Correlation

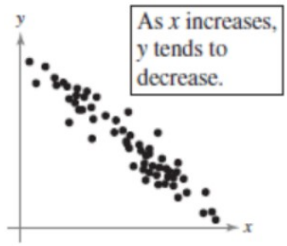
OBJECTIVES

- An introduction to linear correlation, independent and dependent variables, and the types of correlation
- How to find a correlation coefficient
- How to test a population correlation coefficient using a table
- How to perform a hypothesis test for a population correlation coefficient
- How to distinguish between correlation and causation

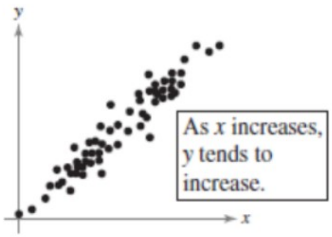
Suppose a safety inspector wants to determine whether a relationship exists between the number of hours of training for an employee and the number of accidents involving that employee. Or suppose a psychologist wants to know whether a relationship exists between the number of hours a person sleeps each night and that person's reaction time. How would he or she determine if any relationship exists? They would use the statistical tool called the **Pearson Correlation Coefficient**.

DEFINITION

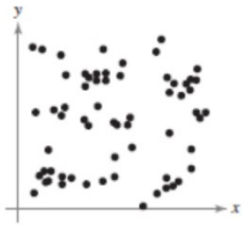
A **correlation** is a relationship between two variables. The data can be represented by the ordered pairs (x, y), where x is the **independent** (or **explanatory**) **variable** and y is the **dependent** (or **response**) **variable.**



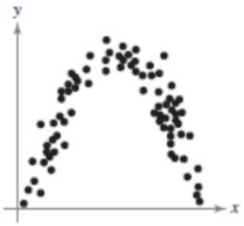
Negative Linear Correlation



Positive Linear Correlation



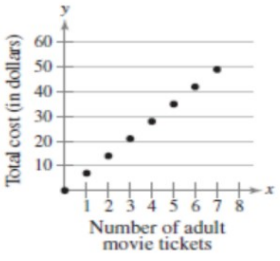
No Correlation



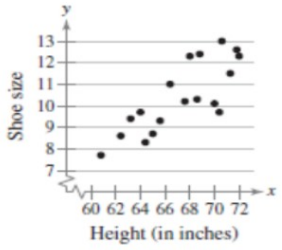
Nonlinear Correlation

Correlation

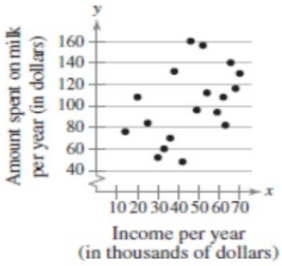
The range of the correlation coefficient is -1 to 1, inclusive. If x and y have a strong positive linear correlation, r is close to 1. If x and y have a strong negative linear correlation, r is close to -1. It is important to remember that if r is close to 0, it does not mean that there is no relation between x and y, just that there is no linear relation. Several examples are shown below.



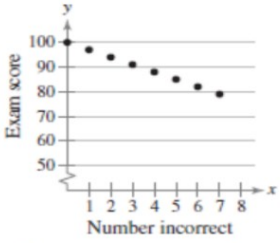
Perfect positive correlation
$r = 1$



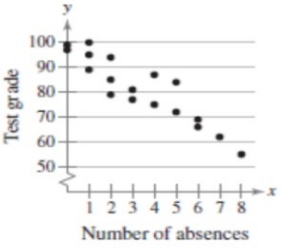
Strong positive correlation
$r = 0.81$



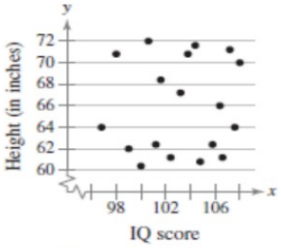
Weak positive correlation
$r = 0.45$



Perfect negative correlation
$r = -1$



Strong negative correlation
$r = -0.92$



No correlation
$r = 0.04$

INSIGHT

The formal name for r is the **Pearson product moment correlation coefficient.** It is named after the English statistician Karl Pearson (1857–1936).
(See page 33.)

Correlation

EXAMPLE 1 Constructing a Scatter Plot

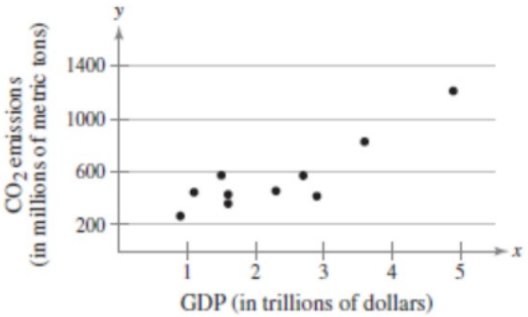
An economist wants to determine whether there is a linear relationship between a country's gross domestic product (GDP) and carbon dioxide (CO_2) emissions. The data are shown in the table at the left. Display the data in a scatter plot and determine whether there appears to be a positive or negative linear correlation or no linear correlation. *(Source: World Bank and U.S. Energy Information Administration)*

GDP (trillions of \$), x	CO_2 emissions (millions of metric tons), y
1.6	428.2
3.6	828.8
4.9	1214.2
1.1	444.6
0.9	264.0
2.9	415.3
2.7	571.8
2.3	454.9
1.6	358.7
1.5	573.5

Solution

The scatter plot is shown at the right. From the scatter plot, it appears that there is a positive linear correlation between the variables.

Interpretation Reading from left to right, as the gross domestic products increase, the carbon dioxide emissions tend to increase.



Correlation

Try It Yourself 1

A director of alumni affairs at a small college wants to determine whether there is a linear relationship between the number of years alumni classes have been out of school and their annual contributions (in thousands of dollars). The data are shown in the table at the left. Display the data in a scatter plot and determine the type of correlation.

a. *Draw* and *label* the x- and y-axes.
b. *Plot* each ordered pair.
c. Does there appear to be a linear correlation? If so, *interpret* the correlation in the context of the data.

STUDY TIP

You can also use MINITAB and Excel to construct scatter plots.

Correlation

EXAMPLE 2 Constructing a Scatter Plot

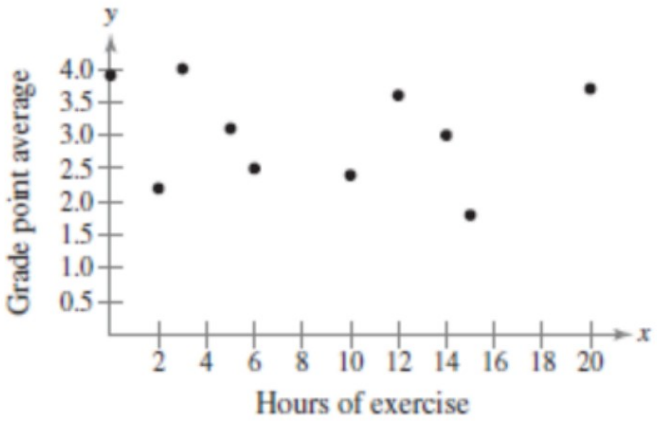
A student conducts a study to determine whether there is a linear relationship between the number of hours a student exercises each week and the student's grade point average (GPA). The data are shown in the following table. Display the data in a scatter plot and describe the type of correlation.

Hours of exercise, x	12	3	0	6	10	2	20	14	15	5
GPA, y	3.6	4.0	3.9	2.5	2.4	2.2	3.7	3.0	1.8	3.1

Solution

The scatter plot is shown at the right. From the scatter plot, it appears that there is no linear correlation between the variables.

Interpretation The number of hours a student exercises each week does not appear to be related to the student's grade point average.



Correlation

Try It Yourself 2

A researcher conducts a study to determine whether there is a linear relationship between a person's height (in inches) and pulse rate (in beats per minute). The data are shown in the following table. Display the data in a scatter plot and describe the type of correlation.

Height, x	68	72	65	70	62	75	78	64	68
Pulse rate, y	90	85	88	100	105	98	70	65	72

a. *Draw* and *label* the x- and y-axes.
b. *Plot* each ordered pair.
c. Does there appear to be a linear correlation? If so, *interpret* the correlation in the context of the data.

STUDY TIP

Save any data put into a technology tool because these data will be used throughout the chapter.

Correlation

EXAMPLE 3 ▸ Constructing a Scatter Plot Using Technology

Old Faithful, located in Yellowstone National Park, is the world's most famous geyser. The durations (in minutes) of several of Old Faithful's eruptions and the times (in minutes) until the next eruption are shown in the table at the left. Using a TI-83/84 Plus, display the data in a scatter plot. Describe the type of correlation.

Duration, x	Time, y	Duration, x	Time, y
1.80	56	3.78	79
1.82	58	3.83	85
1.90	62	3.88	80
1.93	56	4.10	89
1.98	57	4.27	90
2.05	57	4.30	89
2.13	60	4.43	89
2.30	57	4.47	86
2.37	61	4.53	89
2.82	73	4.55	86
3.13	76	4.60	92
3.27	77	4.63	91
3.65	77		

▸ Solution

Begin by entering the x-values into List 1 and the y-values into List 2. Use *Stat Plot* to construct the scatter plot. The plot should look similar to the one shown below. From the scatter plot, it appears that the variables have a positive linear correlation.

Interpretation You can conclude that the longer the duration of the eruption, the longer the time before the next eruption begins.

Correlation

Try It Yourself 3

Consider the data from the Chapter Opener on page 483 on the salaries and average attendances at home games for the teams in Major League Baseball. Use a technology tool to display the data in a scatter plot. Describe the type of correlation.

a. *Enter* the data into List 1 and List 2.
b. *Construct* the scatter plot.
c. Does there appear to be a linear correlation? If so, *interpret* the correlation in the context of the data.

STUDY TIP

You can also use MINITAB and Excel to construct scatter plots.

HYPOTHESIS TESTING FOR A POPULATION CORRELATION COEFFICIENT ρ

In this text, you will consider only two-tailed hypothesis tests for ρ.

THE t-TEST FOR THE CORRELATION COEFFICIENT

A ***t*-test** can be used to test whether the correlation between two variables is significant. The **test statistic** is r and the **standardized test statistic**

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

follows a t-distribution with $n - 2$ degrees of freedom.

GUIDELINES

Using the *t*-Test for the Correlation Coefficient ρ

	IN WORDS	IN SYMBOLS
1.	Identify the null and alternative hypotheses.	State H_0 and H_a.
2.	Specify the level of significance.	Identify α.
3.	Identify the degrees of freedom.	d.f. $= n - 2$
4.	Determine the critical value(s) and the rejection region(s).	Use Table 5 in Appendix B.
5.	Find the standardized test statistic.	$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$
6.	Make a decision to reject or fail to reject the null hypothesis.	If t is in the rejection region, reject H_0. Otherwise, fail to reject H_0.
7.	Interpret the decision in the context of the original claim.	

Correlation

EXAMPLE 4 Finding the Correlation Coefficient

Calculate the correlation coefficient for the gross domestic products and carbon dioxide emissions data given in Example 1. What can you conclude?

Solution Use a table to help calculate the correlation coefficient.

GDP (trillions of $), x	CO_2 emissions (millions of metric tons), y	xy	x^2	y^2
1.6	428.2	685.12	2.56	183,355.24
3.6	828.8	2983.68	12.96	686,909.44
4.9	1214.2	5949.58	24.01	1,474,281.64
1.1	444.6	489.06	1.21	197,669.16
0.9	264.0	237.6	0.81	69,696
2.9	415.3	1204.37	8.41	172,474.09
2.7	571.8	1543.86	7.29	326,955.24
2.3	454.9	1046.27	5.29	206,934.01
1.6	358.7	573.92	2.56	128,665.69
1.5	573.5	860.25	2.25	328,902.25
$\Sigma x = 23.1$	$\Sigma y = 5554$	$\Sigma xy = 15{,}573.71$	$\Sigma x^2 = 67.35$	$\Sigma y^2 = 3{,}775{,}842.76$

With these sums and $n = 10$, the correlation coefficient is

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

$$= \frac{10(15{,}573.71) - (23.1)(5554)}{\sqrt{10(67.35) - 23.1^2}\sqrt{10(3{,}775{,}842.76) - 5554^2}}$$

$$= \frac{27{,}439.7}{\sqrt{139.89}\sqrt{6{,}911{,}511.6}} \approx 0.882.$$

STUDY TIP

Notice that the correlation coefficient r in Example 4 is rounded to three decimal places. This *round-off rule* will be used throughout the text.

The result $r \approx 0.882$ suggests a strong positive linear correlation.

Interpretation As the gross domestic product increases, the carbon dioxide emissions also increase.

Correlation

EXAMPLE 5 The *t*-Test for a Correlation Coefficient

In Example 4, you used 10 pairs of data to find $r \approx 0.882$. Test the significance of this correlation coefficient. Use $\alpha = 0.05$.

Solution

The null and alternative hypotheses are

H_0: $\rho = 0$ (no correlation) and H_a: $\rho \neq 0$ (significant correlation).

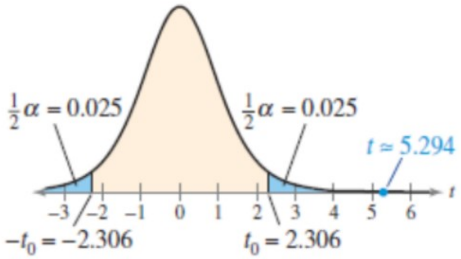
Because there are 10 pairs of data in the sample, there are $10 - 2 = 8$ degrees of freedom. Because the test is a two-tailed test, $\alpha = 0.05$, and d.f. $= 8$, the critical values are $-t_0 = -2.306$ and $t_0 = 2.306$. The rejection regions are $t < -2.306$ and $t > 2.306$. Using the t-test, the standardized test statistic is

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} \approx \frac{0.882}{\sqrt{\frac{1 - (0.882)^2}{10 - 2}}} \approx 5.294.$$

STUDY TIP

Be sure you see in Example 5 that rejecting the null hypothesis means that there is enough evidence that the correlation is significant.

The following graph shows the location of the rejection regions and the standardized test statistic.



Because t is in the rejection region, you should decide to reject the null hypothesis.

Interpretation There is enough evidence at the 5% level of significance to conclude that there is a significant linear correlation between gross domestic products and carbon dioxide emissions.

Correlation

In Try It Yourself 5, you calculated the correlation coefficient of the salaries and average attendances at home games for the teams in Major League Baseball to be $r \approx 0.74972$. Test the significance of this correlation coefficient. Use $\alpha = 0.01$.

a. State the *null* and *alternative hypotheses*.
b. Identify the *level of significance*.
c. Identify the *degrees of freedom*.
d. Determine the *critical values* and the *rejection regions*.
e. Find the *standardized test statistic*.
f. *Make a decision* to reject or fail to reject the null hypothesis.
g. *Interpret* the decision in the context of the original claim.

INSIGHT

In Example 5 you can use Table 11 in Appendix B to test the population correlation coefficient ρ. Given $n = 10$ and $\alpha = 0.05$, the critical value from Table 11 is 0.632. Because

$$|r| \approx 0.882 > 0.632,$$

the correlation is significant. Note that this is the same result you obtained using a t-test for the population correlation coefficient ρ.

CORRELATION AND CAUSATION

The fact that two variables are strongly correlated does not in itself imply a cause-and- effect relationship between the variables. More in-depth study is usually needed to determine whether there is a causal relationship between the variables. If there is a significant correlation between two variables, a researcher should consider the following possibilities

1. **Is there a direct cause-and-effect relationship between the variables?**
   That is, does x cause y? For instance, consider the relationship between gross domestic products and carbon dioxide emissions that has been discussed throughout this section. It is reasonable to conclude that an increase in a country's gross domestic product will result in higher carbon dioxide emissions.

2. **Is there a reverse cause-and-effect relationship between the variables?**
   That is, does y cause x? For instance, consider the Old Faithful data that have been discussed throughout this section. These variables have a positive linear correlation, and it is possible to conclude that the duration of an eruption affects the time before the next eruption. However, it is also possible that the time between eruptions affects the duration of the next eruption.

3. **Is it possible that the relationship between the variables can be caused by a third variable or perhaps a combination of several other variables?**
   For instance, consider the salaries and average attendances per home game for the teams in Major League Baseball listed in the Chapter Opener. Although these variables have a positive linear correlation, it is doubtful that just because a team's salary decreases, the average attendance per home game will also decrease. The relationship is probably due to several other variables, such as the economy, the players on the team, and whether or not the team is winning games.

4. **Is it possible that the relationship between two variables may be a coincidence?**
   For instance, although it may be possible to find a significant correlation between the number of animal species living in certain regions and the number of people who own more than two cars in those regions, it is highly unlikely that the variables are directly related. The relationship is probably due to coincidence.

Determining which of the cases above is valid for a data set can be difficult. For instance, consider the following example. Suppose a person breaks out in a rash each time he eats shrimp at a certain restaurant. The natural conclusion is that the person is allergic to shrimp. However, upon further study by an allergist, it is found that the person is not allergic to shrimp, but to a type of seasoning the chef is putting into the shrimp.

Correlation

OBJECTIVES

- An introduction to linear correlation, independent and dependent variables, and the types of correlation
- How to find a correlation coefficient
- How to test a population correlation coefficient using a table
- How to perform a hypothesis test for a population correlation coefficient
- How to distinguish between correlation and causation

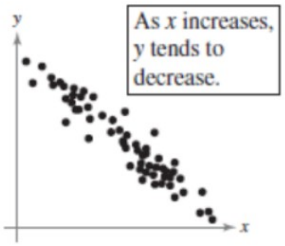
Suppose a safety inspector wants to determine whether a relationship exists between the number of hours of training for an employee and the number of accidents involving that employee. Or suppose a psychologist wants to know whether a relationship exists between the number of hours a person sleeps each night and that person's reaction time. How would he or she determine if any relationship exists? They would use the statistical tool called the **Pearson Correlation Coefficient**.

DEFINITION

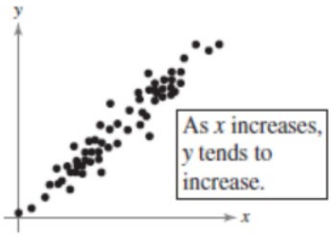
A **correlation** is a relationship between two variables. The data can be represented by the ordered pairs (x, y), where x is the **independent** (or **explanatory**) **variable** and y is the **dependent** (or **response**) **variable.**

Classwork

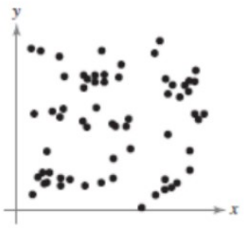
Online text Page 495, #1 -20



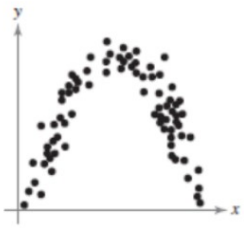
Negative Linear Correlation



Positive Linear Correlation



No Correlation



Nonlinear Correlation