will cause an increase in the number of crimes occurring in that city?

See Statistics Today—Revisited at the end of the chapter for the answers to these questions.

OUTLINE

OBJECTIVES

After completing this chapter, you should be able to

1. Draw a scatter plot for a set of ordered pairs.
2. Compute the correlation coefficient.
3. Test the hypothesis H_0: $\rho = 0$.
4. Compute the equation of the regression line.
5. Compute the coefficient of determination.
6. Compute the standard error of the estimate.
7. Find a prediction interval.
8. Be familiar with the concept of multiple regression.

Introduction

In Chapters 7 and 8, two areas of inferential statistics—confidence intervals and hypothesis testing—were explained. Another area of inferential statistics involves determining whether a relationship exists between two or more numerical or quantitative variables. For example, a businessperson may want to know whether the volume of sales for a given month is related to the amount of advertising the firm does that month. Educators are interested in determining whether the number of hours a student studies is related to the student's score on a particular exam. Medical researchers are interested in questions such as, Is caffeine related to heart damage? or Is there a relationship between a person's age and his or her blood pressure? A zoologist may want to know whether the birth weight of a certain animal is related to its life span. These are only some of the many questions that can be answered by using the techniques of correlation and regression analysis.

The purpose of this chapter then is to answer these questions statistically:

1. Are two or more variables linearly related?
2. If so, what is the strength of the relationship?
3. What type of relationship exists?
4. What kind of predictions can be made from the relationship?

Unusual Stat

A person walks on average 100,000 miles in his or her lifetime. This is about 3.4 miles per day.

10–1 Scatter Plots and Correlation

OBJECTIVE 1

Draw a scatter plot for a set of ordered pairs.

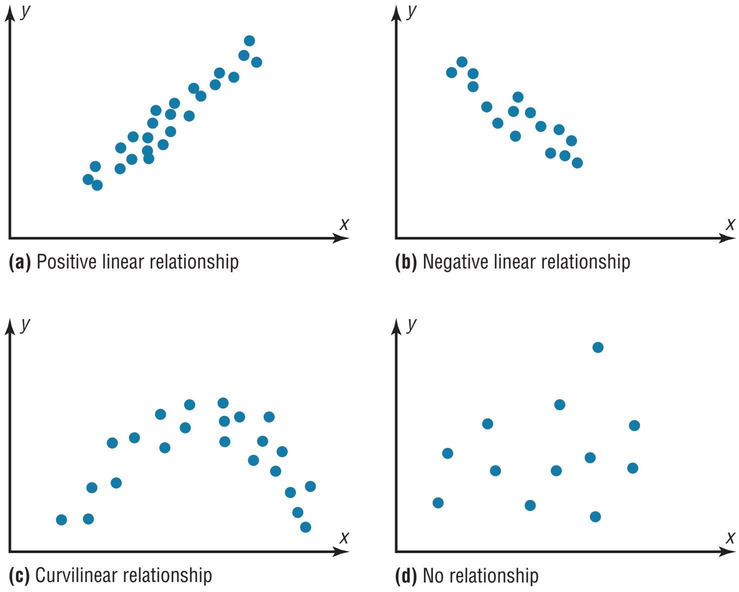
In simple correlation and regression studies, the researcher collects data on two numerical or quantitative variables to see whether a relationship exists between the variables. For example, if a researcher wishes to see whether there is a relationship between number of hours of study and test scores on an exam, she must select a random sample of students, determine the number of hours each studied, and obtain their grades on the exam. A table can be made for the data, as shown here.

Student	Hours of study *x*	Grade *y* (%)
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

The two variables for this study are called the **independent variable** and the **dependent variable.** The independent variable is the variable in regression that can be controlled or manipulated. In this case, the number of hours of study is the independent variable and is designated as the *x* variable. The dependent variable is the variable in regression that cannot be controlled or manipulated. The grade the student received on the exam is the dependent variable, designated as the *y* variable. The reason for this distinction between the variables is that you assume that the grade the student earns *depends* on the number of hours the student studied. Also, you assume that, to some extent, the student can regulate or *control* the number of hours he or she studies for the exam. The independent variable is also known as the *explanatory variable,* and the dependent variable is also called the *response variable*.

The determination of the *x* and *y* variables is not always clear-cut and is sometimes an arbitrary decision. For example, if a researcher studies the effects of age on a person's blood pressure, the researcher can generally assume that age affects blood pressure. Hence, the variable *age* can be called the *independent variable,* and the variable *blood pressure* can be called the *dependent variable.* On the other hand, if a researcher is studying the attitudes of husbands on a certain issue and the attitudes of their wives on the same issue, it is difficult to say which variable is the independent variable and which is the dependent variable. In this study, the researcher can arbitrarily designate the variables as independent and dependent.

FIGURE 10–1
Types of Relationships



(a) Positive linear relationship
(b) Negative linear relationship
(c) Curvilinear relationship
(d) No relationship

The independent and dependent variables can be plotted on a graph called a *scatter plot.* The independent variable *x* is plotted on the horizontal axis, and the dependent variable *y* is plotted on the vertical axis.

A **scatter plot** is a graph of the ordered pairs (*x, y*) of numbers consisting of the independent variable *x* and the dependent variable *y.*

The scatter plot is a visual way to describe the nature of the relationship between the independent and dependent variables. The scales of the variables can be different, and the coordinates of the axes are determined by the smallest and largest data values of the variables.

Researchers look for various types of patterns in scatter plots. For example, in Figure 10–1(a), the pattern in the points of the scatter plot shows a *positive linear relationship.* Here, as the values of the independent variable (*x* variable) increase, the values of the dependent variable (*y* variable) increase. Also, the points form somewhat of a straight line going in an upward direction from left to right.

The pattern of the points of the scatter plot shown in Figure 10–1(b) shows a *negative linear relationship.* In this case, as the values of the independent variable increase, the values of the dependent variable decrease. Also, the points show a somewhat straight line going in a downward direction from left to right.

The pattern of the points of the scatter plot shown in Figure 10–1(c) shows some type of a nonlinear relationship or a curvilinear relationship.

Finally, the scatter plot shown in Figure 10–1(d) shows basically no relationship between the independent variable and the dependent variable since no pattern (line or curve) can be seen.

The procedure table for drawing a scatter plot is given next.

Procedure Table

Drawing a Scatter Plot

Step 1 Draw and label the *x* and *y* axes.

Step 2 Plot each point on the graph.

Step 3 Determine the type of relationship (if any) that exists for the variables.

The procedure for drawing a scatter plot is shown in Examples 10–1 through 10–3.

EXAMPLE 10–1 Car Rental Companies

Construct a scatter plot for the data shown for car rental companies in the United States for a recent year.

Company	Cars (in ten thousands)	Revenue (in billions)
A	63.0	$7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

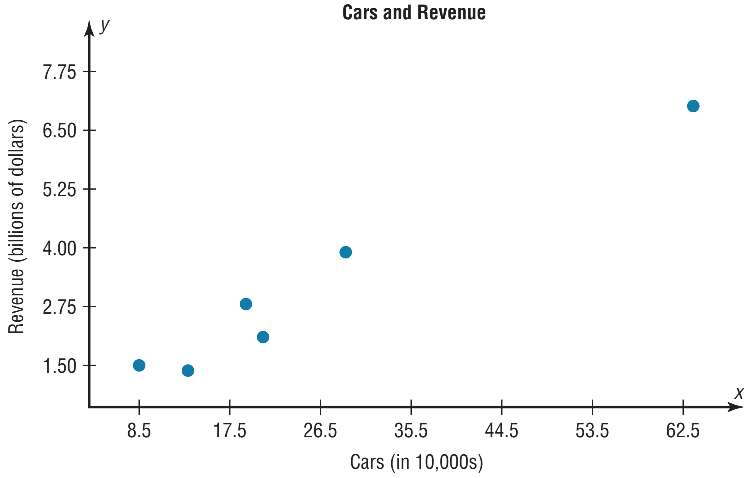
Source: *Auto Rental News.*

SOLUTION

Step 1 Draw and label the *x* and *y* axes.

Step 2 Plot each point on the graph, as shown in Figure 10–2.

FIGURE 10–2 Scatter Plot for Example 10–1



Step 3 Determine the type of relationship (if any) that exists.

In this example, it looks as if a positive linear relationship exists between the number of cars that an agency owns and the total revenue that is made by the company.

EXAMPLE 10–2 Absences and Final Grades

Construct a scatter plot for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class. The data are shown here.

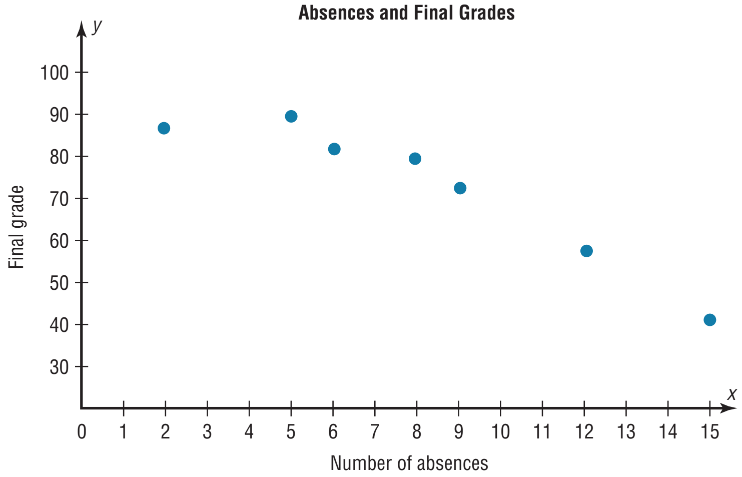
Student	Number of absences *x*	Final grade *y* (%)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

SOLUTION

Step 1 Draw and label the *x* and *y* axes.

Step 2 Plot each point on the graph, as shown in Figure 10–3.

FIGURE 10–3 Scatter Plot for Example 10–2



Step 3 Determine the type of relationship (if any) that exists.

In this example, it looks as if a negative linear relationship exists between the number of student absences and the final grade of the students.

EXAMPLE 10–3 Age and Wealth

A researcher wishes to see if there is a relationship between the ages of the wealthiest people in the world and their net worth. A random sample of 10 persons was selected from the *Forbes* list of the 400 richest people for a recent year. The data are shown. Draw a scatter plot for the data.

Person	Age *x*	Net worth *y* (in billions of dollars)
A	60	11
B	72	69
C	56	11.9
D	55	30
E	83	12.2
F	67	36
G	38	18.7
H	62	10.2
I	62	23.3
J	46	10.6

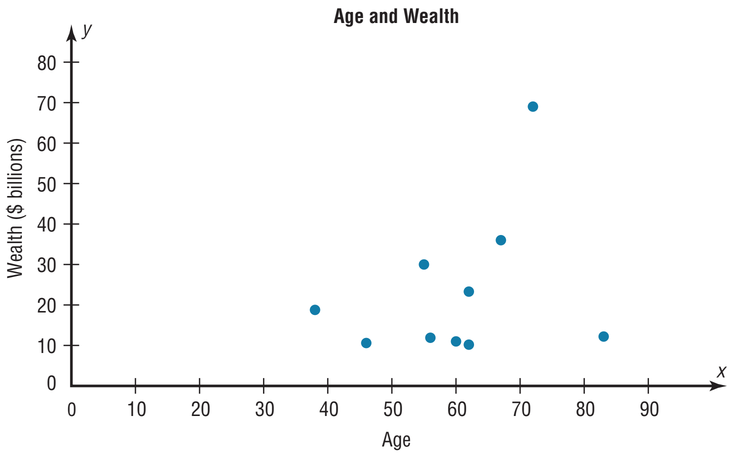
Source: *Forbes* magazine.

SOLUTION

Step 1 Draw and label the *x* and *y* axes.

Step 2 Plot each point on the graph, as shown in Figure 10–4.

FIGURE 10–4 Scatter Plot for Example 10–3



Step 3 Determine the type of relationship (if any) that exists.

In this example, there is no type of a strong linear or curvilinear relationship between a person's age and his or her net worth.

Correlation

OBJECTIVE 2

Compute the correlation coefficient.

Correlation Coefficient

Statisticians use a measure called the *correlation coefficient* to determine the strength of the linear relationship between two variables. There are several types of correlation coefficients.

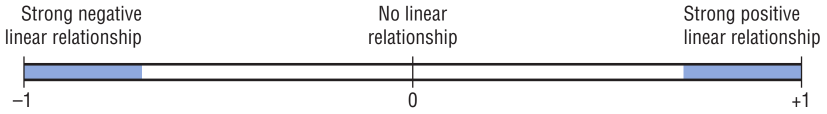
The **population correlation coefficient** denoted by the Greek letter ρ is the correlation computed by using all possible pairs of data values (x, y) taken from a population.

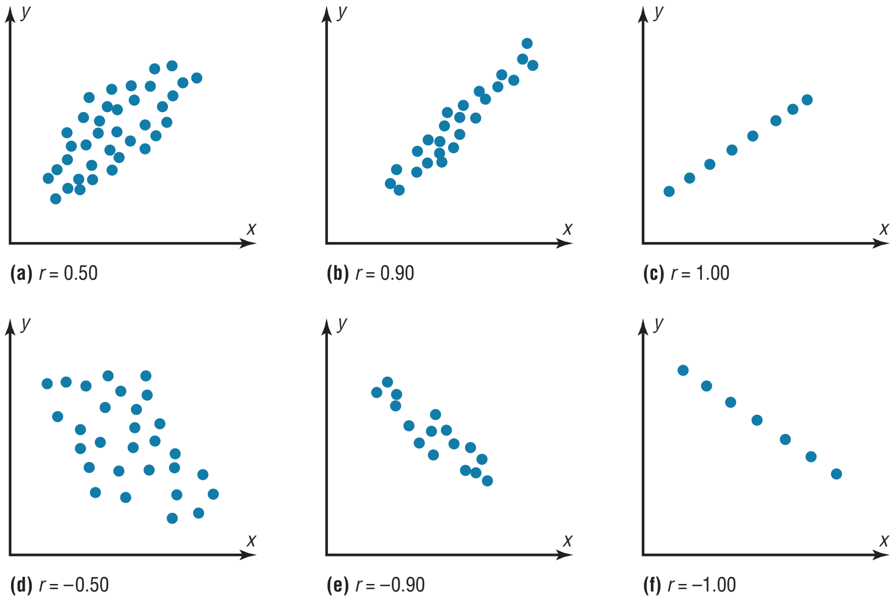
The **linear correlation coefficient** computed from the sample data measures the strength and direction of a linear relationship between two quantitative variables. The symbol for the sample correlation coefficient is r.

The linear correlation coefficient explained in this section is called the **Pearson product moment correlation coefficient (PPMC),** named after statistician Karl Pearson, who pioneered the research in this area.

The *range of the linear correlation coefficient* is from -1 to $+1$. If there is a *strong positive linear relationship* between the variables, the value of r will be close to $+1$. If there is a *strong negative linear relationship* between the variables, the value of r will be close to -1. When there is no linear relationship between the variables or only a weak relationship, the value of r will be close to 0. See Figure 10–5. When the value of r is 0 or close to zero, it implies only that there is no linear relationship between the variables. The data may be related in some other nonlinear way.

FIGURE 10–5 Range of Values for the Correlation Coefficient





(a) $r = 0.50$ **(b)** $r = 0.90$ **(c)** $r = 1.00$

(d) $r = -0.50$ **(e)** $r = -0.90$ **(f)** $r = -1.00$

FIGURE 10–6
Relationship Between the Correlation Coefficient and the Scatter Plot

Properties of the Linear Correlation Coefficient

1. The correlation coefficient is a unitless measure.
2. The value of r will always be between -1 and $+1$ inclusively. That is, $-1 \leq r \leq 1$.
3. If the values of x and y are interchanged, the value of r will be unchanged.
4. If the values of x and/or y are converted to a different scale, the value of r will be unchanged.
5. The value of r is sensitive to outliers and can change dramatically if they are present in the data.

The graphs in Figure 10–6 show the relationship between the correlation coefficients and their corresponding scatter plots. Notice that as the value of the correlation coefficient increases from 0 to $+1$ (parts *a*, *b*, and *c*), data values become closer to a straight line and to an increasingly strong relationship. As the value of the correlation coefficient decreases from 0 to -1 (parts *d*, *e*, and *f*), the data values also become closer to a straight line. Again this suggests a stronger relationship.

Assumptions for the Correlation Coefficient

1. The sample is a random sample.
2. The data pairs fall approximately on a straight line and are measured at the interval or ratio level.
3. The variables have a bivariate normal distribution. (This means that given any specific value of x, the y values are normally distributed; and given any specific value of y, the x values are normally distributed.)

In this book, the assumptions will be stated in the exercises; however, when encountering statistics in other situations, you must check to see that these assumptions have been met before proceeding.

There are several ways to compute the value of the correlation coefficient. One method is to use the formula shown here.

Formula for the Linear Correlation Coefficient *r*

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

where n is the number of data pairs.

Rounding Rule for the Correlation Coefficient Round the value of r to three decimal places.

The formula looks somewhat complicated, but using a table to compute the values, as shown in Example 10–4, makes it somewhat easier to determine the value of r.

There are no units associated with r, and the value of r will remain unchanged if the x and y values are switched.

The procedure for finding the value of the linear correlation coefficient is given next.

Procedure Table

Finding the Value of the Linear Correlation Coefficient

Step 1 Make a table as shown.

x	y	xy	x^2	y^2

Step 2 Place the values of x in the x column and the values of y in the y column. Multiply each x value by the corresponding y value, and place the products in the xy column.

Square each x value and place the squares in the x^2 column.

Square each y value and place the squares in the y^2 column.

Find the sum of each column.

Step 3 Substitute in the formula and find the value for r.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

where n is the number of data pairs.

EXAMPLE 10–4 Car Rental Companies

Compute the linear correlation coefficient for the data in Example 10–1.

SOLUTION

Step 1 Make a table as shown here.

Company	Cars x (in ten thousands)	Revenue y (in billions)	xy	x^2	y^2
A	63.0	$7.0			
B	29.0	3.9			
C	20.8	2.1			
D	19.1	2.8			
E	13.4	1.4			
F	8.5	1.5			

Step 2 Find the values of xy, x^2, and y^2, and place these values in the corresponding columns of the table.

The completed table is shown.

Company	Cars x (in 10,000s)	Revenue y (in billions of dollars)	xy	x^2	y^2
A	63.0	7.0	441.00	3969.00	49.00
B	29.0	3.9	113.10	841.00	15.21
C	20.8	2.1	43.68	432.64	4.41
D	19.1	2.8	53.48	364.81	7.84
E	13.4	1.4	18.76	179.56	1.96
F	8.5	1.5	12.75	72.25	2.25
	$\Sigma x = 153.8$	$\Sigma y = 18.7$	$\Sigma xy = 682.77$	$\Sigma x^2 = 5859.26$	$\Sigma y^2 = 80.67$

Step 3 Substitute in the formula and solve for r.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{[(6)(5859.26) - (153.8)^2][(6)(80.67) - (18.7)^2]}} = 0.982$$

The linear correlation coefficient suggests a strong positive linear relationship between the number of cars a rental agency has and its annual revenue. That is, the more cars a rental agency has, the more annual revenue the company will have.

EXAMPLE 10–5 Absences and Final Grades

Compute the value of the linear correlation coefficient for the data obtained in the study of the number of absences and the final grade of the seven students in the statistics class given in Example 10–2.

SOLUTION

Step 1 Make a table.

Step 2 Find the values of xy, x^2, and y^2; place these values in the corresponding columns of the table.

Student	Number of absences x	Final grade y (%)	xy	x^2	y^2
A	6	82	492	36	6,724
B	2	86	172	4	7,396
C	15	43	645	225	1,849
D	9	74	666	81	5,476
E	12	58	696	144	3,364
F	5	90	450	25	8,100
G	8	78	624	64	6,084
	$\Sigma x = 57$	$\Sigma y = 511$	$\Sigma xy = 3745$	$\Sigma x^2 = 579$	$\Sigma y^2 = 38{,}993$

Step 3 Substitute in the formula and solve for r.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{(7)(3745) - (57)(511)}{\sqrt{[(7)(579) - (57)^2][(7)(38{,}993) - (511)^2]}} = -0.944$$

The value of r suggests a strong negative linear relationship between a student's final grade and the number of absences a student has. That is, the more absences a student has, the lower is his or her grade.

EXAMPLE 10–6 Age and Wealth

Compute the value of the linear correlation coefficient for the data given in Example 10–3 for the age and wealth of the richest persons in the world.

SOLUTION

Step 1 Make a table.

Step 2 Find the values of xy, x^2, and y^2; place these values in the corresponding columns of the table.

Person	Age x	Net wealth y	xy	x^2	y^2
A	60	11	660	3,600	121
B	72	69	4,968	5,184	4,761
C	56	11.9	666.4	3,136	141.61
D	55	30	1,650	3,025	900
E	83	12.2	1,012.6	6,889	148.84
F	67	36	2,412	4,489	1,296
G	38	18.7	710.6	1,444	349.69
H	62	10.2	632.4	3,844	104.04
I	62	23.3	1,444.6	3,844	542.89
J	46	10.6	487.6	2,116	112.36
	$\Sigma x = 601$	$\Sigma y = 232.9$	$\Sigma xy = 14{,}644.2$	$\Sigma x^2 = 37{,}571$	$\Sigma y^2 = 8{,}477.43$

Step 3 Substitute in the formula and solve for r.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{10(14{,}644.2) - (601)(232.9)}{\sqrt{[10(37{,}571) - (601)^2][10(8477.43) - (232.9)^2]}}$$

$$= \frac{6469.1}{\sqrt{(14{,}509)(30{,}531.89)}} = \frac{6469.1}{21{,}047.26091} = 0.307$$

The value of r indicates a weak positive linear relationship between age and wealth of the richest people in the world.

In Example 10–4, the value of r was high (close to 1.00); in Example 10–6, the value of r was much lower (close to 0). This question then arises, When is the value of r due to chance, and when does it suggest a significant linear relationship between the variables? This question will be answered next.

OBJECTIVE 3

Test the hypothesis H_0: $\rho = 0$.

The Significance of the Correlation Coefficient As stated before, the range of the correlation coefficient is between -1 and $+1$. When the value of r is near $+1$ or -1, there is a strong linear relationship. When the value of r is near 0, the linear relationship is weak or nonexistent. Since the value of r is computed from data obtained from samples, there are two possibilities when r is not equal to zero: either the value of r is high enough to conclude that there is a significant linear relationship between the variables, or the value of r is due to chance.

To make this decision, you use a hypothesis-testing procedure. The traditional method is similar to the one used in previous chapters.

Step 1 State the hypotheses.

Step 2 Find the critical values.

Step 3 Compute the test value.

Step 4 Make the decision.

Step 5 Summarize the results.

The sample correlation coefficient can then be used as an estimator of ρ if the following assumptions are valid.

Assumptions for Testing the Significance of the Linear Correlation Coefficient

1. The data are quantitative and are obtained from a simple random sample.
2. The scatter plot shows that the data are approximately linearly related.
3. There are no outliers in the data.
4. The variables x and y must come from normally distributed populations.

In this book, the assumptions will be stated in the exercises; however, when encountering statistics in other situations, you must check to see that these assumptions have been met before proceeding.

In hypothesis testing, one of these is true:

H_0: $\rho = 0$ This null hypothesis means that there is no correlation between the x and y variables in the population.

H_1: $\rho \neq 0$ This alternative hypothesis means that there is a significant correlation between the variables in the population.

When the null hypothesis is rejected at a specific level, it means that there is a significant difference between the value of r and 0. When the null hypothesis is not rejected, it means that the value of r is not significantly different from 0 (zero) and is probably due to chance.

Several methods can be used to test the significance of the correlation coefficient. Three methods will be shown in this section. The first uses the t test.

Interesting Fact

Scientists think that a person is never more than 3 feet away from a spider at any given time!

Historical Notes

A mathematician named Karl Pearson (1857–1936) became interested in Francis Galton's work and saw that his correlation and regression theory could be applied to other areas besides heredity. Pearson developed the correlation coefficient that bears his name.

Formula for the *t* Test for the Correlation Coefficient

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

with degrees of freedom equal to $n - 2$, where n is the number of ordered pairs (x, y).

You do not have to identify the claim here, since the question will always be whether there is a significant linear relationship between the variables.

The two-tailed critical values are used. These values are found in Table F in Appendix A. Also, when you are testing the significance of a correlation coefficient, both variables x and y must come from normally distributed populations.

EXAMPLE 10–7

Test the significance of the correlation coefficient found in Example 10–4. Use $\alpha = 0.05$ and $r = 0.982$.

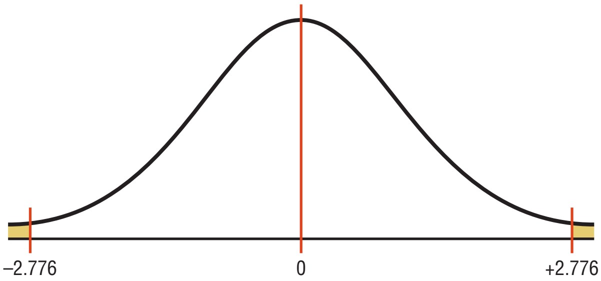
SOLUTION

Step 1 State the hypotheses.

$$H_0: \rho = 0 \qquad \text{and} \qquad H_1: \rho \neq 0$$

Step 2 Find the critical values. Since $\alpha = 0.05$ and there are $6 - 2 = 4$ degrees of freedom, the critical values obtained from Table F are ± 2.776, as shown in Figure 10–7.

FIGURE 10–7
Critical Values for Example 10–7

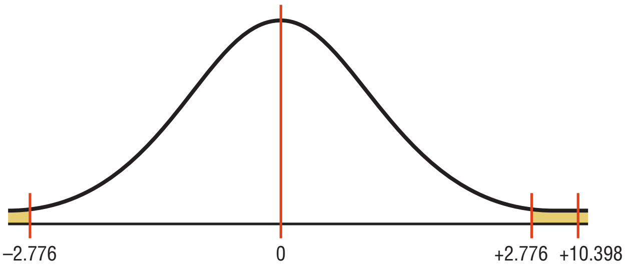


Step 3 Compute the test value.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.982\sqrt{\frac{6-2}{1-(0.982)^2}} = 10.398$$

Step 4 Make the decision. Reject the null hypothesis, since the test value falls in the critical region, as shown in Figure 10–8.

FIGURE 10–8
Test Value for Example 10–7



Step 5 Summarize the results. There is a significant relationship between the number of cars a rental agency owns and its annual income.

The second method that can be used to test the significance of *r* is the *P*-value method. The method is the same as that shown in Chapters 8 and 9. It uses the following steps.

Step 1 State the hypotheses.

Step 2 Find the test value. (In this case, use the *t* test.)

Step 3 Find the *P*-value. (In this case, use Table F.)

Step 4 Make the decision.

Step 5 Summarize the results.

Consider an example where $t = 4.059$, d.f. $= 4$, and $\alpha = 0.05$. Using Table F with d.f. $= 4$ and the row Two tails, the value 4.059 falls between 3.747 and 4.604; hence, $0.01 < P\text{-value} < 0.02$. (The *P*-value obtained from a calculator is 0.015.) That is, the *P*-value falls between 0.01 and 0.02. The decision, then, is to reject the null hypothesis since $P\text{-value} < 0.05$.

The third method of testing the significance of *r* is to use Table I in Appendix A. This table shows the values of the correlation coefficient that are significant for a specific α level and a specific number of degrees of freedom. For example, for 7 degrees of freedom and $\alpha = 0.05$, the table gives a critical value of 0.666. Any value of *r* greater than $+0.666$ or less than -0.666 will be significant, and the null hypothesis will be rejected. See Figure 10–9. When Table I is used, you need not compute the *t* test value. Table I is for two-tailed tests only.

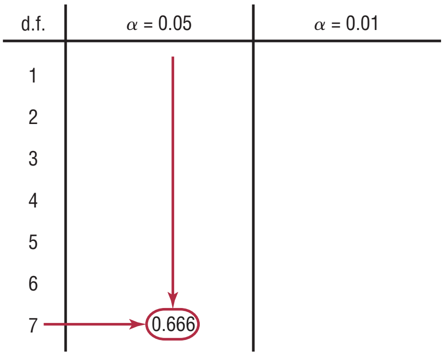


FIGURE 10–9
Finding the Critical Value from Table I

EXAMPLE 10–8

Using Table I, test the significance at $\alpha = 0.01$ of the correlation coefficient $r = 0.307$, obtained in Example 10–6.

SOLUTION

$$H_0\text{: } \rho = 0 \qquad \text{and} \qquad H_1\text{: } \rho \neq 0$$

Since the sample size is 10, there are $n - 2 = 10 - 2 = 8$ degrees of freedom. The critical values obtained from Table I at $\alpha = 0.01$ and 8 degrees of freedom are ± 0.765. Since $0.307 < 0.765$, the decision is to not reject the null hypothesis. Hence, there is not enough evidence to say that there is a significant linear relationship between age and wealth of the richest people in the world. See Figure 10–10.

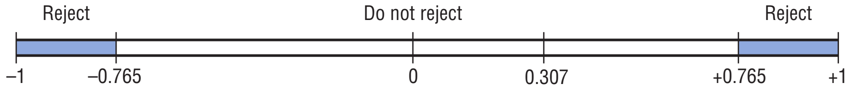


FIGURE 10–10
Rejection and Nonrejection Regions for Example 10–8

Generally, significance tests for correlation coefficients are two-tailed; however, they can be one-tailed. For example, if a researcher hypothesized a positive linear relationship between two variables, the hypotheses would be

$$H_0\text{: } \rho = 0$$
$$H_1\text{: } \rho > 0$$

If the researcher hypothesized a negative linear relationship between two variables, the hypotheses would be

$$H_0\text{: } \rho = 0$$
$$H_1\text{: } \rho < 0$$

In these cases, the t tests and the P-value tests would be one-tailed. Also, tables such as Table I are available for one-tailed tests. In this book, the examples and exercises will involve two-tailed tests.

Correlation and Causation Researchers must understand the nature of the linear relationship between the independent variable x and the dependent variable y. When a hypothesis test indicates that a significant linear relationship exists between the variables, researchers must consider the possibilities outlined next.

When two variables are highly correlated, item 3 in the box states that there exists a possibility that the correlation is due to a third variable. If this is the case and the

Possible Relationships Between Variables

When the null hypothesis has been rejected for a specific α value, any of the following five possibilities can exist.

1. *There is a direct cause-and-effect relationship between the variables.* That is, x causes y. For example, water causes plants to grow, poison causes death, and heat causes ice to melt.
2. *There is a reverse cause-and-effect relationship between the variables.* That is, y causes x. For example, suppose a researcher believes excessive coffee consumption causes nervousness, but the researcher fails to consider that the reverse situation may occur. That is, it may be that an extremely nervous person craves coffee to calm his or her nerves.
3. *The relationship between the variables may be caused by a third variable.* For example, if a statistician correlated the number of deaths due to drowning and the number of cans of soft drink consumed daily during the summer, he or she would probably find a significant relationship. However, the soft drink is not necessarily responsible for the deaths, since both variables may be related to heat and humidity.
4. *There may be a complexity of interrelationships among many variables.* For example, a researcher may find a significant relationship between students' high school grades and college grades. But there probably are many other variables involved, such as IQ, hours of study, influence of parents, motivation, age, and instructors.
5. *The relationship may be coincidental.* For example, a researcher may be able to find a significant relationship between the increase in the number of people who are exercising and the increase in the number of people who are committing crimes. But common sense dictates that any relationship between these two values must be due to coincidence.

third variable is unknown to the researcher or not accounted for in the study, it is called a **lurking variable.** An attempt should be made by the researcher to identify such variables and to use methods to control their influence.

It is important to restate the fact that even if the correlation between two variables is high, it does not necessarily mean causation. There are other possibilities, such as lurking variables or just a coincidental relationship. See the Speaking of Statistics article on page 563.

Also, you should be cautious when the data for one or both of the variables involve averages rather than individual data. It is not wrong to use averages, but the results cannot be generalized to individuals since averaging tends to smooth out the variability among individual data values. The result could be a higher correlation than actually exists.

Thus, when the null hypothesis is rejected, the researcher must consider all possibilities and select the appropriate one as determined by the study. Remember, correlation does not necessarily imply causation.

Applying the Concepts 10–1

Stopping Distances

In a study on speed control, it was found that the main reasons for regulations were to make traffic flow more efficient and to minimize the risk of danger. An area that was focused on in the study was the distance required to completely stop a vehicle at various speeds. Use the following table to answer the questions.

MPH	Braking distance (feet)
20	20
30	45
40	81
50	133
60	205
80	411