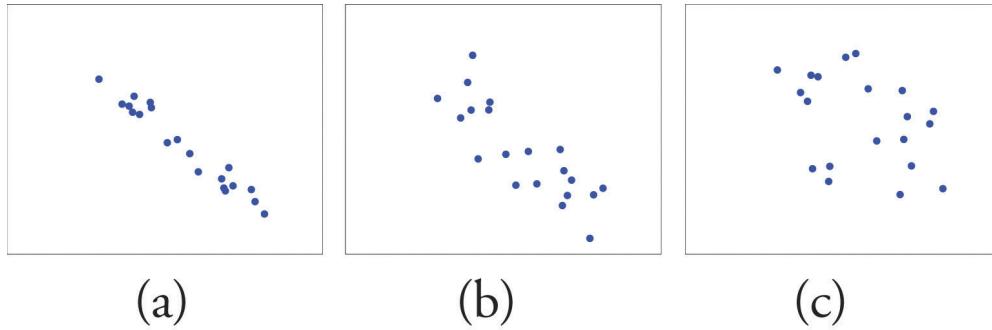# 10.2 The Linear Correlation Coefficient

LEARNING OBJECTIVE

1. To learn what the linear correlation coefficient is, how to compute it, and what it tells us about the relationship between two variables x and y.

Figure 10.3 "Linear Relationships of Varying Strengths" illustrates linear relationships between two variables x and y of varying strengths. It is visually apparent that in the situation in panel (a), x could serve as a useful predictor of y, it would be less useful in the situation illustrated in panel (b), and in the situation of panel (c) the linear relationship is so weak as to be practically nonexistent. The *linear correlation coefficient* is a number computed directly from the data that measures the strength of the linear relationship between the two variables x and y.

Figure 10.3 *Linear Relationships of Varying Strengths*

(a)   (b)   (c)

Definition

The **linear correlation coefficient** *for a collection of n pairs (x, y) of numbers in a sample is the number r given by the formula*

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}}$$

where

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2, \quad SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y), \quad SS_{yy} = \Sigma y^2 - \frac{1}{n}(\Sigma y)^2$$

The linear correlation coefficient has the following properties, illustrated in Figure 10.4 "Linear Correlation Coefficient ":

1. The value of r lies between -1 and 1, inclusive.
2. The sign of r indicates the direction of the linear relationship between x and y:
    1. If $r < 0$ then y tends to decrease as x is increased.
    2. If $r > 0$ then y tends to increase as x is increased.
3. The size of $|r|$ indicates the strength of the linear relationship between x and

y:

1. If $|r|$ is near 1 (that is, if r is near either 1 or −1) then the linear relationship between *x* and *y* is strong.
2. If $|r|$ is near 0 (that is, if r is near 0 and of either sign) then the linear relationship between *x* and *y* is weak.

Figure 10.4 Linear Correlation Coefficient R
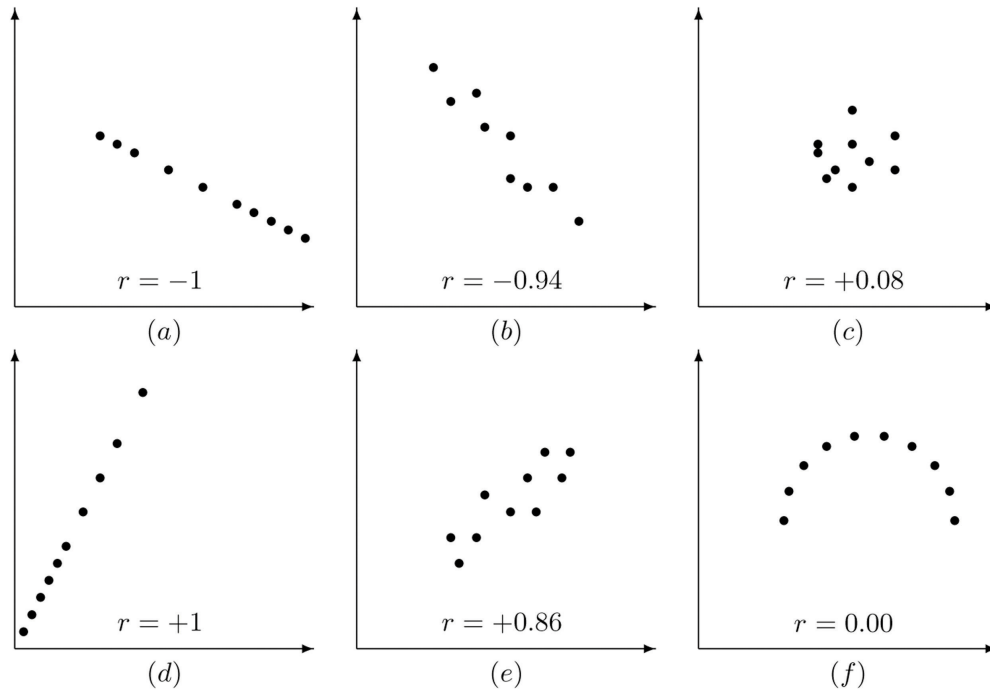
Pay particular attention to panel (f) in Figure 10.4 "Linear Correlation Coefficient ". It shows a perfectly deterministic relationship between *x* and *y*, but $r = 0$ because the relationship is not linear. (In this particular case the points lie on the top half of a circle.)

EXAMPLE 1

Compute the linear correlation coefficient for the height and weight pairs plotted in Figure 10.2 "Plot of Height and Weight Pairs".

Solution:

Even for small data sets like this one computations are too long to do completely by hand. In actual practice the data are entered into a calculator or computer and a statistics program is used. In order to clarify the meaning of the formulas we will display the data and related quantities in tabular form. For each (x, y) pair we compute three numbers: x^2, xy, and y^2, as shown in the table provided. In the last line of the table we have the sum of the numbers in each column. Using them we compute:

	x	y	x²	xy	y²
	68	151	4624	10268	22801
	69	146	4761	10074	21316
	70	157	4900	10990	24649
	70	164	4900	11480	26896
	71	171	5041	12141	29241
	72	160	5184	11520	25600
	72	163	5184	11736	26569
	72	180	5184	12960	32400
	73	170	5329	12410	28900
	73	175	5329	12775	30625
	74	178	5476	13172	31684
	75	188	5625	14100	35344
Σ	859	2003	61537	143626	336025

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 = 61537 - \frac{1}{12}(859)^2 = 46.91\overline{6}$$

$$SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y) = 143626 - \frac{1}{12}(859)(2003) = 244.58\overline{3}$$

$$SS_{yy} = \Sigma y^2 - \frac{1}{n}(\Sigma y)^2 = 336025 - \frac{1}{12}(2003)^2 = 1690.91\overline{6}$$

so that

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{244.58\overline{3}}{\sqrt{(46.91\overline{6})(1690.91\overline{6})}} = 0.868$$

The number $r = 0.868$ quantifies what is visually apparent from <u>Figure 10.2 "Plot of Height and Weight Pairs"</u>: weights tends to increase linearly with height (r is positive) and although the relationship is not perfect, it is reasonably strong (r is near 1).

KEY TAKEAWAYS

- The linear correlation coefficient measures the strength and direction of the linear relationship between two variables x and y.
- The sign of the linear correlation coefficient indicates the direction of the linear relationship between x and y.
- When r is near 1 or −1 the linear relationship is strong; when it is near 0 the linear relationship is weak.

EXERCISES

BASIC

With the exception of the exercises at the end of <u>Section 10.3 "Modelling Linear Relationships with Randomness Present"</u>, the first Basic exercise in each of the following sections through <u>Section 10.7 "Estimation and Prediction"</u> uses the data from the first exercise here, the second Basic exercise uses the data from the second exercise here, and so on, and similarly for the Application exercises. Save your computations done on these exercises so that you do not need to repeat them later.

1. For the sample data

$$
\begin{array}{c|ccccc}
x & 0 & 1 & 3 & 5 & 8 \\
\hline
y & 2 & 4 & 6 & 5 & 9
\end{array}
$$

   a. Draw the scatter plot.
   b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.
   c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

2. For the sample data

$$\begin{array}{c|ccccc} x & 0 & 2 & 3 & 6 & 9 \\ \hline y & 0 & 3 & 3 & 4 & 8 \end{array}$$

a. Draw the scatter plot.

b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.

c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

3. For the sample data

$$\begin{array}{c|ccccc} x & 1 & 3 & 4 & 6 & 8 \\ \hline y & 4 & 1 & 3 & -1 & 0 \end{array}$$

a. Draw the scatter plot.

b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.

c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

4. For the sample data

$$\begin{array}{c|ccccc} x & 1 & 2 & 4 & 7 & 9 \\ \hline y & 5 & 5 & 6 & -3 & 0 \end{array}$$

a. Draw the scatter plot.

b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.

c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

5. For the sample data

$$\begin{array}{c|ccccc} x & 1 & 1 & 3 & 4 & 5 \\ \hline y & 2 & 1 & 5 & 3 & 4 \end{array}$$

a. Draw the scatter plot.

b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.

c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

6. For the sample data

$$\begin{array}{c|ccccc} x & 1 & 3 & 5 & 5 & 8 \\ \hline y & 5 & -2 & 2 & -1 & -3 \end{array}$$

a. Draw the scatter plot.

b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.

c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

7. Compute the linear correlation coefficient for the sample data summarized by the following information:

$$n = 5 \qquad \Sigma\, x = 25 \qquad \Sigma\, x^2 = 165$$
$$\Sigma\, y = 24 \quad \Sigma\, y^2 = 134 \quad \Sigma\, xy = 144$$
$$1 \le x \le 9$$

8. Compute the linear correlation coefficient for the sample data summarized by the following information:

$$n = 5 \qquad \Sigma\, x = 31 \qquad \Sigma\, x^2 = 253$$
$$\Sigma\, y = 18 \quad \Sigma\, y^2 = 90 \qquad \Sigma\, xy = 148$$
$$2 \le x \le 12$$

9. Compute the linear correlation coefficient for the sample data summarized by the following information:

$$n = 10 \qquad \Sigma\, x = 0 \qquad \Sigma\, x^2 = 60$$
$$\Sigma\, y = 24 \quad \Sigma\, y^2 = 234 \quad \Sigma\, xy = -87$$
$$-4 \le x \le 4$$

10. Compute the linear correlation coefficient for the sample data summarized by the following information:

$$n = 10 \qquad \Sigma\, x = -3 \qquad \Sigma\, x^2 = 263$$
$$\Sigma\, y = 55 \quad \Sigma\, y^2 = 917 \quad \Sigma\, xy = -355$$
$$-10 \le x \le 10$$

APPLICATIONS

11. The age x in months and vocabulary y were measured for six children, with the results shown in the table.

x	13	14	15	16	16	18
y	8	10	15	20	27	30

Compute the linear correlation coefficient for these sample data and interpret its meaning in

the context of the problem.

12. The curb weight x in hundreds of pounds and braking distance y in feet, at 50 miles per hour on dry pavement, were measured for five vehicles, with the results shown in the table.

x	25	27.5	32.5	35	45
y	105	125	140	140	150

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

13. The age x and resting heart rate y were measured for ten men, with the results shown in the table.

x	20	23	30	37	35
y	72	71	73	74	74
x	45	51	55	60	63
y	73	72	79	75	77

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

14. The wind speed x in miles per hour and wave height y in feet were measured under various conditions on an enclosed deep water sea, with the results shown in the table,

x	0	0	2	7	7
y	2.0	0.0	0.3	0.7	3.3
x	9	13	20	22	31
y	4.9	4.9	3.0	6.9	5.9

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

15. The advertising expenditure x and sales y in thousands of dollars for a small retail business in its first eight years in operation are shown in the table.

x	1.4	1.6	1.6	2.0
y	180	184	190	220
x	2.0	2.2	2.4	2.6
y	186	215	205	240

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

16. The height x at age 2 and y at age 20, both in inches, for ten women are tabulated in the table.

x	31.3	31.7	32.5	33.5	34.4
y	60.7	61.0	63.1	64.2	65.9
x	35.2	35.8	32.7	33.6	34.8
y	68.2	67.6	62.3	64.9	66.8

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

17. The course average x just before a final exam and the score y on the final exam were recorded for 15 randomly selected students in a large physics class, with the results shown in the table.

x	69.3	87.7	50.5	51.9	82.7
y	56	89	55	49	61
x	70.5	72.4	91.7	83.3	86.5
y	66	72	83	73	82
x	79.3	78.5	75.7	52.3	62.2
y	92	80	64	18	76

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

18. The table shows the acres x of corn planted and acres y of corn harvested, in millions of acres, in a particular country in ten successive years.

x	75.7	78.9	78.6	80.9	81.8
y	68.8	69.3	70.9	73.6	75.1
x	78.3	93.5	85.9	86.4	88.2
y	70.6	86.5	78.6	79.5	81.4

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

19. Fifty male subjects drank a measured amount x (in ounces) of a medication and the concentration y (in percent) in their blood of the active ingredient was measured 30 minutes later. The sample data are summarized by the following information.

$$n = 50 \quad \Sigma x = 112.5 \quad \Sigma y = 4.83$$
$$\Sigma xy = 15.255 \quad 0 \le x \le 4.5$$
$$\Sigma x^2 = 356.25 \quad \Sigma y^2 = 0.667$$

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

20. In an effort to produce a formula for estimating the age of large free-standing oak trees non-

invasively, the girth x (in inches) five feet off the ground of 15 such trees of known age y (in years) was measured. The sample data are summarized by the following information.

$$n = 15 \quad \Sigma x = 3368 \quad \Sigma y = 6496$$
$$\Sigma xy = 1{,}933{,}219 \quad \Sigma x^2 = 917{,}780$$
$$\Sigma y^2 = 4{,}260{,}666 \quad 74 \leq x \leq 395$$

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

21. Construction standards specify the strength of concrete 28 days after it is poured. For 30 samples of various types of concrete the strength x after 3 days and the strength y after 28 days (both in hundreds of pounds per square inch) were measured. The sample data are summarized by the following information.

$$n = 30 \quad \Sigma x = 501.6 \quad \Sigma y = 1338.8$$
$$\Sigma xy = 23{,}246.55 \quad \Sigma x^2 = 8724.74$$
$$\Sigma y^2 = 61{,}980.14 \quad 11 \leq x \leq 22$$

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

22. Power-generating facilities used forecasts of temperature to forecast energy demand. The average temperature x (degrees Fahrenheit) and the day's energy demand y (million watt-hours) were recorded on 40 randomly selected winter days in the region served by a power company. The sample data are summarized by the following information.

$$n = 40 \quad \Sigma x = 2000 \quad \Sigma y = 2969$$
$$\Sigma xy = 143{,}042 \quad \Sigma x^2 = 101{,}340$$
$$\Sigma y^2 = 243{,}027 \quad 40 \leq x \leq 60$$

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

ADDITIONAL EXERCISES

23. In each case state whether you expect the two variables x and y indicated to have positive, negative, or zero correlation.

a. the number x of pages in a book and the age y of the author

b. the number x of pages in a book and the age y of the intended reader

c. the weight x of an automobile and the fuel economy y in miles per gallon

d. the weight x of an automobile and the reading y on its odometer

e. the amount x of a sedative a person took an hour ago and the time y it takes him to respond to a stimulus

24. In each case state whether you expect the two variables x and y indicated to have positive, negative, or zero correlation.

    a. the length x of time an emergency flare will burn and the length y of time the match used to light it burned

    b. the average length x of time that calls to a retail call center are on hold one day and the number y of calls received that day

    c. the length x of a regularly scheduled commercial flight between two cities and the headwind y encountered by the aircraft

    d. the value x of a house and the its size y in square feet

    e. the average temperature x on a winter day and the energy consumption y of the furnace

25. Changing the units of measurement on two variables x and y should not change the linear correlation coefficient. Moreover, most change of units amount to simply multiplying one unit by the other (for example, 1 foot = 12 inches). Multiply each x value in the table in Exercise 1 by two and compute the linear correlation coefficient for the new data set. Compare the new value of r to the one for the original data.

26. Refer to the previous exercise. Multiply each x value in the table in Exercise 2 by two, multiply each y value by three, and compute the linear correlation coefficient for the new data set. Compare the new value of r to the one for the original data.

27. Reversing the roles of x and y in the data set of Exercise 1 produces the data set

$$
\begin{array}{c|ccccc}
x & 2 & 4 & 6 & 5 & 9 \\
\hline
y & 0 & 1 & 3 & 5 & 8 \\
\end{array}
$$

Compute the linear correlation coefficient of the new set of data and compare it to what you got in Exercise 1.

28. In the context of the previous problem, look at the formula for r and see if you can tell why what

you observed there must be true for every data set.

LARGE DATA SET EXERCISES

29. Large Data Set 1 lists the SAT scores and GPAs of 1,000 students. Compute the linear correlation coefficient r. Compare its value to your comments on the appearance and strength of any linear trend in the scatter diagram that you constructed in the first large data set problem for Section 10.1 "Linear Relationships Between Variables".

    http://www.flatworldknowledge.com/sites/all/files/data1.xls

30. Large Data Set 12 lists the golf scores on one round of golf for 75 golfers first using their own original clubs, then using clubs of a new, experimental design (after two months of familiarization with the new clubs). Compute the linear correlation coefficient r. Compare its value to your comments on the appearance and strength of any linear trend in the scatter diagram that you constructed in the second large data set problem for Section 10.1 "Linear Relationships Between Variables".

    http://www.flatworldknowledge.com/sites/all/files/data12.xls

31. Large Data Set 13 records the number of bidders and sales price of a particular type of antique grandfather clock at 60 auctions. Compute the linear correlation coefficient r. Compare its value to your comments on the appearance and strength of any linear trend in the scatter diagram that you constructed in the third large data set problem for Section 10.1 "Linear Relationships Between Variables".

    http://www.flatworldknowledge.com/sites/all/files/data13.xls

ANSWERS

1. $r = 0.921$

3. $r = -0.794$

5. $r = 0.707$

7. 0.875

9. −0.846

11. 0.948

13. 0.709

15. 0.832

17. 0.751

19. 0.965

21. 0.992

23.　　a. zero

　　　b. positive

　　　c. negative

　　　d. zero

　　　e. positive

25. same value

27. same value

29. $r = 0.4601$

31. $r = 0.9002$