CHAPTER 8
CORRELATION

OBJECTIVES

When you complete this chapter, you should be able to

1. recognize the strength and type of relationship between two variables,
2. demonstrate an understanding of a correlation coefficient,
3. select and use appropriate statistical methods to compute a correlation coefficient,
4. read and interpret a scatter plot, and
5. compute Pearson and Spearman rank-order correlation coefficients.

Key Terms

When you complete this chapter, you should understand the following terms

correlation coefficient
- Pearson correlation coefficient
- relationship
- scatter plot
- Spearman rank-order correlation coefficient

Thus far, we have explored individual statistics, which illustrate individual sets of scores through frequencies, central tendency, and variance. Sometimes, however, you will find it necessary to determine the relationship between two sets of scores. For example, are high SAT verbal scores (now called critical reading scores) associated with high college English grades? Do high-ability students tend to excel in all of their academic subjects? This chapter will discuss the concept of *correlation,* which is used in later chapters that will explain the concepts of validity and reliability. Here, we introduce the ***Pearson correlation coefficient***, a statistic that is used with ratio or interval-scaled data. In addition, we introduce the Spearman correlation, which is used with ranked or ordinal scaled data.

THE PEARSON CORRELATION COEFFICIENT

The ***correlation coefficient*** is a statistic that illustrates the existence of a linear ***relationship*** between two variables. It also expresses the strength of the relationship. For example, you might ask the question, are higher teacher salaries linked to higher academic achievement scores among students in a school district?

The numerical values of a correlation coefficient range between –1.00 and +1.00. The higher the numerical value, the stronger the relationship between the two variables (see Table 8.1). If the coefficient has a positive sign, the relationship is positive: If one value is high, the other value is high. Conversely, if the coefficient is negative, the relationship is negative: If one value is high, the other value is low. For example, when a correlation coefficient is positive, if x is high, so is y. You would expect that a person who scores high on an algebra aptitude test would also obtain high grades in algebra. The correlation is also positive when a low value for x corresponds to a low value for y.

Table 8.1

Interpretations of Correlations

Correlation	Interpretation
0.000 to 0.200	Very Weak
0.201 to 0.400	Weak
0.401 to 0.600	Moderate
0.601 to 0.800	Strong
0.801 to 1.000	Very Strong

Note: Numerical values can be plus or minus.

Conversely, if a correlation is negative, high scores on x are associated with low scores on y.

As an example, let's look at the relationship between the SAT math scores and the trigonometry grades of 10th, 11th, and 12th graders. Table 8.2 shows these students' SAT mathematics scores and trigonometry grades. To determine the correlation coefficient, we will calculate the Pearson correlation coefficient for this set of scores.

Table 8.2

SAT Math Scores and Trigonometry Achievement Scores

Forrest	740	95
Jamal	680	90
Alexandra	660	90
Lauren	550	86
Sabrina	500	80
Roberto	480	80
Chang	500	75
Jane	470	75
Rick	480	70
Pete	400	65

The Pearson correlation coefficient, sometimes called the Pearson product-moment correlation, is a measure of the linear relationship between the paired values of two variables (x and y). The equation takes into account each paired value and uses the mean, standard deviation, and z-score formulas in its computation (see Equation 8.1).

Equation 8.1

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{(n-1)\,S_x S_y}$$

In this equation, $(x - x)$ is the deviation of the x variable from the mean, $(y - y)$ is the deviation of the y variable from the mean, S_x is the sample standard deviation for the x variable, S_y is the sample standard deviation for the y variable, and n represents the number of pairs of scores. To simplify this equation, we will represent r as the average value of the products of paired z-scores. This formula is found in Equation 8.2.

Equation 8.2

$$r = \frac{\sum z_x z_y}{n-1}$$

As our first example, we will examine the relationship between SAT mathematics scores and the percentage scores from a 12th-grade trigonometry class. Table 8.3 shows the relationship between the two variables, using the definitional formula given in Equation 8.2.

Table **8.3**

Pearson Correlation Calculation

Student	SAT (x)	Trig. Grade (y)	$z_x = (x - \bar{x})/S_x$	$z_y = (y - \bar{y})/S_y$	$z_x z_y$
Forrest	740	95	1.77	1.50	2.66
Jamal	680	90	1.22	0.98	1.20
Alexandra	660	90	1.04	0.98	1.02
Lauren	550	86	0.04	-0.57	0.02
Sabrina	500	80	-0.42	-0.05	0.02
Rob	480	80	-0.60	-0.05	0.03
Chang	500	75	-0.42	-0.57	0.24
Jane	470	75	-0.69	-0.67	0.46
Rick	480	70	-0.60	-1.08	0.65
Pete	400	65	-1.33	-1.60	2.13
	$\bar{x} = 546$ $S_x = 109.87$	$\bar{y} = 80.50$ $S_y = 9.69$			$\sum z_x z_y = 8.43$

Notice that, for each variable (x and y), z-scores are calculated to transform the data onto the same scale. Next z_x is multiplied by z_y and then summed with the formula $Sz_x z_y$, which will total to a positive sum value when a majority of positive z_x scores is multiplied by positive z_y scores, or, in a negative sum value, when a majority of positive z_x scores is multiplied by negative zy scores. Calculation 8.1 shows that $r = 0.94$, which translates to a very strong positive correlation between SAT math scores and trigonometry scores. There-

fore, we can conclude that high SAT scores are positively related to high trigonometry scores.

Calculation 8.1

$$r = \frac{\sum z_x z_y}{n-1}$$

or

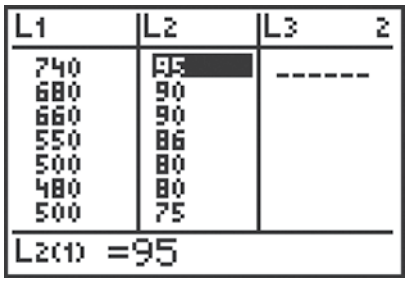
$$r = \frac{8.43}{10-1} = 0.94$$

CALCULATOR EXPLORATION

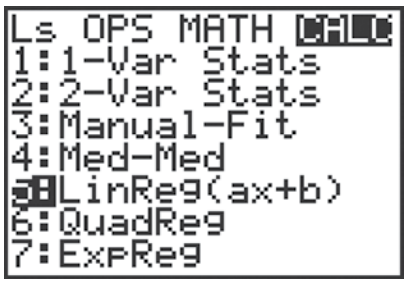
Using either the TI-73, TI-83, or TI-84 graphing calculator, you can easily compute a correlation coefficient by following these keystrokes.

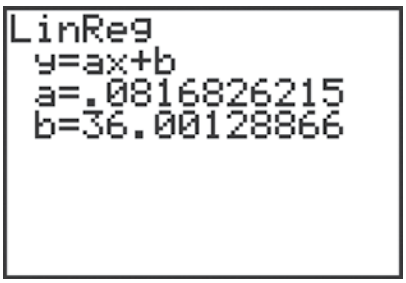
TI-73

Step 1. Display list editor by pressing [LIST]. Here, you can enter up to 999 elements. Under [L1], list all of the scores from the SATs that are listed in Table 8.3. Next, under [L2], list all of the scores listed under trigonometry grades in Table 8.3.

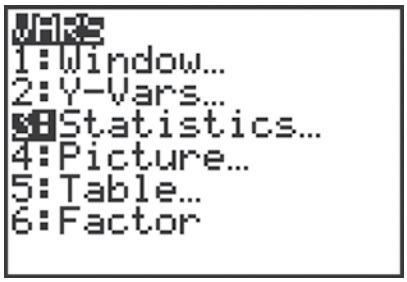


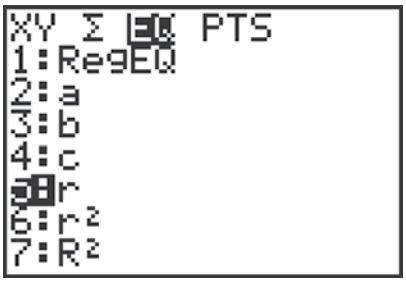
Step 2. Press [2nd LIST] to activate [STAT]. Next, scroll to the right and select [CALC]. Next, scroll down to [5: LinReg (ax+b)], press [5], then press [ENTER].

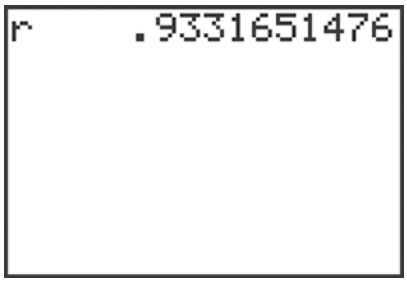




Step 3. Press [2nd APPS] to activate [VARS]. Next, scroll down to [3: Statistics], and press [3]. Next, scroll over to [EQ] and scroll down to [5: r], select [5], and press [ENTER]. Press [ENTER] again to get the correlation coefficient.

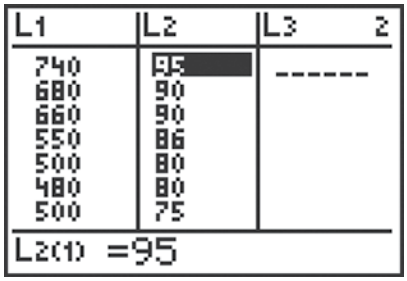




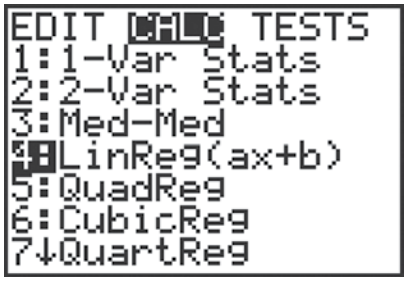


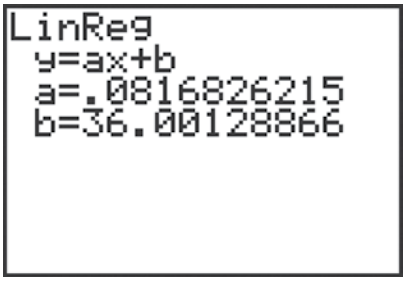
TI-83/TI-84

Step 1. First press [STAT]. Next, select [1: Edit] by pressing [1]. Here you can enter up to 999 elements. Under [L1], list all of the scores from the SATs that are listed in Table 8.3. Next, under [L2], list all of the scores listed under trigonometry grades in Table 8.3.

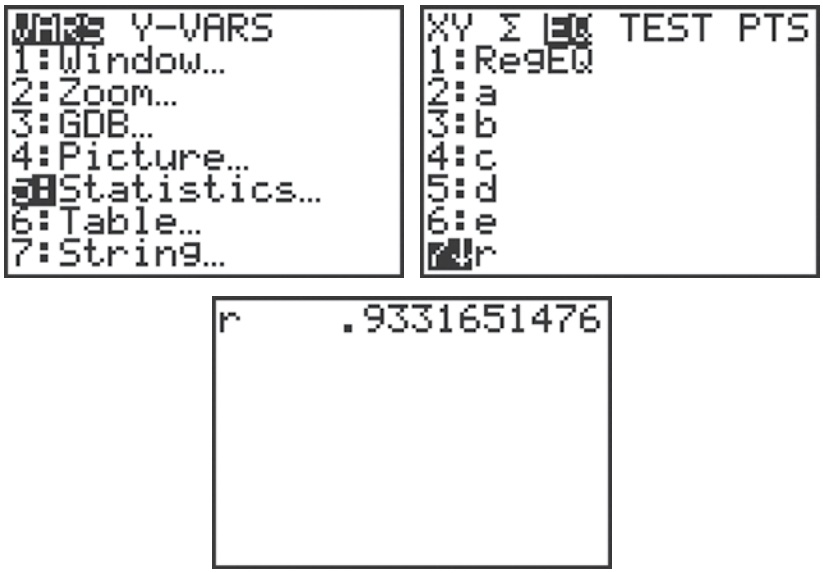


Step 2. Press [STAT] and scroll to the right and select [CALC]. Next, scroll down to [4: LinReg (ax+b)], press [4], and press [ENTER].





Step 3. Press [VARS]. Next, scroll down to [5: Statistics], and press [5]. Next, scroll over to [EQ]. Next, scroll down to [7: r] and press [7]. [ENTER] again to get the validity coefficient.



CORRELATION AND CAUSE AND EFFECT

People often think that correlation means the same thing as causation. Although a correlation means there is a relationship between two variables, it does not mean that one causes the other. For example, there is a positive correlation between ice cream consumption and death by drowning. However, common sense tells us that eating ice cream does not cause death by drowning; simply, when the weather is hot, more people eat ice cream and more people go swimming! As a further example, a plot of monthly sales of ice cream against monthly deaths from heart disease would show a negative correlation. Again, based on hundreds of research studies, it is hardly likely that eating ice cream protects from heart disease. It is simply that the mortality rate from heart disease is inversely related—and ice cream consumption positively related—to a third factor: environmental temperature. It is important to understand that a high correlation between two variables does not imply that one causes the other.

SCATTER PLOTS

The relationship between two variables can be illustrated in a ***scatter plot***, which is a graph showing the paired scores. Figure 8.1 shows a scatter plot for SAT mathematics section scores and trigonometry grades. Notice that, because our relationship is very close to a perfect +1.0, it is almost a straight line.

Figure 8.1

Scatter Plot for SAT Mathematics Section Scores and Trigonometry Grades

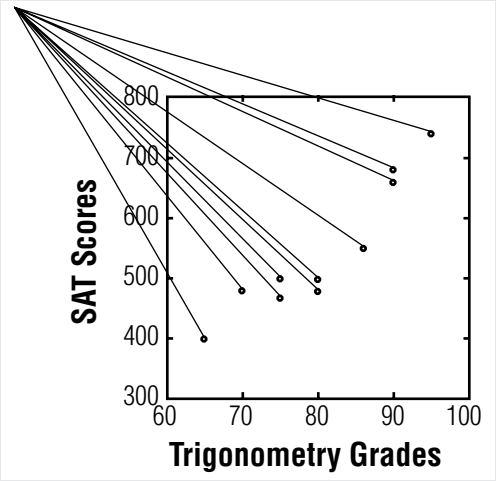


Figure 8.2 is a scatter plot of a perfect +1.00 correlation. In this case, we have positive scores for variables *x* and *y*. Likewise, since variable *x* is Fahrenheit temperature and variable *y* is the corresponding centigrade temperature, this scatter plot shows a perfectly straight line, which is known as a regression line.

Figure 8.2

A Scatter Plot of a Perfect +1.00 Correlation

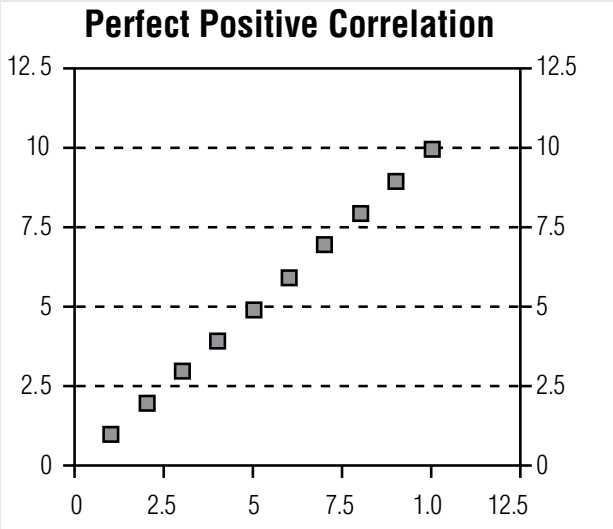


Figure 8.3

A Scatter Plot of a Perfect –1.00 Correlation

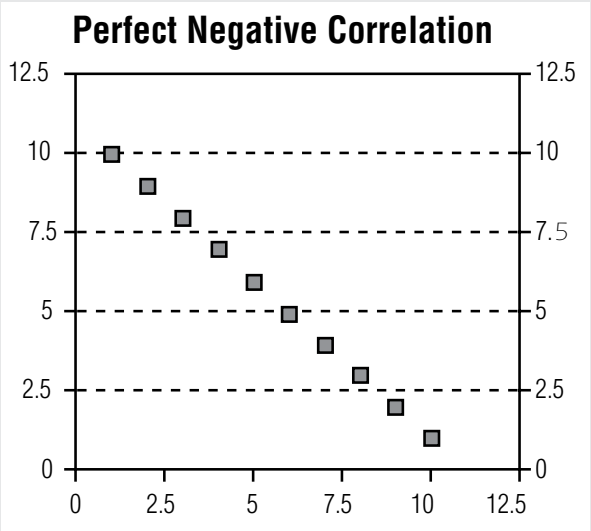


Figure 8.3 shows a scatter plot of a perfect –1.00 correlation. It shows the average velocity (*x*) to the race time (*y*) of a person running a 5k race, we would find that as the runner's running speed increases, his or her time decreases, giving us a perfect negative correlation.

Figure 8.4 shows a scatter plot of a high positive correlation between people's heights (*x*) and weights (*y*).

Figure 8.4

A Scatter Plot of a High Positive Correlation

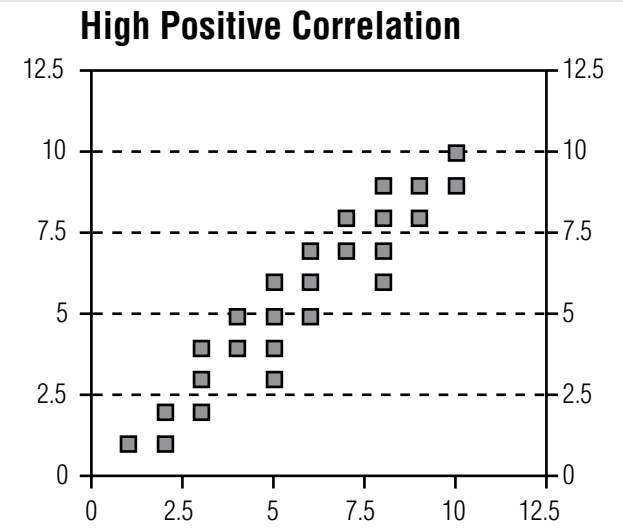


Figure 8.5 shows a scatter plot of a low negative correlation. The scatter plot compares the amount of money spent on education (*x*) with test scores (*y*), showing that as spending increases slightly, academic achievement slightly decreases, giving us a low negative correlation between the two variables (Christmann and Badgett 2000). It is important to remember, however, that a correlation shows only that two variables are related; it does not show a cause and effect relationship. This happens because slight increases in teachers' pay do not affect how hard teachers work—which would often result in higher academic achievement. In essence, most teachers are dedicated professionals who put forth great effort regardless of pay increases.

Figure 8.6 is a scatter plot that shows no correlation. For example, if we were to compare combined SAT scores (*x*) to the heights (*y*) of high school seniors, we would find that there is no relationship between the two variables, which should yield a correlation of 0, for no correlation.

Figure 8.5

A Scatter Plot of a Low Negative Correlation

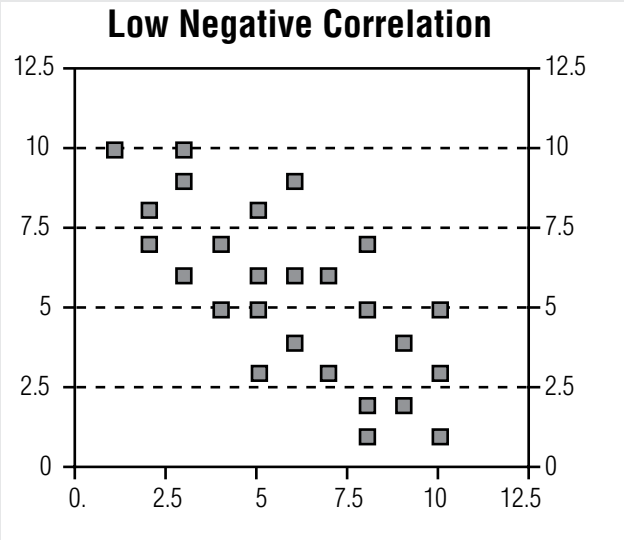
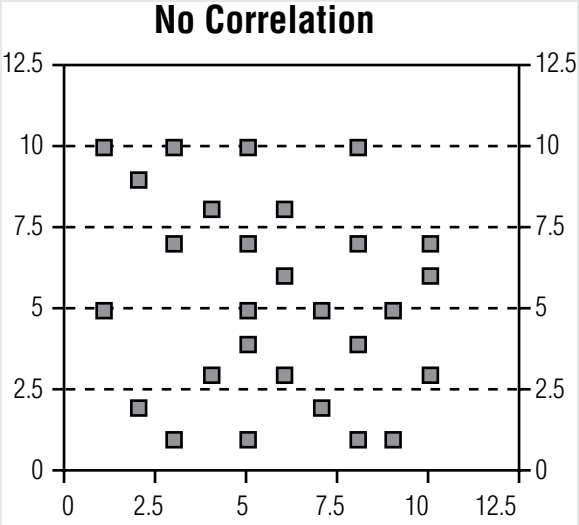


Figure 8.6

A Scatter Plot of a No Correlation



Check for Understanding

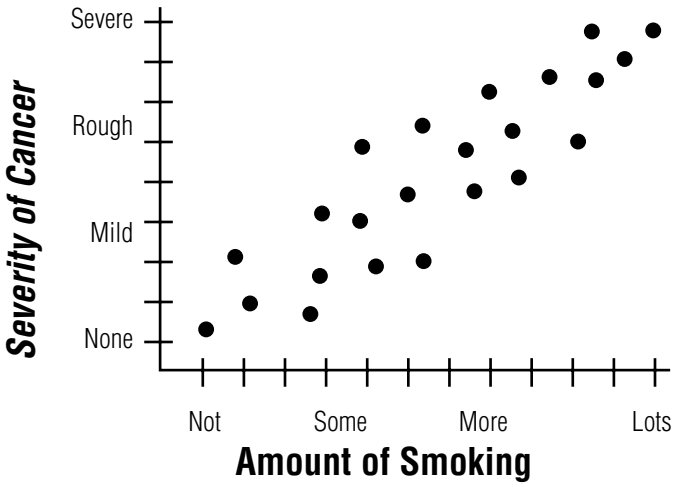
8.1. Which correlation coefficient shows the weakest relationship?
a. + 0.94
b. –0.77
c. + 0.15
d. –0.22

8.2. Calculate a Pearson correlation coefficient between the two variables given in the table to the right.

8.3. How would the calculated correlation coefficient from question 5.2 be classified according to the categories given in Table 5.1?

(x)	(y)
100	77
95	76
90	80
85	70
80	75
75	74
70	72
65	75
60	69
55	60

8.4. How would you best describe the scatter plot below?
a. A scatter plot with no correlation
b. A perfect positive correlation
c. A low negative correlation
d. A high positive correlation



SPEARMAN RANK–ORDER CORRELATION COEFFICIENT

The ***Spearman rank-order correlation coefficient*** describes the linear relationship between two variables that are ranked on an ordinal scale of measurement. For example, you could use the Spearman rank-order coefficient to show such things as academic achievement based on class rank and IQ.

The symbol r_s represents the Spearman correlation coefficient (see Equation 8.3). Just like the Pearson r correlation coefficient, the Spearman correlation ranges between -1.00 and +1.00. We interpret the strength of a Spearman rank-order correlation the same way we interpret the strength of the correlations in Table 8.1.

Equation 8.3

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

For example, suppose we are interested in whether 12th-grade students' GPAs are related to their rankings on the SAT (see Table 8.4, p. 176). The first step in calculating a Spearman rank-order correlation, as shown in Table 8.4, is to rank the GPAs and the SAT scores. Once the scores have been ranked, we subtract the difference between the ranks, D. Next we square the differences (D^2) and compute the sum of the squared differences, SD^2. Then, as illustrated in Equation 8.3, we use the Spearman formula to determine a correlation coefficient:

rs = The Spearman rank-order correlation coefficient
D = The difference between the ranks on the two variables
n = The number of individuals or pairs of ranks

Calculation 8.2 (p. 176) shows how to use the formula to obtain a Spearman rank-order correlation coefficient.

Table 8.4

Twelfth Graders' GPAs and Combined SAT Scores

Student	GPA	Rank	SAT	Rank	D	D^2
John Henry	3.97	1	1180	3	–1	1
Kate	3.85	2.5*	1130	6	–3.50	12.25
Wyatt	3.85	2.5*	1070	8	–5.50	30.25
Mattie	3.67	4	1180	3	1	1
Virgil	3.52	5	1140	5	0	0
Johnny	3.42	6	1290	1	5	25
Josephine	3.36	7	960	9	–2	4
Billy	3.24	8	1180	3	5	25
Jack	2.74	9	1110	7	2	4
Ike	2.56	10	720	10	0	0
						$\Sigma D^2 = 102.50$

When ties occur on x or y, assign the average of the rank involved to each score (e.g., $\frac{2+3}{2} = 2.50$).

Calculation 8.2

$$r_s = 1 - \frac{6\Sigma\,(D^2)}{n(n^2-1)}$$

Step 1

$$r_s = 1 - \frac{6\Sigma\,(105.5)}{10\,(10^2-1)}$$

Step 2

$$r_s = 1 - \frac{633}{990}$$

Step 3

$$r_s = 1 - 0.639 = 0.331$$

Calculation 8.2 shows that $r_s = 0.331$, which is interpreted as a weak positive relationship between GPA and combined SAT scores. Therefore, the example shows that GPA is not a very good predictor of SAT scores and that other predictors, such as motivaton and achievement test results, should be considered.

Check For Understanding

8.5. Which scale of measurement is required for a Spearman rank-order correlation?
a. Ratio
b. Interval
c. Ordinal
d. Nominal

8.6. Find the Spearman rank-order correlation coefficient between the two sets of scores.

Scores

Student	Reading	IQ
A	98	145
B	91	135
C	88	125
D	85	120
E	80	110
F	75	105
G	70	95
H	74	95
I	65	90
J	60	80

8.7. Based on your calculated Spearman correlation coefficient from question 8.6, explain the strength of the relationship between reading scores and IQ scores (very weak, moderate, or strong, for instance) and determine the direction of the relationship (is it positive or negative)?

SUMMARY

The correlation coefficient helps us understand the relationship between two variables. The Pearson correlation coefficient, the most commonly used correlation statistic, is used with either an interval or a ratio scale of measurement. The Spearman rank-order correlation examines the relationship between two variables that are on an ordinal scale.

All coefficients of correlation range from –1.0 to +1.0, with the direction of the relationship shown by the sign (+ or –), and the strength of the relationship is based on its numerical value. For example, a correlation coefficient of zero indicates that there is no correlation between two variables, whereas a positive correlation shows that as the numerical values associated with one variable increase, the numerical values associated with the other variable increase as well. With a negative correlation, as one variable increases, the other variable decreases. A correlation statistic only shows the relative strength of a linear relationship between two variables. It does not imply a cause-and-effect relationship between two variables.

Classroom teachers use correlation to explore relationships between variables. For example, is intelligence related to academic achievement? The next two chapters cover validity and reliability, which deal with the concept of correlation. Therefore, to answer questions related to validity and reliability, an understanding of correlation is necessary.

CHAPTER REVIEW QUESTIONS

8.8. How is a Pearson correlation coefficient different from a Spearman correlation?

8.9. Explain how to calculate a Pearson correlation coefficient.

8.10. The numerical value for a correlation ranges between
a. 0.00 and +1.00.
b. –1.00 and 0.00.
c. –1.00 and +1.00.
d. none of the above.

8.11. What does the Pearson correlation coefficient measure?

8.12. A scatter plot shows a data set spread in a circular pattern. Which correlation coefficient would best describe the correlation for these data?
a. 0.00
b. 1.00
c. –0.50
d. all of the above

8.13. For a Pearson correlation coefficient of –1.90 between test X and test Y, the correlation indicates that
a. as scores on test X increase, scores on test Y decrease.
b. as scores on test X increase, scores on test Y increase.
c. as scores on test X decrease, scores on test Y decrease.
d. none of the above.

8.14. Calculate a Pearson correlation coefficient for the following sets of test scores.

Student	Test A	Test B
A	98	100
B	71	81
C	85	66
D	64	50

8.15. If the original data are measured on an ordinal scale of measurement, what type of correlation should you use?
a. Pearson
b. Spearman
c. Gaussian
d. all of the above

8.16. Calculate a Spearman correlation for the following sets of ranked test scores.

Student	SAT (combined)	GPA
A	1420	3.90
B	1100	3.00
C	1050	2.80
D	980	3.00
E	700	1.68

8.17. A researcher finds a high positive correlation between price of school lunches and achievement test scores. As a result, the researcher concludes that more expensive lunches cause test scores to increase. Do you agree or disagree with the researcher's conclusion? Please explain your decision.

ANSWERS: CHECK FOR UNDERSTANDING

8.1. c. + 0.15
8.2. + 0.733
8.3. A strong positive correlation
8.4. d. A high positive correlation
8.5. c. Ordinal
8.6. $r_s = 0.997$
8.7. There is a very strong positive correlation. This means as one score increases the other increases.

ANSWERS: CHAPTER REVIEW QUESTIONS

8.8. The Pearson and Spearman coefficients are mathematically identical. However, the Spearman rank coefficient is calculated from the ranks of each variable, not the actual values.
8.9. The Pearson correlation coefficient is calculated based on the following formula that uses your z-scores:

$$r = \frac{\sum z_x z_y}{n-1}$$

8.10. c. –1.00 and 1.00.
8.11. The purpose of the Pearson correlation coefficient is to indicate a linear relationship between two measurement variables. This means that if you have two sets of scores, you can determine if one score predicts another score.
8.12. a. 0.00
8.13. d. none of the above.
8.14. Correlation coefficient (r) = 0.7752
8.15. b. Spearman
8.16. Spearman r = 0.825 (corrected for ties)
8.17. Because correlation does not establish cause and effect, the research should not conclude causality.

INTERNET RESOURCES

This website of the University of Florida is linked to universities, statistical journals, mailing lists, and statistical software vendors. It is an excellent site for resource and reference information related to correlation. *www.stat.ufl.edu/vlib/statistics.html*

This Yahoo website offers links to some of the most popular statistics links on the internet. In addition, it provides links to statistics journals and online software. *http://dir.yahoo.com/science/mathematics/statistics*

This website of the American Statistical Association offers updated information about the field of statistics, professional journals, professional development courses, and careers. *www.amstat.org*

This website provides a free 30-day download of the Analyze-it software package. Analyze-it is an add-in for Microsoft Excel (for Windows) and is designed to calculate correlation coefficients, along with other statistical calculations that are covered in this text. *www.analyze-it.com*

REFERENCE

Christmann, E. P., and J. L. Badgett. 2000. A four-year analytic comparison of eleventh grade academic achievement in the Slippery Rock Area High School, and district pupil expenditures. Research report presented to The Slippery Rock School District's Board of Directors and The Pennsylvania Department of Education. (ERIC Document No. ED443824)

FURTHER READING

American Educational Research Association (AERA). 1999. *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Elmore, P. B., and P. L. Woehlke. 1997. *Basic statistics.* New York: Longman.

Gravetter, F. J., and L.B. Wallnau. 2002. *Essentials of statistics for the behavioral sciences.* New York: West.

Pearson, E. S. 1968. Some early correspondence between W. S. Gosset, R. A. Fischer, and Karl Pearson, with notes and comments. In *Studies in the history of statistics and probability,* eds. E. S. Pearson and M. G. Kendall. Vol. 1, pp. 405–417. London: Charles Griffin.

Raymond, J. C. 1999. *Statistical analysis in the behavioral sciences.* New York: McGraw-Hill.

Thorne, B. M., and J. M. Giesen. 2000. *Statistics for the behavioral sciences.* Mountain View, CA: Mayfield.

Zawojewski, J. S., and J. M. Shaughnessy. 2000. Mean and median: Are they really so easy? *Mathematics Teaching in the Middle School* 5 (7): 436–440.