this z score contains the one crucial bit of information common to the three original observations: All are located one standard deviation below the mean. Accordingly, to find the proportion for the shaded areas in Figure 5.4 (that is, the proportion of applicants who are less than exactly 65 inches tall, or light bulbs that burn for fewer than 1080 hours, or fourth graders whose IQ scores are less than 90), we can use the same z score of –1.00 when referring to the table for the standard normal curve, the one table for all normal curves.

Standard Normal Table

Essentially, the standard normal table consists of columns of z scores coordinated with columns of proportions. In a typical problem, access to the table is gained through a z score, such as –1.00, and the answer is read as a proportion, such as the proportion of eligible FBI applicants.

Using the Top Legend of the Table

Table 5.1 shows an abbreviated version of the standard normal curve, while Table A in Appendix C on page 530 shows a more complete version of the same curve. Notice that columns are arranged in sets of three, designated as A, B, and C in the legend at the top of the table. When using the top legend, all entries refer to the upper half of the standard normal curve. The entries in column A are z scores, beginning with 0.00 and ending (in the full-length table of Appendix C) with 4.00. Given a z score of zero or more, columns B and C indicate how the z score splits the area in the upper half of the normal curve. As suggested by the shading in the top legend, column B indicates the proportion of area between the mean and the z score, and column C indicates the proportion of area beyond the z score, in the upper tail of the standard normal curve.

Using the Bottom Legend of the Table

Because of the symmetry of the normal curve, the entries in Table 5.1 and Table A of Appendix C also can refer to the lower half of the normal curve. Now the columns are designated as A′, B′, and C′ in the legend at the bottom of the table. When using the bottom legend, all entries refer to the lower half of the standard normal curve.

Imagine that the nonzero entries in column A′ are negative z scores, beginning with –0.01 and ending (in the full-length table of Appendix C) with –4.00. Given a negative z score, columns B′ and C′ indicate how that z score splits the lower half of the normal curve. As suggested by the shading in the bottom legend of the table, column B′ indicates the proportion of area between the mean and the negative z score, and column C′ indicates the proportion of area beyond the negative z score, in the lower tail of the standard normal curve.

Progress Check *5.2 Using Table A in Appendix C, find the proportion of the total area identified with the following statements:

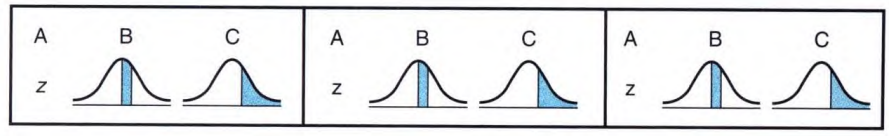
(a) above a z score of 1.80

(b) between the mean and a z score of –0.43

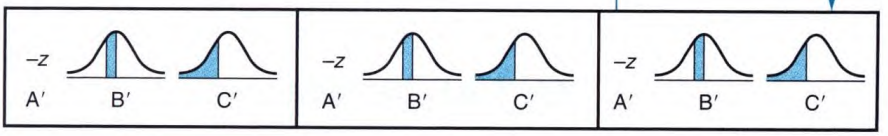
(c) below a z score of –3.00

Reminder:
Use of standard normal table always involves z scores.

Table 5.1
PROPORTIONS (OF AREAS) UNDER THE STANDARD NORMAL CURVE FOR VALUES OF *z* (FROM TABLE A OF APPENDIX C)



z	A	B	C	z	A	B	C	z	A	B	C
0.00	.0000	.5000		0.40	.1554	.3446		0.80	.2881	.2119	
0.01	.0040	.4960		0.41	.1591	.3409		0.81	.2910	.2090	
•	•	•		•	•	•		•	•	•	
								•	•	•	
								•	•	•	
								•	•	•	
								•	•	•	
•	•	•		•	•	•		•	•	•	
								0.99	.3389	.1611	
								1.00	.3413	.1587	
•	•	•		•	•	•		1.01	.3438	.1562	
								•	•	•	
								•	•	•	
•	•	•		•	•	•		•	•	•	
0.38	.1480	.3520		0.78	.2823	.2711		1.18	.3810	.1190	
0.39	.1517	.3483		0.79	.2852	.2148		1.19	.3830	.1170	



(d) between the mean and a *z* score of 1.65

(e) between *z* scores of 0 and −1.96

Answers on page 495.

5.4 SOLVING NORMAL CURVE PROBLEMS

Sections 5.5 and 5.6 give examples of two main types of normal curve problems. In the first type of problem, we use a known score (or scores) to find an unknown *proportion*. For instance, we use the known score of 65 inches to find the unknown proportion of eligible FBI applicants. In the second type of problem, the procedure

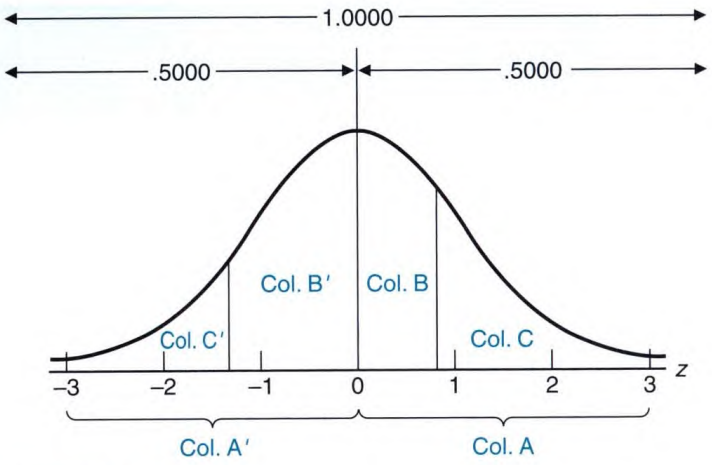


FIGURE 5.5
Interpretation of Table A, Appendix C.

is reversed. Now we use a known proportion to find an unknown *score* (or *scores*). For instance, if the FBI director had specified that applicants' heights must not exceed the 25th percentile (the shortest .25) of the population, we would use the known proportion of .25 to find the unknown cutoff height in inches.

Solve Problems Logically

Do not rush through these examples, memorizing solutions to particular problems or looking for some magic formula. Concentrate on the logic of the solution, *using rough graphs of normal curves as an aid to visualizing the solution.* Only after thinking through to a solution should you do any calculations and consult the normal tables. Then, with just a little practice, you will view the wide variety of normal curve problems not as a bewildering assortment but as many slight variations on two distinctive types.

Key Facts to Remember

When using the standard normal table, it is important to remember that for any z score, the corresponding proportions in columns B and C (or columns B′ and C′) always sum to .5000. Similarly, the total area under the normal curve always equals 1.0000, the sum of the proportions in the lower and upper halves, that is, .5000 + .5000. Finally, although a z score can be either positive or negative, the proportions of area under the curve are always positive or zero but *never* negative (because an area cannot be negative). **Figure 5.5** summarizes how to interpret the normal curve table in this book.

Reminder:
z scores can be negative, but not areas under the normal curve.

5.5 FINDING PROPORTIONS

Example: Finding Proportions for *One* Score

Now we'll use a step-by-step procedure, adopted throughout this chapter, to find the proportion of all FBI applicants who are shorter than exactly 65 inches,

given that the distribution of heights approximates a normal curve with a mean of 68 inches and a standard deviation of 3 inches.

1. **Sketch a normal curve and shade in the target area,** as in the left part of **Figure 5.6**. Being less than the mean of 68, 65 is located to the left of the mean. Furthermore, since the unknown proportion represents those applicants who are shorter than 65 inches, the shaded target sector is located to the left of 65.
2. **Plan your solution according to the normal table.** Decide precisely how you will find the value of the target area. In the present case, the answer will be obtained from column C′ of the standard normal table, since the target area coincides with the type of area identified with column C′, that is, the area in the lower tail beyond a negative z.
3. **Convert X to z.** Express 65 as a z score:

$$z = \frac{X - \mu}{\sigma} = \frac{65 - 68}{3} = \frac{-3}{3} = -1.00$$

4. **Find the target area.** Refer to the standard normal table, using the bottom legend, as the z score is negative. The arrows in Table 5.1 show how to read the table. Look up column A′ to 1.00 (representing a z score of –1.00), and note the corresponding proportion of .1587 in column C′: This is the answer, as suggested in the right part of Figure 5.6. It can be concluded that only .1587 (or .16) of all of the FBI applicants will be shorter than 65 inches.

A Clarification

Because the normal curve is defined for continuous variables, such as height, the same proportion of .1587 would describe not only FBI applicants who are shorter than 65 inches, but also FBI applicants who are shorter than *or equal to* 65 inches. If you think about it, equal to 65 inches translates into a height of exactly 65 inches, that is, 65.0000 with a string of zeros out to infinity! No measured height can coincide with *exactly* 65 inches since, in theory, however long the string of zeros for someone's height, measurement always can be carried additional steps until a nonzero appears.

Exactly 65 inches translates into a point along the horizontal base of the normal curve. The vertical line through this point defines one side of the desired

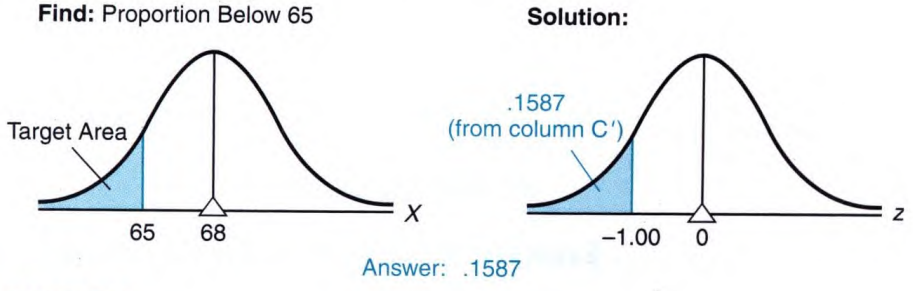


FIGURE 5.6
Finding proportions.

area—the portion below 65 inches—but the line itself has no area. Therefore, when doing normal curve problems, you need not agonize over, for example, whether the desired proportion is below exactly 65 inches or below *and equal to* exactly 65 inches. The answer is the same.

Read Carefully

Carefully read normal curve problems. A single word can change the entire problem as, for example, if you had been asked to find the proportion of applicants who are *taller* than 65 inches. Now we must find the total area to the right, not to the left, of 65 inches (or a z score of –1.00) in Figure 5.6. This requires that we add the proportions for two sectors: the unshaded sector between 65 inches and the mean of 68 inches and the unshaded sector above the mean of 68 inches. To find the proportion between 65 and 68 inches, refer to the standard normal table. Use the bottom legend, as the z score is negative; look up column A′ to 1.00 (representing a z score of –1.00); and note the proportion of .3438 in column B′ (which corresponds to the sector between 65 and 68 inches.) Recalling that .5000 always equals the proportion in the upper half of the curve (above the mean of 68 inches), add these two proportions, .3438 + .5000 = .8438, to determine that .8438 of all FBI applicants will be taller than 65 inches.

Reminder about Interpreting Areas

When read from left to right, the X and z scales along the base of the normal curve, as in Figure 5.6, always increase in value. Accordingly, the area under the normal curve to the left of any given score represents the proportion of shorter applicants (or, more generally, smaller or lower scores), and the area to the right of any given score represents the proportion of taller applicants (or larger or higher scores).

Progress Check *5.3 Assume that GRE scores approximate a normal curve with a mean of 500 and a standard deviation of 100.

(a) Sketch a normal curve and shade in the target area described by each of the following statements:

(a_1) less than 400

(a_2) more than 650

(a_3) less than 700

(b) Plan solutions (in terms of columns B, C, B′, or C′ of the standard normal table, as well as the fact that the proportion for either the entire upper half or lower half always equals .5000) for the target areas in part (a).

(c) Convert to z scores and find the proportions that correspond to the target areas in part (a).

Answers on page 495.

Example: Finding Proportions *between* Two Scores

Assume that, when not interrupted artificially, the gestation periods for human fetuses approximate a normal curve with a mean of 270 days (9 months) and a

standard deviation of 15 days. What proportion of gestation periods will be between 245 and 255 days?

1. **Sketch a normal curve and shade in the target area,** as in the top panel of **Figure 5.7**. Satisfy yourself that, in fact, the shaded area represents just those gestation periods between 245 and 255 days.
2. **Plan your solution according to the normal table.** This type of problem requires more effort to solve because the value of the target area cannot be read directly from Table A. As suggested in the bottom two panels of Figure 5.7, the basic idea is to identify the target area with the difference between two overlapping areas whose values can be read from column C′ of Table A. The larger area (less than 255 days) contains two sectors: the target area (between 245 and 255 days) and a remainder (less than 245 days). The smaller area contains only the remainder (less than 245 days). Subtracting the smaller area (less than 245 days) from the larger area (less than 255 days), therefore, eliminates the common remainder (less than 245 days), leaving only the target area (between 245 and 255 days).
3. **Convert X to z** by expressing 255 as

$$z = \frac{255 - 270}{15} = \frac{-15}{15} = -1.00$$

and by expressing 245 as

$$z = \frac{245 - 270}{15} = \frac{-25}{15} = -1.67$$

4. **Find the target area.** Look up column A′ to a negative z score of −1.00 (remember, you must imagine the negative sign), and note the corresponding

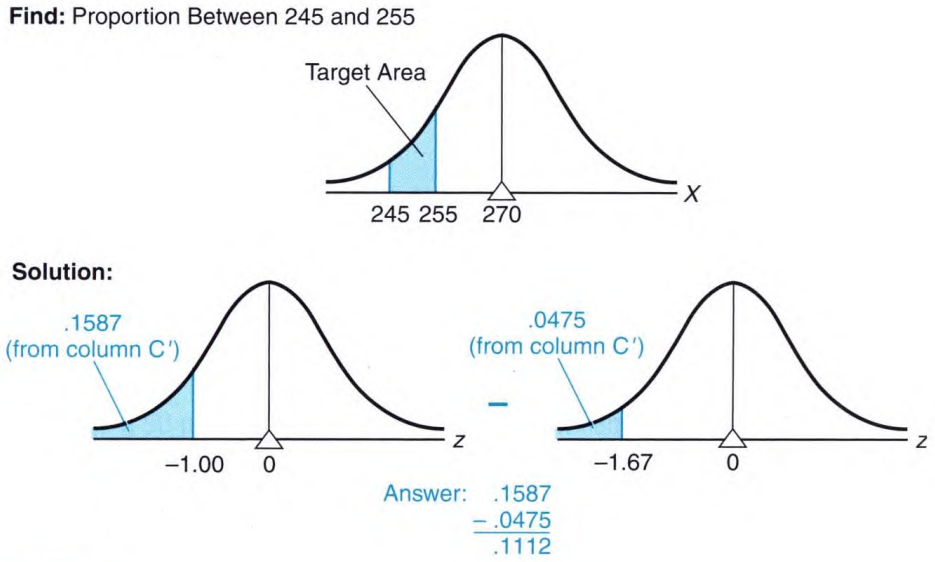


FIGURE 5.7
Finding proportions.

proportion of .1587 in column C′. Likewise, look up column A′ to a z score of –1.67, and note the corresponding proportion of .0475 in column C′. Subtract the smaller proportion from the larger proportion to obtain the answer, .1112. Thus, only .11, or 11 percent, of all gestation periods will be between 245 and 255 days.

Warning: Enter Table Only with Single *z* Score

When solving problems with two z scores, as above, resist the temptation to subtract one z score directly from the other and to enter Table A with this difference. Table A is designed only for individual z scores, not for differences between z scores.

Progress Check 5.4 The problem above can be solved in another way, using entries from column B′ rather than column C′. Visualize this alternative solution as a graph of the normal curve, and verify that, even though column B′ is used, the answer still equals .1112.

Example: Finding Proportions *beyond* Two Scores

Assume that high school students' IQ scores approximate a normal distribution with a mean of 105 and a standard deviation of 15. What proportion of IQs are more than 30 points either above or below the mean?

1. **Sketch a normal curve and shade in the two target areas,** as in the top panel of **Figure 5.8**.
2. **Plan your solution according to the normal table.** The solution to this type of problem is straightforward because each of the target areas can be

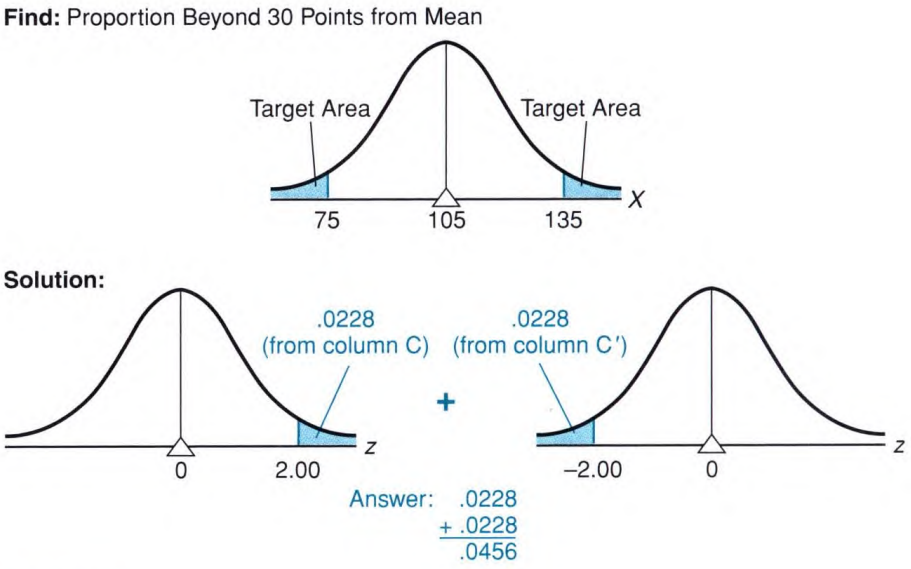


FIGURE 5.8
Finding proportions.

read directly from Table A. The target area in the tail to the right can be obtained from column C, and that in the tail to the left can be obtained from column C′, as shown in the bottom two panels of Figure 5.8.

3. **Convert X to z** by expressing IQ scores of 135 and 75 as

$$z = \frac{135 - 105}{15} = \frac{30}{15} = 2.00$$

$$z = \frac{75 - 105}{15} = \frac{-30}{15} = -2.00$$

4. **Find the target area.** In Table A, locate a z score of 2.00 in column A, and note the corresponding proportion of .0228 in column C. Because of the symmetry of the normal curve, you need not enter the table again to find the proportion below a z score of –2.00. Instead, merely double the above proportion of .0228 to obtain .0456, which represents the proportion of students with IQs more than 30 points either above or below the mean.

Semantic Alert

"*More* than 30 points either above or below the mean" translates into two target areas, one in each tail of the normal curve. "*Within* 30 points either above or below the mean" translates into two entirely new target areas corresponding to the two unshaded sectors in Figure 5.8. Each of these "within" sectors shares a common boundary at the mean, but one sector extends 30 points above the mean and the other sector extends 30 points below the mean.

Progress Check *5.5 Assume that SAT math scores approximate a normal curve with a mean of 500 and a standard deviation of 100.

(a) Sketch a normal curve and shade in the target area(s) described by each of the following statements:

(a_1) *more* than 570

(a_2) *less* than 515

(a_3) *between* 520 and 540

(a_4) between 470 and 520

(a_5) more than 50 points above the mean

(a_6) more than 100 points either above or below the mean

(a_7) within 50 points either above or below the mean

(b) Plan solutions (in terms of columns B, C, B′, and C′) for the target areas in part (a).

(c) Convert to z scores and find the target areas in part (a).

Answers on page 495.

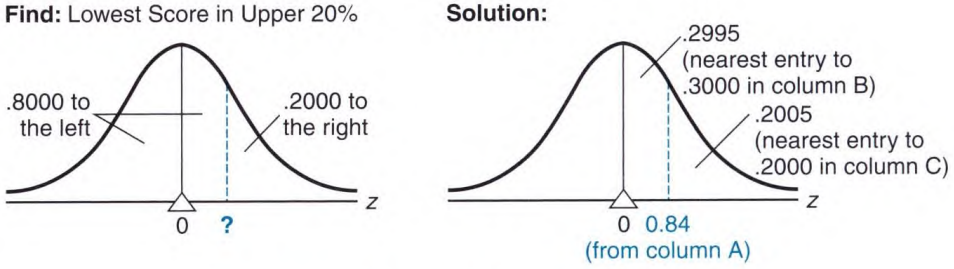
5.6 FINDING SCORES

So far, we have concentrated on normal curve problems for which Table A must be consulted to find the unknown proportion (of area) associated with some known score or pair of known scores. For instance, given a GRE score of 650, we found that the unknown proportion of scores larger than 650 equals .07. *Now we will concentrate on the opposite type of normal curve problem for which Table A must be consulted to find the unknown score or scores associated with some known proportion.* For instance, given that a GRE score must be in the upper 25 percent of the distribution (in order for an applicant to be considered for admission to graduate school), we must find the unknown minimum GRE score. Essentially, this type of problem requires that we reverse our use of Table A by entering proportions in columns B, C, B′, or C′ and finding z scores listed in columns A or A′.

Example: Finding *One* Score

Exam scores for a large psychology class approximate a normal curve with a mean of 230 and a standard deviation of 50. Furthermore, students are graded "on a curve," with only the upper 20 percent being awarded grades of A. What is the lowest score on the exam that receives an A?

1. **Sketch a normal curve and, on the correct side of the mean, draw a line representing the target score,** as in **Figure 5.9**. This is often the most difficult step, and it involves semantics rather than statistics. It's often helpful to visualize the target score as splitting the total area into two sectors—one to the left of (below) the target score and one to the right of (above) the target score. For example, in the present case, the target score is the point along the base of the curve that splits the total area into 80 percent, or .8000 to the left, and 20 percent, or .2000 to the right. The mean of a normal curve serves as a valuable frame of reference since it always splits the total area into two equal halves—.5000 to the left of the mean and .5000 to the right of the mean. Since more than .5000—that



Answer: $X = \mu + (z)(\sigma)$
$= 230 + (0.84)(50)$
$= 230 + 42$
$= 272$

FIGURE 5.9
Finding scores.

is, .8000—of the total area is to the left of the target score, this score must be on the upper or right side of the mean. On the other hand, if less than .5000 of the total area had been to the left of the target score, this score would have been placed on the lower or left side of the mean.

2. **Plan your solution according to the normal table.** In problems of this type, you must plan how to find the z score for the target score. Because the target score is on the right side of the mean, concentrate on the area in the upper half of the normal curve, as described in columns B and C. The right panel of Figure 5.9 indicates that either column B or C can be used to locate a z score in column A. It is crucial, however, to search for the single value (.3000) that is valid for column B or the single value (.2000) that is valid for column C. Note that we look in column B for .3000, not for .8000. Table A is not designed for sectors, such as the lower .8000, that span the mean of the normal curve.
3. **Find z.** Refer to Table A. Scan column C to find .2000. If this value does not appear in column C, as typically will be the case, approximate the desired value (and the correct score) by locating the entry in column C nearest to .2000. If adjacent entries are equally close to the target value, use either entry—it is your choice. As shown in the right panel of Figure 5.9, the entry in column C closest to .2000 is .2005, and the corresponding z score in column A equals 0.84. Verify this by checking Table A. Also note that exactly the same z score of 0.84 would have been identified if column B had been searched to find the entry (.2995) nearest to .3000. The z score of 0.84 represents the point that separates the upper 20 percent of the area from the rest of the area under the normal curve.
4. **Convert z to the target score.** Finally, convert the z score of 0.84 into an exam score, given a distribution with a mean of 230 and a standard deviation of 50. You'll recall that a z score indicates how many standard deviations the original score is above or below its mean. In the present case, the target score must be located .84 of a standard deviation above its mean. The distance of the target score above its mean equals 42 ($.84 \times 50$), which, when added to the mean of 230, yields a value of 272. Therefore, 272 is the lowest score on the exam that receives an A.

When converting z scores to original scores, you will probably find it more efficient to use the following equation (derived from the z score equation on page 103):

CONVERTING z SCORE TO ORIGINAL SCORE

$$X = \mu + (z)(\sigma) \tag{5.2}$$

in which X is the target score, expressed in original units of measurement; μ and σ are the mean and the standard deviation, respectively, for the original normal curve; and z is the standard score read from column A or A′ of Table A. When appropriate numerical substitutions are made, as shown in the bottom of Figure 5.9, 272 is found to be the answer, in agreement with our earlier conclusion.

Comment: Place Target Score on Correct Side of Mean

When finding scores, it is crucial that the target score be placed on the correct side of the mean. This placement dictates how the normal table will be read—whether down from the top legend, with entries in column A interpreted as positive z scores, or up from the bottom legend, with entries in column A′ interpreted as negative z scores. In the previous problem, the incorrect placement of the target score on the left side of the mean would have led to a z score of –0.84, rather than 0.84, and an erroneous answer of 188 (230 – 42), rather than the correct answer of 272 (230 + 42).

To make correct placements, you must properly interpret the specifications for the target score. Expand potentially confusing one-sided specifications, such as the "upper 20 percent, or upper .2000," into "left .8000 and right .2000." Having identified the left and right areas of the target score, which sum to 1.0000, you can compare the specifications of the target score with those of the mean. Remember that the mean of a normal curve always splits the total area into .5000 to the left of the mean and .5000 to the right of the mean. Accordingly, if the area to the left of the target score is more than .5000, the target score should be placed on the upper or right side of the mean. Otherwise, if the area to the left of the target score is less than .5000, the target score should be placed on the lower or left side of the mean.

Example: Finding *Two* Scores

Assume that the annual rainfall in the San Francisco area approximates a normal curve with a mean of 22 inches and a standard deviation of 4 inches. What are the rainfalls for the more atypical years, defined as the driest 2.5 percent of all years and the wettest 2.5 percent of all years?

1. **Sketch a normal curve. On either side of the mean, draw two lines representing the two target scores,** as in **Figure 5.10**. The smaller (driest) target score splits the total area into .0250 to the left and .9750 to the right, and the larger (wettest) target score does the exact opposite.

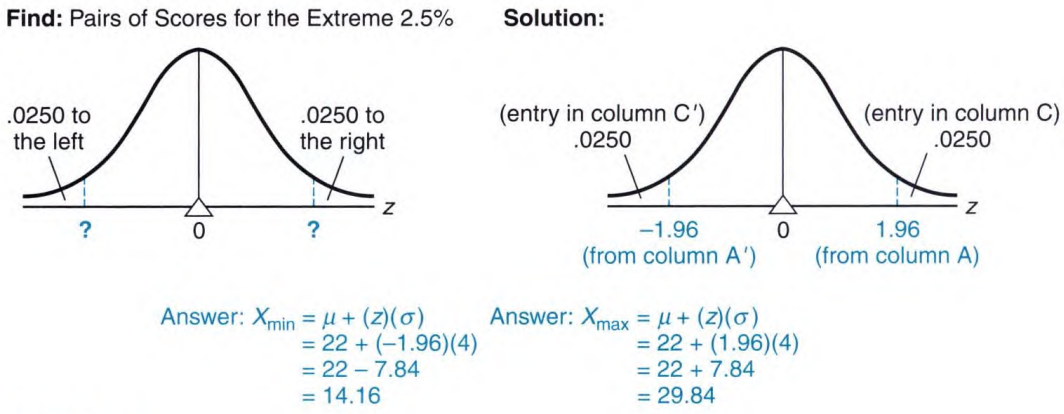


FIGURE 5.10
Finding scores.

2. **Plan your solution according to the normal table.** Because the smaller target score is located on the lower or left side of the mean, we will concentrate on the area in the lower half of the normal curve, as described in columns B′ and C′. The target z score can be found by scanning either column B′ for .4750 or column C′ for .0250. After finding the smaller target score, we will capitalize on the symmetrical properties of normal curves to find the value of the larger target score.
3. **Find z.** Referring to Table A, we can scan column B′ for .4750, or the entry nearest to .4750. In this case, .4750 appears in column B′, and the corresponding z score in column A′ equals −1.96. The same z score of −1.96 would have been obtained if column C′ had been searched for a value of .0250.
4. **Convert z to the target score.** When the appropriate numbers are substituted in Formula 5.2, as shown in the bottom panel of Figure 5.10, the smaller target score equals 14.16 inches, the amount of annual rainfall that separates the driest 2.5 percent of all years from all of the other years.

The location of the larger target score is the mirror image of that for the smaller target score. Therefore, we need not even consult Table A to establish that its z score equals 1.96—that is, the same value as the smaller target score, but without the negative sign. When 1.96 is converted to inches of rainfall, as shown in the bottom of Figure 5.10, the larger target equals 29.84 inches, the amount of annual rainfall that separates the wettest 2.5 percent of all years from all other years.

Comment: Common and Rare Events

In the above problem, we drew attention to the atypical, or rare years, by concluding that 2.5 percent of the driest years registered less than 14.16 inches of rainfall, whereas 2.5 percent of the wettest years registered more than 29.84 inches. Had we wished, we could also have drawn attention to the typical, or common years, by concluding that the most moderate, "middle" 95 percent of all years registered between 14.16 and 29.84 inches of rainfall. The middle 95 percent straddles the line perpendicular to the mean, or 50th percentile, with half, or 47.5 percent, above this line and the other half, or 47.5 percent, below this line.

Later in inferential statistics, we'll judge whether, for instance, an observed mean difference is real or transitory. As you'll see, this decision will depend on whether the one observed mean difference can be viewed as a common outcome or as a rare outcome in the distribution of all possible mean differences that could happen just by chance. Since common events tend to be identified with the middle 95 percent of the area under the normal curve and rare events with the extreme 2.5 percent in each tail, you'll often use z scores of ± 1.96 in inferential statistics.

Progress Check *5.6 Assume that the burning times of electric light bulbs approximate a normal curve with a mean of 1200 hours and a standard deviation of 120 hours. If a large number of new lights are installed at the same time (possibly along a newly opened freeway), at what time will

(a) 1 percent fail? (**Hint:** This splits the total area into .0100 to the left and .9900 to the right.)

(b) 50 percent fail?

(c) 95 percent fail?

(d) If a new inspection procedure eliminates the weakest 8 percent of all lights before they are marketed, the manufacturer can safely offer customers a money-back guarantee on all lights that fail before _____ hours of burning time.

Answers on page 495.

DOING NORMAL CURVE PROBLEMS

Read the problem carefully to determine whether a proportion or a score is to be found.

----------FINDING PROPORTIONS ----------

1. Sketch the normal curve and shade in the target area.

Examples: One Area Two Areas

2. Plan the solution in terms of the normal table.

3. Convert *X* to *z*: $z = \dfrac{X - \mu}{\sigma}$

4. Find the target area by entering either column A or A′ with *z*, and noting the corresponding proportion from column B, C, B′, or C′.

----------FINDING SCORES----------

1. Sketch the normal curve and, on the correct side of the mean, draw a line representing the target score.

Examples: To Left of Mean (area to left less than .5000) To Right of Mean (area to left more than .5000)

2. Plan the solution in terms of the normal table.

3. Find *z* by locating the entry nearest to that desired in column B, C, B′, or C′ and reading out the corresponding *z* score.

4. Convert *z* to the target score: $X = \mu + (z)(\sigma)$

Guidelines for Normal Curve Problems

Reminder:
Refer to the "Doing Normal Curve Problems" box when doing normal curve problems.

You now have the necessary information for solving most normal curve problems, but there is no substitute for actually working problems, such as those offered at the end of this chapter. For your convenience, a complete set of guidelines appears in the "Doing Normal Curve Problems" box on page 118. Before reading on, spend a few moments studying it, and then refer back to it whenever necessary.

5.7 MORE ABOUT *z* SCORES

z Scores for Non-normal Distributions

z scores are not limited to normal distributions. Non-normal distributions also can be transformed into sets of unit-free, standardized z scores. *In this case, the standard normal table cannot be consulted*, since the shape of the distribution of z scores is the same as that for the original non-normal distribution. For instance, if the original distribution is positively skewed, the distribution of z scores also will be positively skewed. *Regardless of the shape of the distribution, the shift to z scores always produces a distribution of standard scores with a mean of 0 and a standard deviation of 1.*

Interpreting Test Scores

Under most circumstances, z scores provide efficient descriptions of relative performance on one or more tests. Without additional information, it is meaningless to know that Sharon earned a raw score of 159 on a math test, but it is very informative to know that she earned a z score of 1.80. The latter score suggests that she did relatively well on the math test, being almost two standard deviation units above the mean. More precise interpretations of this score could be made, of course, if it is known that the test scores approximate a normal curve.

The use of z scores can help you identify a person's relative strengths and weaknesses on several different tests. For instance, **Table 5.2** shows Sharon's scores on college achievement tests in three different subjects. The evaluation of her test performance is greatly facilitated by converting her raw scores into the z scores listed in the final column of Table 5.2. A glance at the z scores suggests that although she did relatively well on the math test, her performance on the English test was only slightly above average, as indicated by a z score of 0.50, and her performance on the psychology test was slightly below average, as indicated by a z score of –0.67.

Table 5.2
SHARON'S ACHIEVEMENT TEST SCORES

SUBJECT	RAW SCORE	MEAN	STANDARD DEVIATION	*z* SCORE
Math	159	141	10	1.80
English	83	75	16	0.50
Psych	23	27	6	–0.67

Importance of Reference Group

Remember that z scores reflect performance relative to some group rather than an absolute standard. A meaningful interpretation of z scores requires, therefore, that the nature of the reference group be specified. In the present example, it is important to know whether Sharon's scores were relative to those of the other students at her college or to those of students at a wide variety of colleges, as well as to any other special characteristics of the reference group.

Progress Check *5.7 Convert each of the following test scores to z scores:

	TEST SCORE	MEAN	STANDARD DEVIATION
(a)	53	50	9
(b)	38	40	10
(c)	45	30	20
(d)	28	20	20

Progress Check *5.8

(a) Referring to Question 5.7, which one test score would you prefer?

(b) Referring to Question 5.7, if you had earned a score of 64 on some test, which of the four distributions (a, b, c, or d) would have permitted the most favorable interpretation of this score?

Answers on page 495.

Standard Scores

Standard Score
Any unit-free scores expressed relative to a known mean and a known standard deviation.

Whenever any unit-free scores are expressed relative to a known mean and a known standard deviation, they are referred to as **standard scores.** Although z scores qualify as standard scores because they are unit-free and expressed relative to a known mean of 0 and a known standard deviation of 1, other scores also qualify as standard scores.

Transformed Standard Scores

Transformed Standard Score
A standard score that, unlike a z score, usually lacks negative signs and decimal points.

Being by far the most important standard score, z scores are often viewed as synonymous with standard scores. *For convenience, particularly when reporting test results to a wide audience, z scores can be changed to* **transformed standard scores,** *other types of unit-free standard scores that lack negative signs and decimal points.* These transformations change neither the shape of the original distribution nor the relative standing of any test score within the distribution. For example, a test score located one standard deviation below the mean might be reported not as a z score of –1.00 but as a T score of 40 in a distribution of T scores with a mean of 50 and a standard deviation of 10. The important point to realize is that although reported as a score of 40, this T score accurately reflects the relative location of the original z score of –1.00: A T score of 40 is located at a distance of one standard deviation (of size 10) below the mean (of size 50). **Figure 5.11** shows the values of some of the more common types of transformed standard scores relative to the various portions of the area under the normal curve.

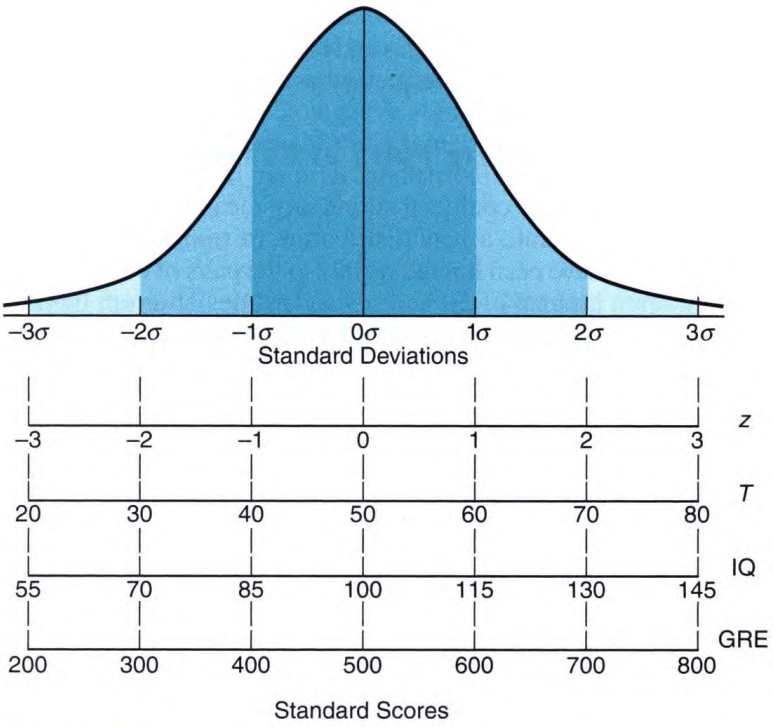


FIGURE 5.11
Common transformed standard scores associated with normal curves.

Converting to Transformed Standard Scores

Use the following formula to convert any original standard score, z, into a transformed standard score, z′, having a distribution with any desired mean and standard deviation.

TRANSFORMED STANDARD SCORE

$$z' = \text{desired mean} + (z)\,(\text{desired standard deviation}) \qquad (5.3)$$

where z' (called *z prime*) is the transformed standard score and z is the original standard score.

For instance, if you wish to convert a z score of −1.50 into a new distribution of z' scores for which the desired mean equals 500 and the desired standard deviation equals 100, substitute these numbers into the above formula to obtain

$$\begin{aligned} z' &= 500 + (-1.50)\,(100) \\ &= 500 - 150 \\ &= 350 \end{aligned}$$

Again, notice that the transformed standard score accurately reflects the relative location of the original standard score of −1.50: The transformed score of 350 is located at a distance of 1.5 standard deviation units (each of size 100) below the

mean (of size 500). The change from a z score of -1.50 to a z' score of 350 eliminates negative signs and decimal points without distorting the relative location of the original score, expressed as a distance from the mean in standard deviation units.

Substitute Pairs of Convenient Numbers

You could substitute any mean or any standard deviation in Formula 5.3 to generate a new distribution of transformed scores. Traditionally, substitutions have been limited mainly to the pairs of convenient numbers shown in Figure 5.11: a mean of 50 and a standard deviation of 10 (T scores), a mean of 100 and a standard deviation of 15 (IQ scores), and a mean of 500 and a standard deviation of 100 (GRE scores). The substitution of other arbitrary pairs of numbers serves no purpose; indeed, because of their peculiarity, they might make the new distribution, even though it lacks the negative signs and decimal points common to z scores, slightly less comprehensible to people who have been exposed to the traditional pairs of numbers.

Progress Check *5.9 Assume that each of the raw scores listed below originates from a distribution with the specified mean and standard deviation. After converting each raw score into a z score, transform each z score into a series of new standard scores with means and standard deviations of 50 and 10, 100 and 15, and 500 and 100, respectively. (In practice, you would transform a particular z into only one new standard score.)

	RAW SCORE	MEAN	STANDARD DEVIATION
(a)	24	20	5
(b)	37	42	3

Answers on page 494.

Summary

Many observed frequency distributions approximate the well-documented normal curve, an important theoretical curve noted for its symmetrical bell-shaped form. The normal curve can be used to obtain answers to a wide variety of questions.

Although there are infinite numbers of normal curves, each with its own mean and standard deviation, there is only one standard normal curve, with its mean of 0 and its standard deviation of 1. Only the standard normal curve is actually tabled. The standard normal table (Table A in Appendix C), requires the use of z scores, that is, original scores expressed as deviations, in standard deviation units, above or below its mean.

There are two general types of normal curve problems: (1) those that require you to find the unknown proportion (of area) associated with some score or pair of scores and (2) those that require you to find the unknown score or scores associated with some area. Answers to the first type of problem usually require you to convert original scores into z scores (Formula 5.1), and answers to the second type of problem usually require you to translate a z score back into an original score (Formula 5.2).

Even when distributions fail to approximate normal curves, z scores can provide efficient descriptions of relative performance on one or more tests.

When reporting test results, z scores are often transformed into other types of standard scores that lack negative signs and decimal points. These conversions change neither the shape of the original distribution nor the relative standing of any test score within the original distribution.

Important Terms

Normal curve
z score
Standard score
Standard normal curve
Transformed standard score

Key Equations

z SCORE

$$z = \frac{X - \mu}{\sigma}$$

CONVERTING z TO X

$$X = \mu + z\sigma$$

REVIEW QUESTIONS

***5.10** Fill in the blank spaces.

To identify a particular normal curve, you must know the __(a)__ and __(b)__ for that distribution. To convert a particular normal curve to the standard normal curve, you must convert original scores into __(c)__ scores. A z score indicates how many __(d)__ a score is __(e)__ or __(f)__ the mean of the distribution. Although there are infinite numbers of normal curves, there is __(g)__ standard normal curve. The standard normal curve has a __(h)__ of 0 and a __(i)__ of 1.

The total area under the standard normal curve equals __(j)__. When using the standard normal table, it is important to remember that for any z score, the corresponding proportions in columns B and C (or columns B′ and C′) always sum to __(k)__. Furthermore, the proportion in column B (or B′) always specifies the proportion of area between the __(l)__ and the z score, while the proportion in column C (or C′) always specifies the proportion of area __(m)__ the z score. Although any z score can be either positive or negative, the proportions of area, specified in columns B and C (or columns B′ and C′), are never __(n)__.

Standard scores are unit-free scores expressed relative to a known __(o)__ and __(p)__. The most important standard score is a __(q)__ score. Unlike z

scores, transformed standard scores usually lack __(r)__ signs and __(s)__ points. Transformed standard scores accurately reflect the relative standing of the original __(t)__ score.

Answers on page 496.

Finding Proportions

5.11 Scores on the Wechsler Adult Intelligence Scale (WAIS) approximate a normal curve with a mean of 100 and a standard deviation of 15. What proportion of IQ scores are

(a) above 125?

(b) below 82?

(c) within 9 points of the mean?

(d) more than 40 points from the mean?

5.12 Suppose that the burning times of electric light bulbs approximate a normal curve with a mean of 1200 hours and a standard deviation of 120 hours. What proportion of lights burn for

(a) less than 960 hours?

(b) more than 1500 hours?

(c) within 50 hours of the mean?

(d) between 1300 and 1400 hours?

Finding Scores

5.13 IQ scores on the WAIS test approximate a normal curve with a mean of 100 and a standard deviation of 15. What IQ score is identified with

(a) the upper 2 percent, that is, 2 percent to the right (and 98 percent to the left)?

(b) the lower 10 percent?

(c) the upper 60 percent?

(d) the middle 95 percent? [Remember, the middle 95 percent straddles the line perpendicular to the mean (or the 50th percentile), with half of 95 percent, or 47.5 percent, above this line and the remaining 47.5 percent below this line.]

(e) the middle 99 percent?

Finding Proportions and Scores

IMPORTANT NOTE: When doing Questions **5.14** and **5.15,** remember to decide first whether a proportion or a score is to be found.

***5.14** An investigator polls common cold sufferers, asking them to estimate the number of hours of physical discomfort caused by their most recent colds. Assume that their estimates approximate a normal curve with a mean of 83 hours and a standard deviation of 20 hours.

(a) What is the estimated number of hours for the shortest-suffering 5 percent?

(b) What proportion of sufferers estimate that their colds lasted longer than 48 hours?

(c) What proportion suffered for fewer than 61 hours?

(d) What is the estimated number of hours suffered by the extreme 1 percent either above or below the mean?

(e) What proportion suffered for between 1 and 3 days, that is, between 24 and 72 hours?

(f) What is the estimated number of hours suffered by the middle 95 percent? [See the comment about "middle 95 percent" in Question **5.13(d)**.]

(g) What proportion suffered for between 2 and 4 days?

(h) A medical researcher wishes to concentrate on the 20 percent who suffered the most. She will work only with those who estimate that they suffered for more than _____ hours.

(i) Another researcher wishes to compare those who suffered least with those who suffered most. If each group is to consist of only the extreme 3 percent, the mild group will consist of those who suffered for fewer than _____ hours, and the severe group will consist of those who suffered for more than _____ hours.

(j) Another survey found that people with colds who took daily doses of vitamin C suffered, on the average, for 61 hours. What proportion of the original survey (with a mean of 83 hours and a standard deviation of 20 hours) suffered for more than 61 hours?

(k) What proportion of the original survey suffered for *exactly* 61 hours? (Be careful!)

Answers on page 496.

5.15 Admission to a state university depends partially on the applicant's high school GPA. Assume that the applicants' GPAs approximate a normal curve with a mean of 3.20 and a standard deviation of 0.30.

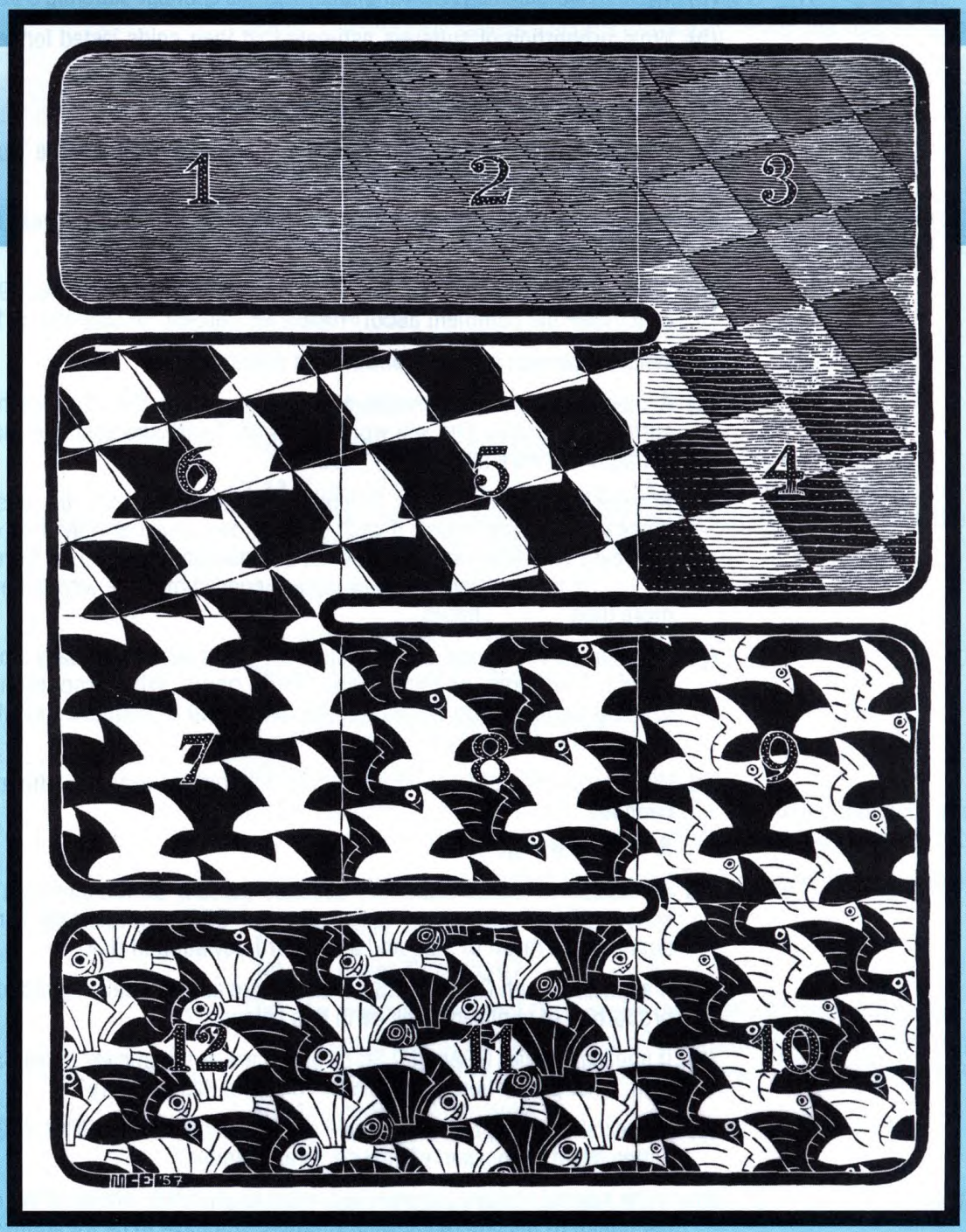
(a) If applicants with GPAs of 3.50 or above are automatically admitted, what proportion of applicants will be in this category?

(b) If applicants with GPAs of 2.50 or below are automatically denied admission, what proportion of applicants will be in this category?

(c) A special honors program is open to all applicants with GPAs of 3.75 or better. What proportion of applicants are eligible?

(d) If the special honors program is limited to students whose GPAs rank in the upper 10 percent, what will Brittany's GPA have to be for admission to this program?

5.16 When describing test results, someone objects to the conversion of raw scores into standard scores, claiming that this constitutes an arbitrary change in the value of the test score. How might you respond to this objection?


1
2
3
4
5
6
7
8
9
10
11
12
M C E 57

CHAPTER 6

Describing Relationships: Correlation

Preview

Is there a relationship between your IQ and the wealth of your parents? Between your computer skills and your GPA? Between your anxiety level and your perceived social attractiveness? Answers to these questions require us to describe the relationship between pairs of variables. The original data must consist of actual pairs of observations, such as, for example, IQ scores and parents' wealth for each member of the freshman class. Two variables are related if pairs of scores show an orderliness that can be depicted graphically with a ***scatterplot*** *and numerically with a* ***correlation coefficient.***

Table 6.1
GREETING CARDS SENT AND RECEIVED BY FIVE FRIENDS

FRIEND	NUMBER OF CARDS SENT	NUMBER OF CARDS RECEIVED
Andrea	5	10
Mike	7	12
Doris	13	14
Steve	9	18
John	1	6

Does the familiar saying "You get what you give" accurately describe the exchange of holiday greeting cards? An investigator suspects that a relationship exists between the number of greeting cards *sent* and the number of greeting cards *received* by individuals. Prior to a full-fledged survey—and also prior to any statistical analysis based on variability, as described later in Section 15.9—he obtains the estimates for the most recent holiday season from five friends, as shown in **Table 6.1**. (The data in Table 6.1 represent a very simple observational study with two dependent variables, as defined in Section 1.6, since numbers of cards sent and received are not under the investigator's control.)

6.1 AN INTUITIVE APPROACH

If the suspected relationship does exist between cards sent and cards received, then an inspection of the data might reveal, as one possibility, a tendency for "big senders" to be "big receivers" and for "small senders" to be "small receivers." More generally, there is a tendency for pairs of scores to occupy similar relative positions in their respective distributions.

Positive Relationship

Trends among pairs of scores can be detected most easily by constructing a list of paired scores in which the scores along one variable are arranged from largest to smallest. In panel A of **Table 6.2**, the five pairs of scores are arranged from the largest (13) to the smallest (1) number of cards sent. This table reveals a pronounced tendency for pairs of scores to occupy similar *relative* positions in their respective distributions. For example, John sent relatively few cards (1) and received relatively few cards (6), whereas Doris sent relatively many cards (13) and received relatively many cards (14). We can conclude, therefore, that the two variables are related. Furthermore, this relationship implies that "You get what you give." *Insofar as relatively low values are paired with relatively low values, and relatively high values are paired with relatively high values, the relationship is* **positive.**

In panels B and C of Table 6.2, each of the five friends continues to send the same number of cards as in panel A, but new pairs are created to illustrate two other possibilities—a negative relationship and little or no relationship. (In real applications, of course, the pairs are fixed by the data and cannot be changed.)

Positive Relationship
Occurs insofar as pairs of scores tend to occupy similar relative positions (high with high and low with low) in their respective distributions.

Negative Relationship

Notice the pattern among the pairs in panel B. Now there is a pronounced tendency for pairs of scores to occupy dissimilar and opposite relative positions in their respective distributions. For example, although John sent relatively few cards (1), he received relatively many (18). From this pattern, we can conclude that the two variables are related. Furthermore, this relationship implies that "You get the opposite of what you give." *Insofar as relatively low values are paired with relatively high values, and relatively high values are paired with relatively low values, the relationship is* **negative.**

Negative Relationship
Occurs insofar as pairs of scores tend to occupy dissimilar relative positions (high with low and vice versa) in their respective distributions.

Table 6.2
THREE TYPES OF RELATIONSHIPS

A. POSITIVE RELATIONSHIP

FRIEND	SENT	RECEIVED
Doris	13	14
Steve	9	18
Mike	7	12
Andrea	5	10
John	1	6

B. NEGATIVE RELATIONSHIP

FRIEND	SENT	RECEIVED
Doris	13	6
Steve	9	10
Mike	7	14
Andrea	5	12
John	1	18

C. LITTLE OR NO RELATIONSHIP

FRIEND	SENT	RECEIVED
Doris	13	10
Steve	9	18
Mike	7	12
Andrea	5	6
John	1	14

Little or No Relationship

No regularity is apparent among the pairs of scores in panel C. For instance, although both Andrea and John sent relatively few cards (5 and 1, respectively), Andrea received relatively few cards (6) and John received relatively many cards (14). Given this lack of regularity, we can conclude that little, if any, relationship exists between the two variables and that "What you get has no bearing on what you give."

Review

Whether we are concerned about the relationship between cards sent and cards received, years of heavy smoking and life expectancy, educational level and annual income, or scores on a vocational screening test and subsequent ratings as a police officer,

> **two variables are *positively* related if pairs of scores tend to occupy similar relative positions (*high with high and low with low*) in their respective distributions,**
> **and they are *negatively* related if pairs of scores tend to occupy dissimilar relative positions (*high with low and vice versa*) in their respective distributions.**

The remainder of this chapter deals with how best to describe and interpret a relationship between pairs of variables. The intuitive method of searching for regularity among pairs of scores is cumbersome and inexact when the analysis involves more than a few pairs of scores. Although this technique has much appeal, it must be abandoned in favor of several other, more efficient and exact statistical techniques, namely, a special graph known as a *scatterplot* and a measure known as a *correlation coefficient.*

It will become apparent in the next chapter that once a relationship has been identified, it can be used for predictive purposes. Having established that years of heavy smoking is negatively related to length of life (because heavier smokers tend to have shorter lives), we can use this relationship to predict the life expectancy of someone who has smoked heavily for the past 10 years. This type of prediction could serve a variety of purposes, such as calculating a life insurance premium or supplying extra motivation in an antismoking workshop.

Progress Check ***6.1** Indicate whether the following statements suggest a positive or negative relationship:

(a) More densely populated areas have higher crime rates.

(b) Schoolchildren who often watch TV perform more poorly on academic achievement tests.

(c) Heavier automobiles yield poorer gas mileage.

(d) Better-educated people have higher incomes.

(e) More anxious people voluntarily spend more time performing a simple repetitive task.

Answers on page 496.

6.2 SCATTERPLOTS

Scatterplot
A graph containing a cluster of dots that represents all pairs of scores.

A **scatterplot** *is a graph containing a cluster of dots that represents all pairs of scores.* With a little training, you can use any dot cluster as a preview of a fully measured relationship.

Construction

To construct a scatterplot, as in **Figure 6.1**, scale each of the two variables along the horizontal (X) and vertical (Y) axes, and use each pair of scores to locate a dot within the scatterplot. For example, the pair of numbers for Mike, 7 and 12, define points along the X and Y axes, respectively. Using these points to anchor lines perpendicular (at right angles) to each axis, locate Mike's dot where the two lines intersect. Repeat this process, with imaginary lines, for each of the four remaining pairs of scores to create the scatterplot of Figure 6.1.

Our simple example involving greeting cards has shown the basic idea of correlation and the construction of a scatterplot. Now we'll examine more complex sets of data in order to learn how to interpret scatterplots.

Positive, Negative, or Little or No Relationship?

The first step is to note the tilt or slope, if any, of a dot cluster. *A dot cluster that has a slope from the lower left to the upper right,* as in panel A of **Figure 6.2**, *reflects a positive relationship.* Small values of one variable are paired with small values of the other variable, and large values are paired with large values. In panel A, short people tend to be light, and tall people tend to be heavy.

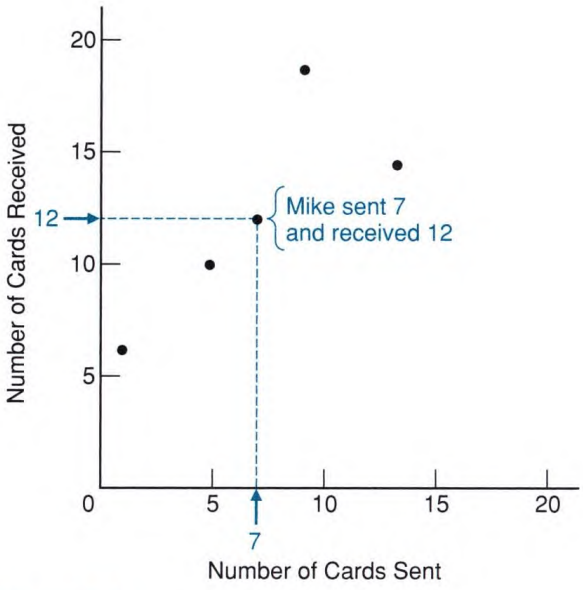


FIGURE 6.1
Scatterplot for greeting card exchange.

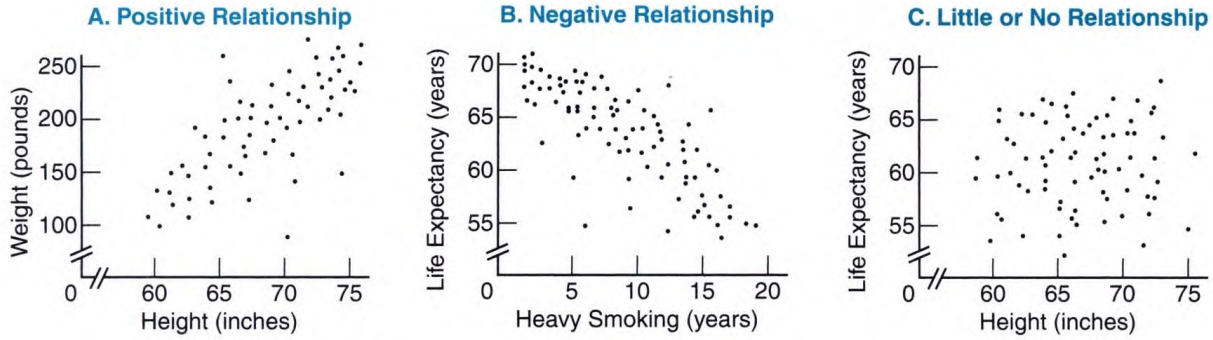


FIGURE 6.2
Three types of relationships.

On the other hand, *a dot cluster that has a slope from the upper left to the lower right,* as in panel B of Figure 6.2, *reflects a negative relationship.* Small values of one variable tend to be paired with large values of the other variable, and vice versa. In panel B, people who have smoked heavily for few years or not at all tend to have longer lives, and people who have smoked heavily for many years tend to have shorter lives.

Finally, *a dot cluster that lacks any apparent slope,* as in panel C of Figure 6.2, *reflects little or no relationship.* Small values of one variable are just as likely to be paired with small, medium, or large values of the other variable. In panel C, notice that the dots are strewn about in an irregular shotgun fashion, suggesting that there is little or no relationship between the height of young adults and their life expectancies.

Strong or Weak Relationship?

Having established that a relationship is either positive or negative, note how closely the dot cluster approximates a straight line. *The more closely the dot cluster approximates a straight line, the stronger (the more regular) the relationship will be.* **Figure 6.3** shows a series of scatterplots, each representing a

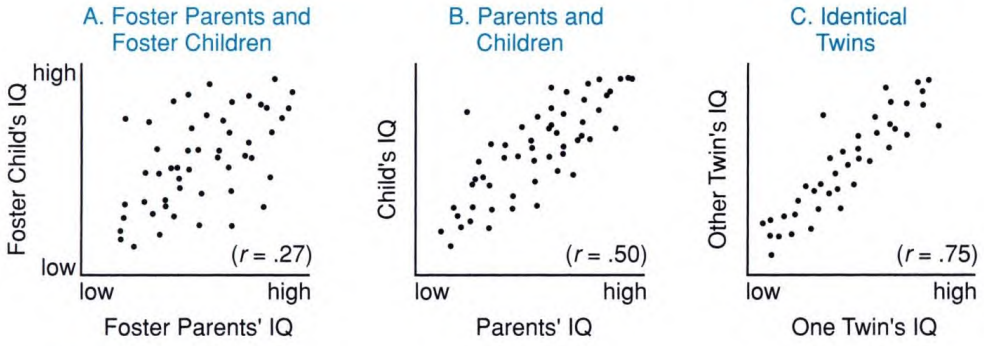


FIGURE 6.3
Three positive relationships. (Scatterplots simulated from a 50-year literature survey.) Source: L. Erlenmeyer-Kimling and L. F. Jarvik. "Genetics and Intelligence: A Review." Science, 142, 1477–1479.)

different positive relationship between IQ scores for pairs of people whose backgrounds reflect different degrees of genetic overlap, ranging from minimum overlap between foster parents and foster children to maximum overlap between identical twins. (Ignore the parenthetical expressions involving *r*, to be discussed later.) Notice that the dot cluster more closely approximates a straight line for people with greater degrees of genetic overlap—for parents and children in panel B of Figure 6.3 and even more so for identical twins in panel C.

Perfect Relationship

A dot cluster that equals (rather than merely approximates) a straight line reflects a perfect relationship between two variables. In practice, perfect relationships are most unlikely.

Curvilinear Relationship

Linear Relationship
A relationship that can be described best with a straight line.

Curvilinear Relationship
A relationship that can be described best with a curved line.

The previous discussion assumes that a dot cluster approximates a *straight* line and, therefore, reflects a **linear relationship.** But this is not always the case. Sometimes a dot cluster approximates a *bent* or *curved* line, as in **Figure 6.4**, and therefore reflects a **curvilinear relationship.** Descriptions of these relationships are more complex than those of linear relationships. For instance, we see in Figure 6.4 that physical strength, as measured by the force of a person's handgrip, is less for children, more for adults, and then less again for older people. Otherwise, the scatterplot can be interpreted as before, that is, the more closely the dot cluster approximates a curved line, the stronger the curvilinear relationship will be.

Look again at the scatterplot in Figure 6.1 for the greeting card data. Although the small number of dots in Figure 6.1 hinders any interpretation, the dot cluster appears to approximate a straight line, stretching from the lower left to the upper right. This suggests a positive relationship between greeting cards sent and received, in agreement with the earlier intuitive analysis of these data.

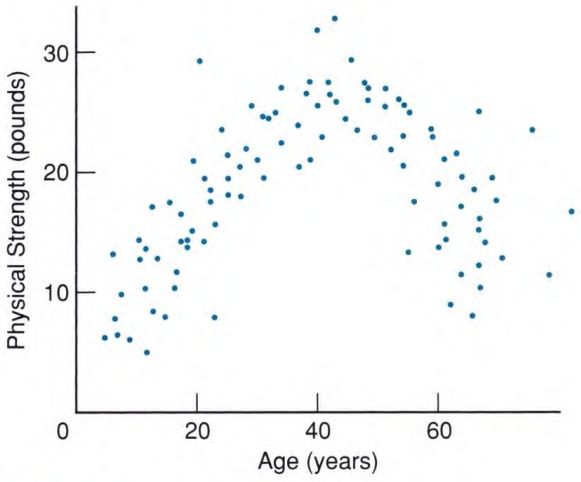
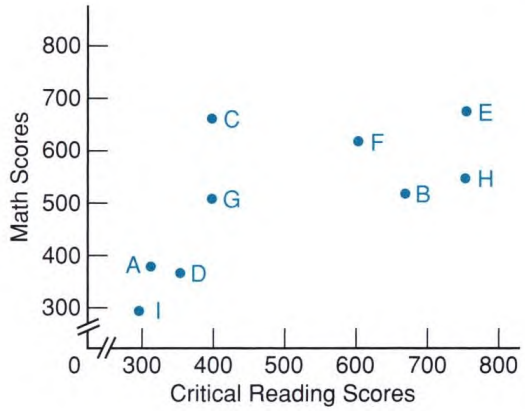


FIGURE 6.4
Curvilinear relationship.

Progress Check *6.2 Critical reading and math scores on the SAT test for students A, B, C, D, E, F, G, and H are shown in the following scatterplot:



(a) Which student(s) scored about the same on both tests?

(b) Which student(s) scored higher on the critical reading test than on the math test?

(c) Which student(s) will be eligible for an honors program that requires minimum scores of 700 in critical reading and 500 in math?

(d) Is there a negative relationship between the critical reading and math scores?

Answers on page 496.

6.3 A CORRELATION COEFFICIENT FOR QUANTITATIVE DATA: *r*

Correlation Coefficient
A number between –1 and 1 that describes the relationship between pairs of variables.

A *correlation coefficient* is a number between –1 and 1 that describes the relationship between pairs of variables.

The next few sections concentrate on the type of correlation coefficient, designated as *r*, that *describes the linear relationship between pairs of variables for quantitative data*. Many other types of correlation coefficients have been introduced to handle specific types of data, including ranked and qualitative data, and a few of these will be described briefly in Section 6.7.

Key Properties of *r*

Pearson Correlation Coefficient (r)
A number between –1.00 and +1.00 that describes the linear relationship between pairs of quantitative variables.

Named in honor of the British scientist Karl Pearson, the **Pearson correlation coefficient, *r*,** can equal any value between –1.00 and +1.00 . Furthermore, the following two properties apply:

1. *The sign of* r *indicates the type of linear relationship, whether positive or negative.*
2. *The numerical value of* r, *without regard to sign, indicates the strength of the linear relationship.*

Sign of *r*

A number with a plus sign (or no sign) indicates a positive relationship, and a number with a minus sign indicates a negative relationship. For example, an *r* with a plus sign describes the positive relationship between height and weight shown in panel A of Figure 6.2, and an *r* with a minus sign describes the negative relationship between heavy smoking and life expectancy shown in panel B.

Numerical Value of *r*

The more closely a value of *r* approaches either –1.00 or +1.00, the stronger (more regular) the relationship. Conversely, the more closely the value of *r* approaches 0, the weaker (less regular) the relationship. For example, an *r* of –.90 indicates a stronger relationship than does an *r* of –.70, and an *r* of –.70 indicates a stronger relationship than does an *r* of .50. (Remember, if no sign appears, it is understood to be plus.) In Figure 6.3, notice that the values of *r* shift from .75 to .27 as the analysis for pairs of IQ scores shifts from a relatively strong relationship for identical twins to a relatively weak relationship for foster parents and foster children.

From a slightly different perspective, the value of *r* is a measure of how well a straight line (representing the linear relationship) describes the cluster of dots in the scatterplot. Again referring to Figure 6.3, notice that an imaginary straight line describes the dot cluster less well as the values of *r* shift from .75 to .27.

WWW

INTERNET DEMONSTRATION
Go to the Web site for this book (http://www.wiley.com/college/witte). Click on the *Student Companion Site,* then *Internet Demonstrations,* and finally ***Guessing Correlations*** to practice matching values of correlation coefficients with various scatterplots.

Interpretation of *r*

Located along a scale from –1.00 to +1.00, the value of *r* supplies information about the direction of a linear relationship—whether positive or negative—and, generally, information about the relative strength of a linear relationship—whether relatively weak (and a poor describer of the data) because *r* is in the vicinity of 0, or relatively strong (and a good describer of the data) because *r* deviates from 0 in the direction of either +1.00 or –1.00.

If, as usually is the case, we wish to generalize beyond the limited sample of actual paired scores, *r* can't be interpreted at face value. Viewed as the product of chance sampling variability (see Section 15.9), the value of *r* must be evaluated with tools from inferential statistics to establish whether the relationship is real or merely transitory. This evaluation depends not only on the value of *r* but also on the actual number of pairs of scores used to calculate *r*. On the assumption that reasonably large numbers of pairs of scores are involved (preferably

hundreds and certainly many more than the five pairs of scores in our purposely simple greeting card example), an r of .50 or more, in either the positive or the negative direction, would represent a *very strong* relationship in most areas of behavioral and educational research.* But there are exceptions. An r of at least .80 or more would be expected when correlation coefficients measure "test reliability," as determined, for example, from pairs of IQ scores for people who take the same IQ test twice or take two forms of the same test (to establish that any person's two scores tend to be similar and, therefore, that the test scores are reproducible, or "reliable").

Range Restrictions

Except for special circumstances, the value of the correlation coefficient declines whenever the range of possible X or Y scores is restricted. Range restriction is analogous to magnifying a subset of the original dot cluster and, in the process, losing much of the orderly and predictable pattern in the original dot cluster. For example, **Figure 6.5** shows a dot cluster with an obvious slope, represented by an r of .70 for the positive relationship between height and weight for all college students. If, however, the range of heights along Y is restricted to students who stand over 6 feet 2 inches (or 74 inches) tall, the abbreviated dot cluster loses its obvious slope because of the more homogeneous weights among tall students. Therefore, as depicted in Figure 6.5, the value of r drops to .10.

Sometimes it's impossible to avoid a range restriction. For example, some colleges only admit students with SAT test scores above some minimum value.

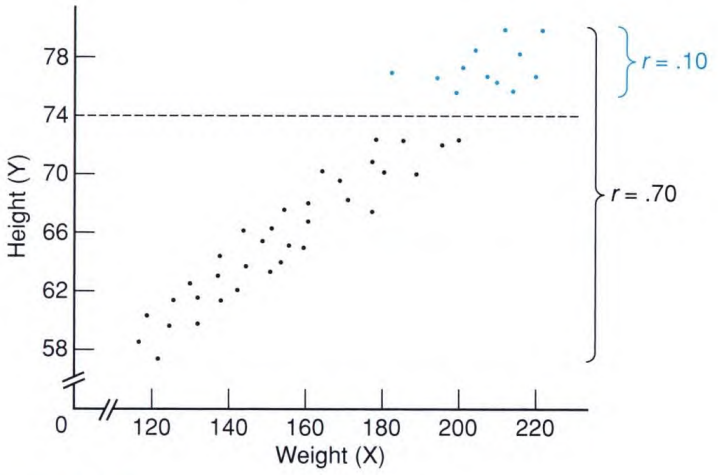


FIGURE 6.5
Effect of range restriction on the value of r.

* In his landmark book *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed. (Hillsdale, NJ: Erlbaum, 1988), Jacob Cohen suggests that a value of r in the vicinity of .10 or less reflects a small (weak) relationship; a value in the vicinity of .30 reflects a medium (moderate) relationship; and a value in the vicinity of .50 or more reflects a large (strong) relationship.

Subsequently, the value of any correlation between SAT scores and college GPAs for these students will be lower because of the absence of any students with SAT scores below the minimum score required for admission. Always check for any possible restriction on the ranges of X or Y scores—whether by design or accident—that could lower the value of r.

Caution

Be careful when interpreting the actual numerical value of r. An r of .70 for height and weight doesn't signify that the strength of this relationship equals either .70 or 70 percent of the strength of a perfect relationship. *The value of* r *can't be interpreted as a proportion or percentage of some perfect relationship.*

Verbal Descriptions

When interpreting a brand new r, you'll find it helpful to translate the numerical value of r into a verbal description of the relationship. An r of .70 for the height and weight of college students could be translated into "Taller students tend to weigh more" (or some other equally valid statement, such as "Lighter students tend to be shorter"); an r of –.42 for time spent taking an exam and the subsequent exam score could be translated into "Students who take less time tend to make higher scores"; and an r in the neighborhood of 0 for shoe size and IQ could be translated into "Little, if any, relationship exists between shoe size and IQ."

If you have trouble verbalizing the value of r, refer back to the original scatterplot or, if necessary, visualize a rough scatterplot corresponding to the value of r. Use any detectable dot cluster to think your way through the relationship. Does the dot cluster have a slope from the lower left to the upper right—that is, does low go with low and high go with high? Or does the dot cluster have a slope from the upper left to the lower right—that is, does low go with high and vice versa? It is crucial that you translate abstractions such as "Low goes with low and high goes with high" into concrete terms such as "Shorter students tend to weigh less, and taller students tend to weigh more."

Progress Check *6.3 Supply a verbal description for each of the following correlations. (If necessary, visualize a rough scatterplot for r, using the scatterplots in Figure 6.3 as a frame of reference.)

(a) an r of –.84 between total mileage and automobile resale value

(b) an r of –.35 between the number of days absent from school and performance on a math achievement test

(c) an r of .03 between anxiety level and college GPA

(d) an r of .56 between age of schoolchildren and reading comprehension

Answers on page 496.

Correlation Not Necessarily Cause-Effect

Given a correlation between the prevalence of poverty and crime in U.S. cities, you can *speculate* that poverty causes crime—that is, poverty produces crime with the same degree of inevitability as the flip of a light switch illuminates a room.

According to this view, any widespread reduction in poverty should cause a corresponding decrease in crime. As suggested in Chapter 1, you can also *speculate* that a common cause, such as inadequate education, overpopulation, racial discrimination, etc., or some combination of these factors produces both poverty and crime. According to this view, a widespread reduction in poverty should have no effect on crime. Which speculation is correct? Unfortunately, this issue cannot be resolved merely on the basis of an observed correlation.

A correlation coefficient, regardless of size, never provides information about whether an observed relationship reflects a simple cause-effect relationship or some more complex state of affairs.

In the past, the interpretation of the correlation between cigarette smoking and lung cancer was vigorously disputed. American Cancer Society representatives interpreted the correlation as a causal relationship: Smoking produces lung cancer. On the other hand, tobacco industry representatives interpreted the correlation as, at most, an indication that both the desire to smoke cigarettes and lung cancer are caused by some more basic but yet unidentified factor or factors, such as the body metabolism or personality of some people. According to this reasoning, people with a high body metabolism might be more prone to smoke and, quite independent of their smoking, more vulnerable to lung cancer. Therefore, smoking correlates with lung cancer because both are effects of some common cause or causes.

Role of Experimentation

Sometimes experimentation can resolve this kind of controversy. In the present case, laboratory animals were trained to inhale different amounts of tobacco tars and were then euthanized. Autopsies revealed that the observed incidence of lung cancer (the dependent variable) varied directly with the amount of inhaled tobacco tars (the independent variable), even though possible "contaminating" factors, such as different body metabolisms or personalities, had been neutralized either through experimental control or by random assignment of the subjects to different test conditions. As was noted in Chapter 1, experimental confirmation of a correlation can provide strong evidence in favor of a cause-effect interpretation of the observed relationship; indeed, in the smoking-cancer controversy, cumulative experimental findings overwhelmingly support the conclusion that smoking causes lung cancer.

Progress Check ***6.4** Speculate on whether the following correlations reflect simple cause-effect relationships or more complex states of affairs. (**Hint:** A cause-effect relationship implies that, if all else remains the same, any change in the causal variable should always produce a predictable change in the other variable.)

(a) caloric intake and body weight

(b) height and weight

(c) SAT math score and score on a calculus test

(d) poverty and crime

Answers on page 497.

6.4 DETAILS: *z* SCORE FORMULA FOR *r*

The simplest formula for r is:

Reminder:
Always use Formula 6.2 on p. 141 to calculate r.

CORRELATION COEFFICIENT (*z* SCORE FORMULA)

$$r = \frac{\Sigma z_x z_y}{n-1} \qquad (6.1)$$

where z_x and z_y are the z score equivalents for each pair of original scores, X and Y, and n refers to the number of pairs of scores. The term $\Sigma z_x z_y$ is found by first multiplying each pair of z_x and z_y scores and then adding the products for all pairs. Formula 6.1 directs us to divide $\Sigma z_x z_y$ by the total number of pairs minus one, $n - 1$. (The $n - 1$ in the denominator reflects the fact that the z scores are calculated using sample standard deviations.)

Calculating *r* (*z* Score Formula)

In actual practice, you should *never* use the z score formula to calculate the value of r because of several complications, including the extra effort of converting original scores to z scores. (For example, to obtain Doris's z_x score of 1.34 in panel A of **Table 6.3**, subtract 7, the sample mean $\bar{X}$, from 13, her X score, then divide the resulting deviation of 6 by 4.47, the sample standard deviation s_x. Similarly, to obtain her z_y of 0.45, subtract 12, the sample mean $\bar{Y}$, from

Table 6.3
THREE RELATIONSHIPS: *z* SCORE FORMULA FOR *r*

	A POSITIVE RELATIONSHIP			B NEGATIVE RELATIONSHIP			C LITTLE OR NO RELATIONSHIP		
	z_x	z_y	$z_x z_y$	z_x	z_y	$z_x z_y$	z_x	z_y	$z_x z_y$
Doris	1.34	0.45	0.60	1.34	−0.45	−0.60	1.34	−0.45	−0.60
Steve	0.45	1.34	0.60	0.45	−1.34	−0.60	0.45	1.34	0.60
Mike	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Andrea	−0.45	−0.45	0.20	−0.45	0.45	−0.20	−0.45	−1.34	0.60
John	−1.34	−1.34	1.80	1.34	−1.34	−1.80	−1.34	0.45	−0.60
			$\Sigma z_x z_y = 3.20$			$\Sigma z_x z_y = -3.20$			$\Sigma z_x z_y = 0.00$
	$r = \frac{\Sigma z_x z_y}{n-1}$			$r = \frac{\Sigma z_x z_y}{n-1}$			$r = \frac{\Sigma z_x z_y}{n-1}$		
	$= \frac{3.20}{4}$			$= \frac{-3.20}{4}$			$= \frac{0}{4}$		
	$r = .80$			$r = -.80$			$r = 0$		

14, her Y score, then divide the resulting deviation of 2 by 4.47, the sample standard deviation s_y.)

In order to clarify some important properties of r, Table 6.3 illustrates how, once z scores replace original scores for the three sets of paired scores in Table 6.2, values of r are calculated for each of the three types of relationships for the greeting card example. Let's look more closely at how Formula 6.1 processes a positive relationship, a negative relationship, and little or no relationship.

Positive Relationship

If there is a positive relationship, as in panel A of Table 6.3, pairs of z scores will tend to have similar relative locations in their respective distributions, and therefore positive z scores will tend to be paired with positive z scores and negative z scores will tend to be paired with negative z scores. Consequently, and this is a crucial point, the products of paired z scores, $z_x z_y$, will tend to be positive because multiplication involves pairs of numbers with like signs, either both plus or both minus. As a result, the numerator term in Formula 6.1, $\Sigma z_x z_y$, becomes a relatively large positive number, which, when divided by $n - 1$, yields a positive r—in the present case, an r of .80.

Perfect Positive Relationship

If pairs of z scores have both the same magnitude and the same sign (for instance, one pair might be 1.34 and 1.34, and another might be 0.45 and 0.45), r would equal 1.00, indicating a perfect positive relationship. Under these circumstances, the relationship would display total regularity, with pairs of z scores occupying *exactly* the same relative locations in their respective distributions. Returning to panel A of Table 6.3, you might verify that if the top two entries in the z_x column had been reversed, causing each pair of z scores to have both the same magnitude and the same sign, then

$$r = \frac{\Sigma z_x z_y}{n-1} = \frac{4.00}{4} = 1.00$$

Negative Relationship

If there is a negative relationship, as in panel B of Table 6.3, pairs of z scores will tend to have relative locations that are reversed in their respective distributions, and therefore positive z scores will tend to be paired with negative scores, and vice versa. Consequently, the products of paired z scores, $z_x z_y$, will tend to be negative because multiplication involves pairs of numbers with unlike signs, one plus and the other minus. As a result, the numerator term in Formula 6.1, $\Sigma z_x z_y$, will tend to be a relatively large negative number, which, when divided by $n - 1$, will be a negative r—in the present case, an r of –.80.

Perfect Negative Relationship

If pairs of z scores have the same magnitude but unlike signs (for instance, one pair might be 1.34 and –1.34, and another pair might be 0.45 and –0.45), r would equal –1.00, indicating a perfect negative correlation. Under these circumstances, the relationship would also display perfect regularity, with pairs of

z scores occupying relative locations that are *exactly reversed* in their respective distributions.

Little or No Relationship

If there is little or no relationship, as in panel C of Table 6.3, no consistent pattern will describe the relative locations of pairs of z scores in their respective distributions. Therefore, a positive z score is equally likely to be paired with either a positive or a negative z score, and vice versa. Consequently, about half of all products of paired z scores, $z_x z_y$, are positive because multiplication involves numbers with like signs, and about half of all products of paired z scores, $z_x z_y$, are negative because multiplication involves numbers with unlike signs. Since positive and negative products tend to cancel each other, the numerator term in Formula 6.1, $\Sigma z_x z_y$, tends toward a small positive or negative number that, when divided by $n - 1$, yields a value of r near 0.

Review

To summarize, an understanding of correlation, as measured by r, can be gained from the z score formula. The pattern among pairs of z scores can be used to anticipate the value of r. If pairs of z scores are similar in both magnitude and sign, the value of r will tend toward 1.00, indicating a strong positive correlation. But if pairs of z scores are similar in magnitude but opposite in sign, the value of r will tend toward –1.00, indicating a strong negative correlation. As the pattern among pairs of z scores becomes less apparent, the value of r tends toward 0, indicating a weak or nonexistent correlation.

r Is Independent of Units of Measurement

The z score formula also pinpoints another important property of r—its independence of the original units of measurement. In fact, the same value of r describes the correlation between height and weight for a group of adults, regardless of whether height is measured in inches or centimeters or whether weight is measured in pounds or grams. In effect, the value of r depends only on the pattern among pairs of z scores, which in turn show no traces of the units of measurement for the original X and Y scores. If you think about it, this is the same as saying that

a positive value of *r* reflects a tendency for pairs of scores to occupy *similar* relative locations (high with high and low with low) in their respective distributions, while a negative value of *r* reflects a tendency for pairs of scores to occupy *dissimilar* relative locations (high with low and vice versa) in their respective distributions.

Progress Check *6.5 Pretend that there is a perfect positive (+1.00) relationship between height and weight for adults. (Actually, you might recall, it's in the vicinity of .70 for college students.) In this case, if John stands two standard deviations above the mean height, his weight will be ___(a)___ standard deviation units ___(b)___ the mean. If Kristen is one and one-half standard deviation units below the mean height, her weight will be ___(c)___ standard deviation units ___(d)___ the mean. If Carson is one-third of a standard deviation above the mean weight, his height will be ___(e)___ of a standard deviation ___(f)___ the mean.

Progress Check *6.6 Repeat Question 6.5, assuming a perfect negative (–1.00) relationship between height and weight.

Answers on page 497.

6.5 DETAILS: COMPUTATION FORMULA FOR *r*

Except to provide a more intuitive picture of correlation, *the z score formula should not be used to calculate r*. Converting each *X* and *Y* score into an equivalent *z* score not only is laborious but often produces an appreciable rounding error if means and standard deviations are approximate numbers. It is more efficient and usually more accurate to calculate a value for *r* by using the computation formula:

Reminder:
Use this formula to calculate r.

CORRELATION COEFFICIENT (COMPUTATION FORMULA)

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} \tag{6.2}$$

where the two sum of squares terms in the denominator are defined as

$$SS_x = \Sigma\left(X - \overline{X}\right)^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

$$SS_y = \Sigma\left(Y - \overline{Y}\right)^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$$

and the sum of the products term in the numerator, SP_{xy}, is defined in Formula 6.3.

SUM OF PRODUCTS (DEFINITION AND COMPUTATION FORMULAS)

$$SP_{xy} = \Sigma\left(X - \overline{X}\right)\left(Y - \overline{Y}\right) = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n} \tag{6.3}$$

In the case of SP_{xy}, instead of summing the squared deviation scores for either *X* or *Y*, as with SS_x and SS_y, we find the sum of the products for each pair of deviation score. Notice in Formula 6.2 that, since the terms in the denominator must be positive, only the sum of the products, SP_{xy}, determines whether the value of *r* is positive or negative. Furthermore, the size of SP_{xy} mirrors that of $\Sigma z_x z_y$ discussed in the previous section; stronger relationships are associated with larger positive or negative sums of products. **Table 6.4** illustrates the calculation of *r* for the original greeting card data by using the computation formula.

Progress Check *6.7 Couples who attend a clinic for first pregnancies are asked to estimate (independently of each other) the ideal number of children. Given that *X* and *Y* represent the estimates of females and males, respectively, the

Table 6.4
CALCULATION OF *r*: COMPUTATION FORMULA

A. COMPUTATIONAL SEQUENCE

Assign a value to *n* 1, representing the number of pairs of scores.
Sum all scores for *X* 2 and for *Y* 3.
Find the product of each pair of *X* and *Y* scores 4, one at a time, then add all of these products 5.
Square each *X* score 6, one at a time, then add all squared *X* scores 7.
Square each *Y* score 8, one at a time, then add all squared *Y* scores 9.
Substitute numbers into formulas 10 and solve for SP_{xy}, SS_x, and SS_y.
Substitute into formula 11 and solve for *r*.

B. DATA AND COMPUTATIONS

	CARDS		4	6	8
FRIEND	**SENT, *X***	**RECEIVED, *Y***	***XY***	***X*²**	***Y*²**
Doris	13	14	182	169	196
Steve	9	18	162	81	324
Mike	7	12	84	49	144
Andrea	5	10	50	25	100
John	1	6	6	1	36

1 $n = 5$ 2 $\Sigma X = 35$ 3 $\Sigma Y = 60$ 5 $\Sigma XY = 484$ 7 $\Sigma X^2 = 325$ 9 $\Sigma Y^2 = 800$

10 $$SP_{xy} = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n} = 484 - \frac{(35)(60)}{5} = 484 - 420 = 64$$

$$SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 325 - \frac{(35)^2}{5} = 325 - 245 = 80$$

$$SS_y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = 800 - \frac{(60)^2}{5} = 800 - 720 = 80$$

11 $$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} = \frac{64}{\sqrt{(80)(80)}} = \frac{64}{80} = .80$$

results are as follows:

COUPLE	*X*	*Y*
A	1	2
B	3	4
C	2	3
D	3	2
E	1	0
F	2	3

Calculate a value for *r*, using the computation formula (6.2).

Answer on page 497.

6.6 OUTLIERS AGAIN

In Section 2.3, *outliers* were defined as very extreme scores that require special attention because of their potential impact on a summary of data. This is also true when outliers appear among sets of paired scores. Although quantitative techniques can be used to detect these outliers, we simply focus on dots in scatterplots that deviate conspicuously from the main dot cluster.

Greeting Card Study Revisited

Figure 6.6 shows the effect of each of two possible outliers, substituted one at a time for Doris's dot (13, 14), on the original value of r (.80) for the greeting card data. Although both outliers A and B deviate conspicuously from the dot cluster, they have radically different effects on the value of r. Outlier A (33, 34) contributes to a new value of .98 for r that merely reaffirms the original positive relationship between cards sent and received. On the other hand, outlier B (13, 4) causes a dramatically new value of .04 for r that entirely neutralizes the original positive relationship. Neither of the values for outlier B, taken singularly, is extreme. Rather, it is their unusual combination—13 cards sent and only 4 received—that yields the radically different value of .04 for r, indicating that the new dot cluster is not remotely approximated by a straight line.

Dealing with Outliers

Of course, serious investigators would use many more than five pairs of scores, and therefore the effect of outliers on the value of r would tend not to be as dramatic as the one above. Nevertheless, outliers can have a considerable impact on the value of r and, therefore, pose problems of interpretation. Unless there is some reason for discarding an outlier—because of a failed accuracy check or because, for example, you establish that the friend who received only 4 cards had sent 13 cards that failed to include an expected monetary gift—the

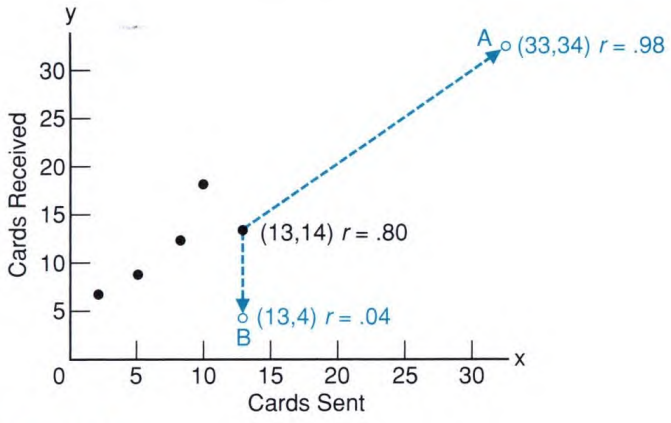


FIGURE 6.6
Effect of each of two outliers on the value of r.

most defensible strategy is to report the values of r both with and without any outliers.*

WWW

INTERNET DEMONSTRATION
Go to the Web site for this book (http://www.wiley.com/college/witte). Click on the *Student Companion Site,* then *Internet Demonstrations,* and finally ***Outliers*** to see the effect of outliers on the correlation coefficient.

6.7 OTHER TYPES OF CORRELATION COEFFICIENTS

There are many other types of correlation coefficients, but we will discuss only several that are direct descendants of the Pearson correlation coefficient. Although designed originally for use with quantitative data, the Pearson r has been extended, sometimes under the guise of new names and customized versions of Formula 6.2, to other kinds of situations. For example, to describe the correlation between *ranks* assigned independently by two judges to a set of science projects, simply substitute the numerical ranks into Formula 6.2, then solve for a value of the Pearson r (also referred to as *Spearman's rho* coefficient for ranked or ordinal data). To describe the correlation between quantitative data (for example, annual income) and *qualitative or nominal data with only two categories* (for example, male and female), assign arbitrary numerical codes, such as 1 and 2, to the two qualitative categories, then solve Formula 6.2 for a value of the Pearson r (also referred to as a *point biserial* correlation coefficient). Or to describe the relationship between *two ordered qualitative variables,* such as the attitude toward legal abortion (favorable, neutral, or opposed) and educational level (high school only, some college, college graduate), assign any *ordered* numerical codes to the categories for both qualitative variables, then solve Formula 6.2 for a value of the Pearson r (also referred to as *Cramer's phi* coefficient).

Most computer outputs would simply report each of the above correlations as a Pearson r. Given the widespread use of computers, the more specialized names for the Pearson r will probably survive, if at all, as artifacts of an earlier age, when calculations were manual and some computational relief was obtained by customizing Formula 6.2 for situations involving ranks and qualitative data.

6.8 COMPUTER OUTPUT

Most analyses in this book are performed by hand on small batches of data. When analyses are based on large batches of data, as often happens in practice, it is much more efficient to use a computer. Although we will not show how to enter commands and data into a computer, we will describe the most relevant

*For more information about the quantitative detection of outliers among sets of paired scores, see Chapter 15 in D.C. Howell's, *Statistical Methods for Psychology,* 7th ed. (Belmont, CA: Wadsworth, 2010).

portions of some computer outputs. Once you have learned to ignore irrelevant details and references to more advanced statistical procedures, you'll find that statistical results produced by computers are as easy to interpret as those produced by hand.

Three of the most widely used statistical programs, Minitab, SPSS (Statistical Package for the Social Sciences), and SAS (Statistical Analysis System), generate the computer outputs in this book. As interpretive aids, some outputs are cross-referenced with explanatory comments at the bottom of the printout. Since these outputs are based on data already analyzed by hand, computer-produced results can be compared with familiar results. For example, the computer-produced scatterplot, as well as the correlation of .800 in Table 6.5 can be compared with the manually produced scatterplot in Figure 6.1 and the correlation of .80 in Table 6.4.

WWW

INTERNET SITES

Go to the Web site for this book (http://www.wiley.com/college/witte). Click on the *Student Companion Site,* then *Internet Sites,* and finally ***Minitab, SPSS, or SAS*** to obtain more information about these statistical packages, as well as demonstration software.

Correlation Matrix

Table showing correlations for all possible pairs of variables.

When every possible pairing of variables is reported, as in lower half of the output in **Table 6.5**, a **correlation matrix** is produced. The value of .800 occurs twice in the matrix, since the correlation is the same whether the relationship is described as that between cards sent and cards received or vice versa. The value of 1.000, which also occurs twice, reflects the trivial fact that any variable correlates perfectly with itself.

Reading a Larger Correlation Matrix

Since correlation matrices can be expanded to incorporate any number of variables, they are useful devices for showing correlations between all possible pairs of variables when, in fact, many variables are being studied. For example, in **Table 6.6**, four variables generate a correlation matrix with 4×4, or 16, correlation coefficients. The four perfect (but trivial) correlations of 1.000, produced by pairing each variable with itself, split the remainder of the matrix into two triangular sections, each containing six nontrivial correlations. Since the correlations in these two sectors are mirror images, you can attend to just the values of the six correlations in one sector in order to evaluate all relevant correlations among the four original variables.

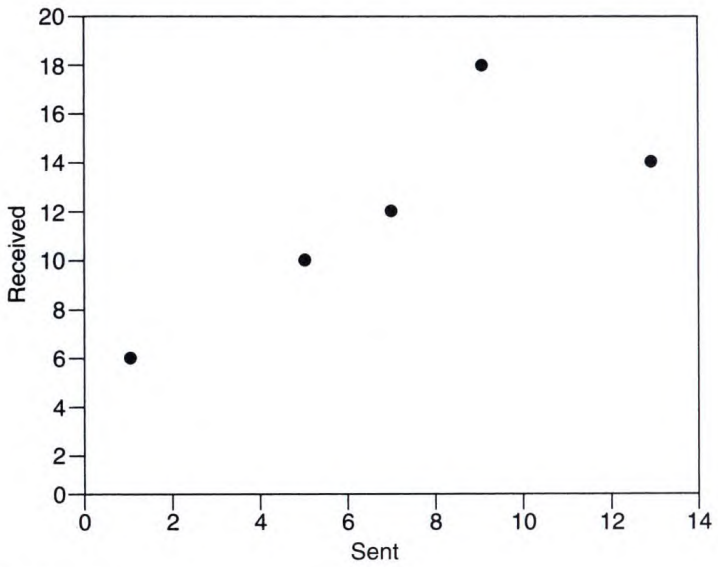
Interpreting a Larger Correlation Matrix

Three of the six color-coded correlations in Table 6.6 involve GENDER. GENDER qualifies for a correlation analysis once arbitrary numerical codes (1 for male and 2 for female) have been assigned. Looking across the bottom

Table 6.5
SPSS OUTPUT: SCATTERPLOT AND CORRELATION FOR GREETING CARD DATA

GRAPH

1



CORRELATIONS

		SENT	RECEIVED
Sent	Pearson Correlation	1.000	.800
	Sig. (2-tailed)	–	.104
	N	5	5
Received	Pearson Correlation	2 .800	1.000
	Sig. (2-tailed)	3 .104	–
	N	4 5	5

Comments:

1 Scatterplot for greeting card data (using slightly different scales than in Figure 6.1).

2 The correlation for cards sent and cards received equals .800, in agreement with the calculations in Table 6.4.

3 The value of Sig. *helps us interpret the statistical significance of a correlation by evaluating the observed value of* r *relative to the actual number of pairs of scores used to calculate* r. *Discussed later in Section 14.6,* Sig.*-values are referred to as* p*-values in this book. At this point, perhaps the easiest way to view a* Sig.*-value is as follows: The smaller the value of* Sig. *(on a scale from 0 to 1), the more likely that you would observe a correlation with the same sign, either positive or negative, if the study were repeated with new observations. Investigators often focus only on those correlations with* Sig.*-values smaller than .05.*

4 Number of cases or paired scores.

Table 6.6
SPSS OUTPUT: CORRELATION MATRIX FOR FOUR VARIABLES (BASED ON 336 STATISTICS STUDENTS)

CORRELATIONS

		AGE	COLLEGE GPA	HIGH SCHOOL GPA	GENDER
AGE	Pearson Correlation	1.000	.2228	−.0376	.0813
	Sig. (2-tailed)	—	.000	.511	.138
	N	335	333	307	335
COLLEGE GPA	Pearson Correlation	.2228	1.000	.2521	.2069
	Sig. (2-tailed)	.000	—	.000	.000
	N	333	334	306	334
HIGH SCHOOL GPA	Pearson Correlation	−.0376	.2521	1.000	.2981
	Sig. (2-tailed)	.511	.000	—	.000
	N	307	306	307	307
GENDER	Pearson Correlation	.0813	.2069	.2981	1.000
	Sig. (2-tailed)	.138	.000	.000	—
	N	335	334	307	336

row, GENDER is positively correlated with AGE (.0813); with COLLEGE GPA (.2069); and with HIGH SCHOOL GPA (.2981). Looking across the next row, HIGH SCHOOL GPA is negatively correlated with AGE (–.0376) and positively correlated with COLLEGE GPA (.2521). Lastly, COLLEGE GPA is positively correlated with AGE (.2228).

As suggested in Comment 3 at the bottom of Table 6.5, values of *Sig.* help us judge the statistical significance of the various correlations. A smaller value of *Sig.* implies that if the study were repeated, the same positive or negative sign of the corresponding correlation would probably reappear, even though calculations are based on an entirely new group of similarly selected students. Therefore, we can conclude that the four correlations with *Sig.*-values close to zero (.000) probably would reappear as positive relationships. In a new group, female students would tend to have higher high school and college GPAs, and students with higher college GPAs would tend to have higher high school GPAs and to be older. Because of the larger *Sig.*-value of .138 for the correlation between GENDER and AGE we cannot be as confident that female students would be older than male students. Because of the even larger *Sig.*-value of .511 for the small negative correlation between AGE and HIGH SCHOOL GPA, this correlation would be just as likely to reappear as either a positive or negative relationship and should not be taken seriously.

Finally, the numbers in the last row of each cell in Table 6.6 show the total number of cases actually used to calculate the corresponding correlation. Excluded from these totals are those cases in which students failed to supply the requested information.

Progress Check *6.8 Refer to Table 6.6 when answering the following questions.

(a) Would the same positive correlation of .2981 have been obtained between GENDER and HIGH SCHOOL GPA if the assignment of codes had been reversed, with females being coded as 1 and males coded as 2? Explain your answer.

(b) Given the new coding of females as 1 and males as 2, would the results still permit you to conclude that females tend to have higher high school GPAs than do males?

(c) Would the original positive correlation of .2981 have been obtained if, instead of the original coding of males as 1 and females as 2, males were coded as 10 and females as 20? Explain your answer.

(d) Assume that the correlation matrix includes a fifth variable. What would be the total number of relevant correlations in the expanded matrix?

Answers on pages 497.

Summary

The presence of regularity among pairs of X and Y scores indicates that the two variables are related, and the absence of any regularity suggests that the two variables are, at most, only slightly related. When the regularity consists of relatively low X scores being paired with relatively low Y scores and relatively high X scores being paired with relatively high Y scores, the relationship is positive. When it consists of relatively low X scores being paired with relatively high Y scores and vice versa, the relationship is negative.

A scatterplot is a graph with a cluster of dots that represents all pairs of scores. A dot cluster that has a slope from the lower left to the upper right reflects a positive relationship, and a dot cluster that has a slope from the upper left to the lower right reflects a negative relationship. A dot cluster that lacks any apparent slope reflects little or no relationship.

In a positive or negative relationship, the more closely the dot cluster approximates a straight line, the stronger the relationship will be.

When the dot cluster approximates a straight line, the relationship is linear; when it approximates a bent line, the relationship is curvilinear.

Located on a scale from –1.00 to +1.00, the value of r indicates both the direction of a linear relationship—whether positive or negative—and, generally, the relative strength of a linear relationship. Values of r in the general vicinity of either –1.00 or +1.00 indicate a relatively strong relationship, and values of r in the neighborhood of 0 indicate a relatively weak relationship.

Although the value of r can be used to formulate a verbal description of the relationship, the numerical value of r does not indicate a proportion or percentage of a perfect relationship.

Always check for any possible restriction on the ranges of X and Y scores that could lower the value of r.

The presence of a correlation, by itself, does not resolve the issue of whether it reflects a simple cause-effect relationship or a more complex state of affairs.

The Pearson correlation coefficient, r, describes the linear relationship between pairs of variables for quantitative data. An understanding of correlation,

as described by r, can be gained from the z score formula (6.1). In practice, it is both more efficient and more accurate to calculate r by using the computation formula (6.2).

Outliers can have a considerable impact on the value of r and, therefore, pose problems of interpretation.

Although designed originally for use with quantitative data, the Pearson r has been extended to other kinds of situations, including those with ranked and qualitative data.

Whenever there are more than two variables, correlation matrices can be useful devices for showing correlations between all possible pairs of variables.

Important Terms and Symbols

Positive relationship
Scatterplot
Curvilinear relationship
Pearson correlation coefficient (*r*)
Negative relationship
Linear relationship
Correlation coefficient
Correlation matrix

Key Equations

CORRELATION COEFFICIENT

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

where $SP_{xy} = \Sigma(X - \overline{X})(Y - \overline{Y}) = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}$

REVIEW QUESTIONS

6.9 **(a)** Estimate whether the following pairs of scores for X and Y reflect a positive relationship, a negative relationship, or no relationship. **Hint:** Note any tendency for pairs of X and Y scores to occupy similar or dissimilar relative locations.

X	Y
64	66
40	79
30	98
71	65
55	76
31	83
61	68
42	80
57	72

(b) Construct a scatterplot for X and Y. Verify that the scatterplot does not describe a pronounced curvilinear trend.

(c) Calculate r using the computation formula (6.2).

***6.10** On the basis of an extensive survey, the California Department of Education reported an r of –.32 for the relationship between the amount of time spent watching TV and the achievement test scores of schoolchildren. Each of the following statements represents a possible interpretation of this finding. Indicate whether each is True or False.

(a) *Every* child who watches a lot of TV will perform poorly on the achievement tests.

(b) Extensive TV viewing causes a decline in test scores.

(c) Children who watch little TV will tend to perform well on the tests.

(d) Children who perform well on the tests will tend to watch little TV.

(e) If Gretchen's TV-viewing time is reduced by one-half, we can expect a substantial improvement in her test scores.

(f) TV viewing could not possibly cause a decline in test scores.

Answers on Page 497.

6.11 Assume that an r of .80 describes the relationship between daily food intake, measured in ounces, and body weight, measured in pounds, for a group of adults. Would a shift in the units of measurement from ounces to grams and from pounds to kilograms change the value of r? Justify your answer.

6.12 An extensive correlation study indicates that a longer life is experienced by people who follow the seven "golden rules" of behavior, including moderate drinking, no smoking, regular meals, some exercise, and eight hours of sleep each night. Can we conclude, therefore, that this type of behavior *causes* a longer life?

CHAPTER 7 Regression

Preview

If two variables are correlated, description can lead to prediction. For example, if computer skills and GPAs are related, level of computer skills can be used to predict GPAs. Predictive accuracy increases with the strength of the underlying correlation.

Also discussed is a very prevalent phenomenon known as "regression toward the mean." It often occurs over time to subsets of extreme observations, such as after the superior performance of professional athletes or after the poor performance of learning-challenged children. If misinterpreted as a real effect, regression toward the mean can lead to erroneous conclusions.

A correlation analysis of the exchange of greeting cards by five friends for the most recent holiday season suggests a strong positive relationship between cards sent and cards received. When informed of these results, another friend, Emma, who enjoys receiving greeting cards, asks you to predict how many cards she will receive during the next holiday season, assuming that she plans to send 11 cards.

7.1 TWO ROUGH PREDICTIONS

Predict "Relatively Large Number"

You could offer Emma a very rough prediction by recalling that cards sent and received tend to occupy *similar* relative locations in their respective distributions. Therefore, Emma can expect to receive a *relatively large* number of cards, since she plans to send a *relatively large* number of cards.

Predict "between 14 and 18 Cards"

To obtain a slightly more precise prediction for Emma, refer to the scatterplot for the original five friends shown in **Figure 7.1**. Notice that Emma's plan to send 11 cards locates her along the X axis between the 9 cards sent by Steve

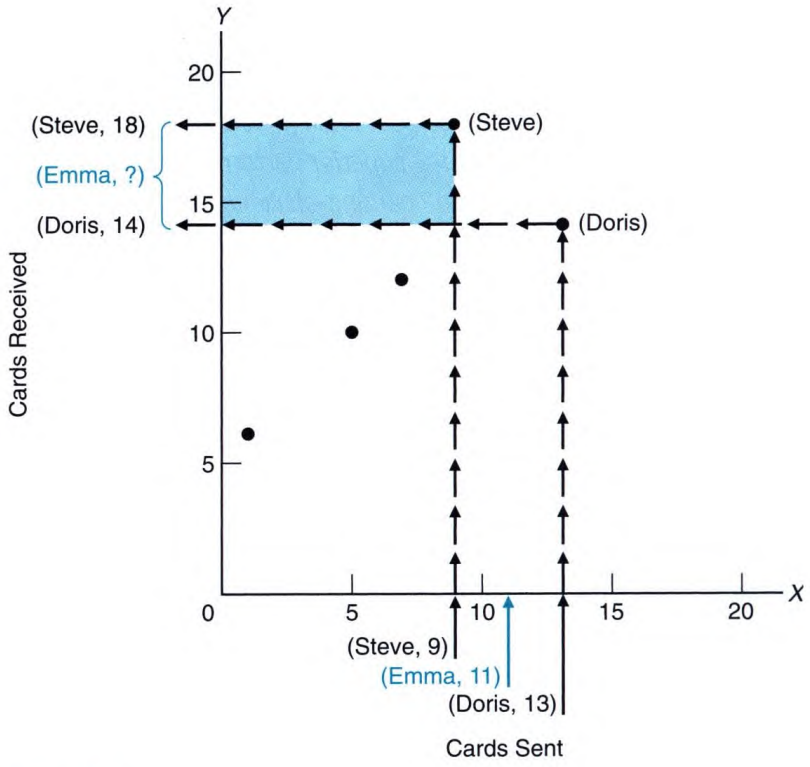


FIGURE 7.1
A rough prediction for Emma (using dots for Steve and Doris).

and the 13 sent by Doris. Using the dots for Steve and Doris as guides, construct two strings of arrows, one beginning at 9 and ending at 18 for Steve and the other beginning at 13 and ending at 14 for Doris. [The direction of the arrows reflects our attempt to predict cards received (Y) from cards sent (X). Although not required, it is customary to predict from X to Y.] Focusing on the interval along the Y axis between the two strings of arrows, you could predict that Emma's return should be between 14 and 18 cards, the numbers received by Doris and Steve.

The latter prediction might satisfy Emma, but it would not win any statistical awards. Although each of the five dots in Figure 7.1 supplies valuable information about the exchange of greeting cards, our prediction for Emma is based only on the two dots for Steve and Doris.

7.2 A REGRESSION LINE

All five dots contribute to the more precise prediction, illustrated in **Figure 7.2**, that Emma will receive 15.20 cards. Look more closely at the solid line designated as the regression line in Figure 7.2, which guides the string of arrows, beginning at 11, toward the predicted value of 15.20. The regression line is a straight line rather than a curved line because of the linear relationship between cards sent and cards received. As will become apparent, it can be used repeatedly to predict cards received. Regardless of whether Emma decides to send 5, 15, or 25 cards, it will guide a new string of arrows, beginning at 5 or 15 or 25, toward a new predicted value along the Y axis.

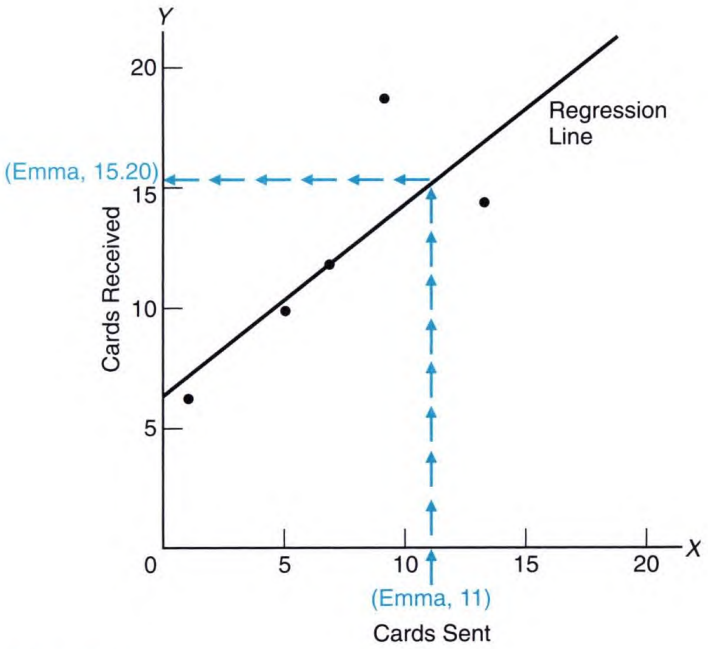


FIGURE 7.2
Prediction of 15.20 for Emma (using the regression line).

Placement of Line

For the time being, forget about any prediction for Emma and concentrate on how the five dots dictate the placement of the regression line. If all five dots had defined a single straight line, placement of the regression line would have been simple; merely let it pass through all dots. When the dots fail to define a single straight line, as in the scatterplot for the five friends, placement of the regression line represents a compromise. It passes through the main cluster, possibly touching some dots but missing others.

Predictive Errors

Figure 7.3 illustrates the predictive errors that would have occurred if the regression line had been used to predict the number of cards received by the five friends. Solid dots reflect the *actual* number of cards received, and open dots, always located along the regression line, reflect the *predicted* number of cards received. (To avoid clutter in Figure 7.3, the strings of arrows have been omitted. However, you might find it helpful to imagine a string of arrows, ending along the *Y* axis, for each dot, whether solid or open.) The largest predictive error, shown as a broken vertical line, occurs for Steve, who sent 9 cards. Although he actually received 18 cards, he should have received slightly fewer than 14 cards, according to the regression line. The smallest predictive error—none whatsoever—occurs for Mike, who sent 7 cards. He actually received the 12 cards that he should have received, according to the regression line.

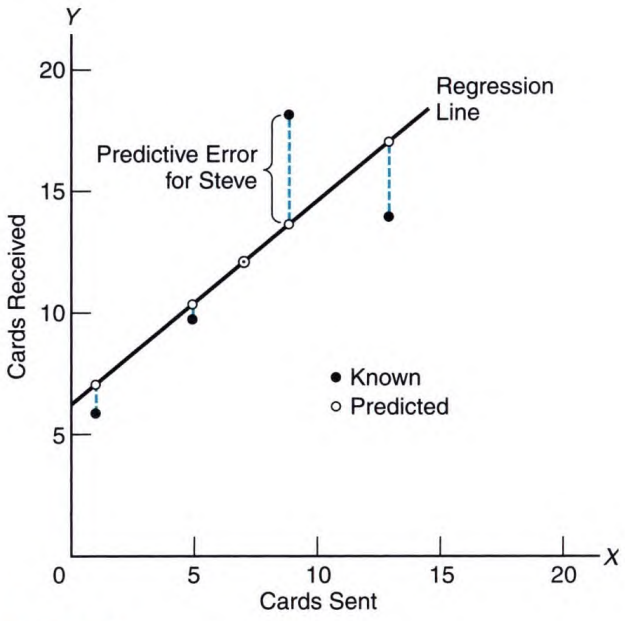
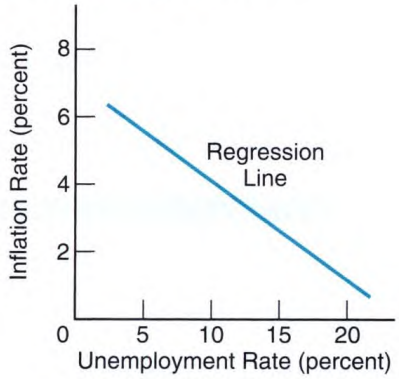


FIGURE 7.3
Predictive errors.

Total Predictive Error

We engage in the seemingly silly activity of predicting what is known already for the five friends to check the adequacy of our predictive effort. The smaller the total for all predictive errors in Figure 7.3, the more favorable will be the prognosis for our predictions. Clearly, it is desirable for the regression line to be placed in a position that *minimizes* the total predictive error, that is, that minimizes the total of the vertical discrepancies between the solid and open dots shown in Figure 7.3.

Progress Check *7.1 To check your understanding of the first part of this chapter, make predictions using the graph below.



(a) Predict the approximate rate of inflation, given an unemployment rate of 5 percent.

(b) Predict the approximate rate of inflation, given an unemployment rate of 15 percent.

Answers on page 497.

7.3 LEAST SQUARES REGRESSION LINE

To avoid the arithmetic standoff of zero always produced by adding positive and negative predictive errors (associated with errors above and below the regression line, respectively), *the placement of the regression line minimizes* not the total predictive error but *the total squared predictive error*, that is, the total for all squared predictive errors. When located in this fashion, the *regression* line is often referred to as the *least squares regression line*. Although more difficult to visualize, this approach is consistent with the original aim—to minimize the total predictive error or some version of the total predictive error, thereby providing a more favorable prognosis for our predictions.

Need a Mathematical Solution

Without the aid of mathematics, the search for a least squares regression line would be frustrating. Scatterplots would be proving grounds cluttered with tentative regression lines, discarded because of their excessively large totals for

squared discrepancies. Even the most time-consuming, conscientious effort would culminate in only a close approximation to the least squares regression line.

WWW

INTERNET DEMONSTRATION
Go to the Web site for this book (http://www.wiley.com/college/witte) and click on **Regression** to try fitting by eye the least squares regression line to a cluster of dots.

Least Squares Regression Equation

Happily, an equation pinpoints the exact least squares regression line for any scatterplot. Most generally, this equation reads:

LEAST SQUARES REGRESSION EQUATION

$$Y' = bX + a \tag{7.1}$$

where Y' represents the predicted value (the predicted number of cards that will be received by any new friend, such as Emma); X represents the known value (the known number of cards sent by any new friend); and b and a represent numbers calculated from the original correlation analysis, as described below.*

Finding Values of *b* and *a*

To obtain a working regression equation, solve each of the following expressions, first for b and then for a, using data from the original correlation analysis. The expression for b reads:

SOLVING FOR *b*

$$b = \sqrt{\frac{SS_y}{SS_x}}r \tag{7.2}$$

where SS_y represents the sum of squares for all Y scores (the cards received by the five friends); SS_x represents the sum of squares for all X scores (the cards sent by the five friends); and r represents the correlation between X and Y (cards sent and received by the five friends).

The expression for a reads:

*You might recognize that the least squares equation describes a straight line with a slope of b and a Y-intercept of a.

SOLVING FOR *a*

$$a = \bar{Y} - b\bar{X} \tag{7.3}$$

where $\bar{Y}$ and $\bar{X}$ refer to the sample means for all Y and X scores, respectively, and b is defined by the preceding expression.

The values of all terms in the expressions for b and a can be obtained from the original correlation analysis either directly, as with the value of r, or indirectly, as with the values of the remaining terms: SS_y, SS_x, $\bar{Y}$, and $\bar{X}$. **Table 7.1** illustrates the computational sequence that produces a least squares regression equation for the greeting card example, namely,

$$Y' = .80\,(X) + 6.40$$

where .80 and 6.40 represent the values computed for b and a, respectively.

Table 7.1
DETERMINING THE LEAST SQUARES REGRESSION EQUATION

A. COMPUTATIONAL SEQUENCE

Determine values of SS_x, SS_y, and r 1 by referring to the original correlation analysis in Table 6.4.
Substitute numbers into the formula 2 and solve for b.
Assign values to $\bar{X}$ and $\bar{Y}$ 3 by referring to the original correlation analysis in Table 6.4.
Substitute numbers into the formula 4 and solve for a.
Substitute numbers for b and a in the least squares regression equation 5.

B. COMPUTATIONS

1 $SS_x = 80^*$
$SS_y = 80^*$
$r = .80$

2 $$b = \sqrt{\frac{SS_y}{SS_x}}(r) = \sqrt{\frac{80}{80}}(.80) = .80$$

3 $\bar{X} = 7^{**}$
$\bar{Y} = 12^{**}$

4 $a = \bar{Y} - (b)(\bar{X}) = 12 - (.80)(7) = 12 - 5.60 = 6.40$

5 $Y' = (b)(X) + a$
$= (.80)(X) + 6.40$

* *Computations not shown. Verify, if you wish, using Formula 4.4.*
** *Computations not shown. Verify, if you wish, using Formula 3.1.*

Least Squares Regression Equation

The equation that minimizes the total of all squared prediction errors for known Y scores in the original correlation analysis.

Key Property

Once numbers have been assigned to b and a, as just described, the **least squares regression equation** emerges as a working equation with a most desirable property: It automatically *minimizes the total of all squared predictive errors for known Y scores in the original correlation analysis.*

Solving for Y'

In its present form, the regression equation can be used to predict the number of cards that Emma will receive, assuming that she plans to send 11 cards. Simply substitute 11 for X and solve for the value of Y' as follows:

$$\begin{aligned} Y' &= .80(11) + 6.40 \\ &= 8.80 + 6.40 \\ &= 15.20 \end{aligned}$$

Notice that the predicted card return for Emma, 15.20, qualifies as a genuine prediction, that is, a forecast of an unknown event based on information about some known event. This prediction appeared earlier in Figure 7.2.

Our working regression equation provides an inexhaustible supply of predictions for the card exchange. Each prediction emerges simply by substituting some value for X and solving the equation for Y', as described above. **Table 7.2** lists the predicted card returns for a number of different card investments. Verify that you can obtain a few of the Y' values shown in Table 7.2 from the regression equation.

Table 7.2
PREDICTED CARD RETURNS(Y') FOR DIFFERENT CARD INVESTMENTS (X)

X	Y′
0	6.40
4	9.60
8	12.80
10	14.40
12	16.00
20	22.40
30	30.40

Notice that, even when no cards are sent ($X = 0$), we predict a return of 6.40 cards because of the value of a. Also, notice that sending each additional card translates into an increment of only .80 in the predicted return because of the value of b. In other words, whenever b has a value less than 1.00, increments in the predicted return will lag—by an amount equal to the value of b, that is, .80 in the present case—behind increments in cards sent. If the value of b had been greater than 1.00, then increments in the predicted return would have exceeded increments in cards sent. (If the value of b had been negative, because of an underlying negative correlation, then sending additional cards would have triggered decrements, not increments, in the predicted return—and the tradition of sending holiday greeting cards probably would disappear.)

A Limitation

Emma might survey these predicted card returns before committing herself to a particular card investment. However, this strategy could backfire because there is no evidence of a simple *cause-effect* relationship between cards sent and cards received. The desired effect might be completely missing if, for instance, Emma expands her usual card distribution to include casual acquaintances and even strangers, as well as her friends and relatives.

Progress Check *7.2 Assume that an r of .30 describes the relationship between educational level (highest grade completed) and estimated number of hours spent reading each week. More specifically:

EDUCATIONAL LEVEL (X)	WEEKLY READING TIME (Y)
$\bar{X} = 13$	$\bar{Y} = 8$
$SS_x = 25$	$SS_y = 50$
$r = .30$	

(a) Determine the least squares equation for predicting weekly reading time from educational level.

(b) Faith's education level is 15. What is her predicted reading time?

(c) Keegan's educational level is 11. What is his predicted reading time?

Answers on page 497.

Graphs or Equations?

Encouraged by Figures 7.2 and 7.3, you might be tempted to generate predictions from graphs rather than equations. However, unless constructed skillfully, graphs yield less accurate predictions than do equations. In the long run, it is more accurate and easier to generate predictions from equations.

7.4 STANDARD ERROR OF ESTIMATE, $s_{y|x}$

Although we predicted that Emma's investment of 11 cards will yield a return of 15.20 cards, we would be surprised if she actually received 15 cards. It is more likely that because of the imperfect relationship between cards sent and cards received, Emma's return will be some number other than 15. Although designed to minimize predictive error, the least squares equation does not eliminate it. Therefore, our next task is to estimate the amount of error associated with our predictions. The smaller the estimated error is, the better the prognosis will be for our predictions.

Finding the Standard Error of Estimate

The estimate of error for new predictions reflects our failure to predict the number of cards received by the original five friends, as depicted by the discrepancies between solid and open dots in Figure 7.3. Known as the *standard error of estimate* and symbolized as $s_{y|x}$, this estimate of predictive error complies with the general format for any sample standard deviation, that is, the square root of some sum of squares term divided by its degrees of freedom. (See Formula 4.10 on page 91.) The formula for $s_{y|x}$ reads:

STANDARD ERROR OF ESTIMATE (DEFINITION FORMULA)

$$s_{y|x} = \sqrt{\frac{SS_{y|x}}{n-2}} = \sqrt{\frac{\Sigma(Y-Y')^2}{n-2}} \qquad (7.4)$$

where the sum of squares term in the numerator, $SS_{y|x}$, represents the sum of the squares for predictive errors, $Y - Y'$, and the degrees of freedom term in the denominator, $n - 2$, reflects the loss of two degrees of freedom because any straight line, including the regression line, can be made to coincide with two data points. The symbol $s_{y|x}$ is read as "s sub y given x."

Although we can estimate the overall predictive error by dealing directly with predictive errors, $Y - Y'$, it is more efficient to use the following computation formula:

STANDARD ERROR OF ESTIMATE (COMPUTATION FORMULA)

$$s_{y|x} = \sqrt{\frac{SS_y\left(1 - r^2\right)}{n - 2}} \tag{7.5}$$

where SS_y is the sum of the squares for Y scores (cards received by the five friends), that is,

$$SS_y = \sum\left(Y - \bar{Y}\right) = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n}$$

and r is the correlation coefficient (cards sent and received).

Key Property

Standard Error of Estimate ($s_{y|x}$)
A rough measure of the average amount of predictive error.

The **standard error of estimate** represents a special kind of standard deviation that reflects the magnitude of predictive error.

You might find it helpful to think of the standard error of estimate, $s_{y|x}$, as a rough measure of the average amount of predictive error—that is, as a rough measure of the average amount by which known Y values deviate from their predicted Y' values.*

The value of 3.10 for $s_{y|x}$, as calculated in **Table 7.3**, represents the standard deviation for the discrepancies between known and predicted card returns originally shown in Figure 7.3. In its role as an estimate of predictive error, the value of $s_{y|x}$ can be attached to any new prediction. Thus, a concise prediction statement may read: "The predicted card return for Emma equals 15.20 ± 3.10," in which the latter term serves as a rough estimate of the average amount of predictive error, that is, the average amount by which 15.20 will either overestimate or underestimate Emma's true card return.

* Strictly speaking, the standard error of estimate exceeds the average predictive error by 10 to 20 percent. Nevertheless, it is reasonable to describe the standard error in this fashion—as long as you remember that, as with the corresponding definition for the standard deviation in Chapter 4, an approximation is involved.

Table 7.3
CALCULATION OF THE STANDARD ERROR OF ESTIMATE, $s_{y|x}$

A. COMPUTATIONAL SEQUENCE

Assign values to SS_y and r 1 by referring to previous work with the least squares regression equation in Table 7.1.

Substitute numbers into the formula 2 and solve for $s_{y|x}$.

B. COMPUTATIONS

1 $SS_y = 80$

$r = .80$

2 $$s_{y|x} = \sqrt{\frac{SS_y\left(1-r^2\right)}{n-2}} = \sqrt{\frac{80\left(1-[.80]^2\right)}{5-2}} = \sqrt{\frac{80(.36)}{3}} = \sqrt{\frac{28.80}{3}} = \sqrt{9.60}$$

$$= 3.10$$

Importance of r

To appreciate the importance of the correlation coefficient in any predictive effort, let's substitute a few extreme values for r in the numerator of Formula 7.5 and note the resulting effect on the sum of squares for predictive errors, $SS_{y|x}$. Substituting a value of 1 for r, we obtain

$$SS_{y|x} = SS_y\,(1 - r^2) = SS_y[1 - (1)^2] = SS_y[1 - 1] = SS_y[0] = 0$$

As expected, when predictions are based on perfect relationships, the sum of squares for predictive errors equals zero, and there is no predictive error. At the other extreme, substituting a value of 0 for r in the numerator of Formula 7.5, we obtain

$$SS_{y|x} = SS_y(1 - r^2) = SS_y[1 - (0)^2] = SS_y[1 - 0] = SS_y[1] = SS_y$$

Again, as expected, when predictions are based on a nonexistent relationship, the sum of squares for predictive errors equals SS_y, the sum of squares of Y scores about $\overline{Y}$, and there is no reduction in predictive error. Clearly, the prognosis for a predictive effort is most favorable when predictions are based on strong relationships, as reflected by a sizable positive or negative value of r. The prognosis is most dismal—and a predictive effort should not even be attempted—when predictions must be based on a weak or nonexistent relationship, as reflected by a value of r near 0.

Progress Check *7.3

(a) Calculate the standard error of estimate for the data in Question 7.2 on page 161, assuming that the correlation of .30 is based on $n = 35$ pairs of observations.

(b) Supply a rough interpretation of the standard error of estimate.

Answers on page 497.

7.5 ASSUMPTIONS

Linearity

Use of the regression equation requires that the underlying relationship be linear. You need to worry about violating this assumption only when the scatterplot for the original correlation analysis reveals an obviously bent or curvilinear dot cluster, such as illustrated in Figure 6.4 on page 132. In the unlikely event that a dot cluster describes a pronounced curvilinear trend, consult more advanced statistics books for appropriate procedures.

Homoscedasticity

Use of the standard error of estimate, s_{yix}, assumes that except for chance, *the dots in the original scatterplot will be dispersed equally about all segments of the regression line.* You need to worry about violating this assumption, officially known by its tongue-twisting designation as the assumption of *homoscedasticity* (pronounced "ho-mo-skee-das-ti-ci-ty"), only when the scatterplot reveals a dramatically different type of dot cluster, such as that shown in **Figure 7.4**. At the very least, the standard error of estimate for the data in Figure 7.4 should be used cautiously, since its value overestimates the variability of dots about the lower half of the regression line and underestimates the variability of dots about the upper half of the regression line.

7.6 MULTIPLE REGRESSION EQUATIONS

Any serious predictive effort usually culminates in a more complex equation that contains not just one but several X, or *predictor variables*. For instance, a serious effort to predict college GPA might culminate in the following equation:

$$Y' = .410(X_1) + .005(X_2) + .001(X_3) + 1.03$$

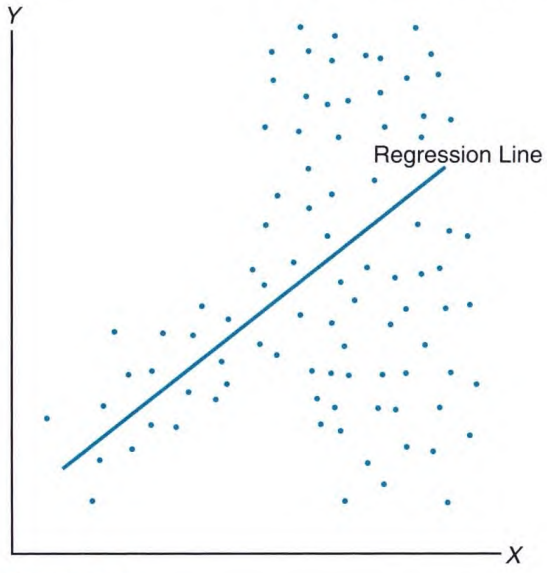


FIGURE 7.4
Violation of homoscedasticity assumption. (Dots lack equal variability about all line segments.)

Multiple Regression Equation
A least squares equation that contains more than one predictor or X variable.

where Y' represents predicted college GPA and X_1, X_2, and X_3 refer to high school GPA, IQ score, and SAT score, respectively. By capitalizing on the combined predictive power of several predictor variables, these **multiple regression equations** supply more accurate predictions for Y' (often referred to as the *criterion variable*) than could be obtained from a simple regression equation.

Common Features

Although more difficult to visualize, multiple regression equations possess many features in common with their simple counterparts. For instance, they still qualify as least squares equations, since they minimize the sum of the squared predictive errors. By the same token, they are accompanied by standard errors of estimate that roughly measure the average amounts of predictive error. Be assured, therefore, that this chapter will serve as a good point of departure if, sometime in the future, you must deal with multiple regression equations.

7.7 REGRESSION TOWARD THE MEAN

Regression Toward the Mean
A tendency for scores, particularly extreme scores, to shrink toward the mean.

Regression toward the mean *refers to a tendency for scores, particularly extreme scores, to shrink toward the mean.* This tendency often appears among subsets of observations whose values are extreme and at least partly due to chance. For example, because of regression toward the mean, we would expect that students who made the top five scores on the first statistics exam would not make the top five scores on the second statistics exam. Although all five students might score above the mean on the second exam, some of their scores would regress back toward the mean. Most likely, the top five scores on the first exam reflect two components. One relatively permanent component reflects the fact that these students are superior because of good study habits, a strong aptitude for quantitative reasoning, and so forth. The other relatively transitory component reflects the fact that, on the day of the exam, at least some of these students were very lucky because all sorts of little chance factors, such as restful sleep, a pleasant commute to campus, etc., worked in their favor. On the second test, even though the scores of these five students continue to reflect an above-average permanent component, some of their scores will suffer because of less good luck or even bad luck. The net effect is that the scores of at least some of the original five top students will drop below the top five scores—that is, regress *back* toward the mean—on the second exam. (When significant regression toward the mean occurs after a spectacular performance by, for example, a rookie athlete or a first-time author, the term *sophomore jinx* often is invoked.)

There is good news for those students who made the five lowest scores on the first exam. Although all five students might score below the mean on the second exam, some of their scores probably will regress *up* toward the mean. On the second exam, some of them will not be as unlucky. The net effect is that the scores of at least some of the original five poorest students will move above the bottom five scores—that is, regress up toward the mean—on the second exam.

Appears in Many Distributions

Regression toward the mean appears among subsets of extreme observations for a wide variety of distributions. For example, it appears for the subset of best (or worst) performing stocks on the New York Stock Exchange across any period, such as a week, month, or year. It also appears for the top (or bottom) major league baseball hitters during consecutive seasons. **Table 7.4** lists the top 10 hitters in the major leagues during 2007 and shows how they fared during 2008. Notice that, with the exception of Chipper Jones, each of their batting averages regressed downward, toward .264, the mean for all hitters during 2008. Incidentally, it is not true that, viewed as a group, all major league hitters are headed toward mediocrity. Hitters among the top 10 in 2007, who were not among the top 10 in 2008, were replaced by other mostly above-average hitters, who also were very lucky during 2008. Observed regression toward the mean occurs for individuals or subsets of individuals, not for entire groups.

The Regression Fallacy

Regression Fallacy
Occurs whenever regression toward the mean is interpreted as a real, rather than a chance, effect.

The **regression fallacy** *is committed whenever regression toward the mean is interpreted as a real, rather than a chance, effect.* A classic example of the regression fallacy occurred in an Israeli Air Force study of pilot training reported in a 1974 issue of *Science* by Amos Tversky and Daniel Kahnemann. Some trainees were praised after very good landings, while others were reprimanded after very bad landings. On their next landings, praised trainees did more poorly and reprimanded trainees did better. It was concluded, therefore, that praise hinders but a reprimand helps performance!

A valid conclusion considers regression toward the mean. It's reasonable to assume that, in addition to skill, chance plays a role in landings. Some trainees

Table 7.4
REGRESSION EFFECT: BATTING AVERAGES OF TOP 10 HITTERS IN MAJOR LEAGUE BASEBALL DURING 2007 AND HOW THEY FARED DURING 2008

	BATTING AVERAGES*		
TOP TEN HITTERS (2007)	2007	2008	REGRESSION EFFECT
1. Ordonez	.363	.317	Yes
2. Suzuki	.351	.310	Yes
3. Polanco	.341	.307	Yes
4. Holliday	.340	.321	Yes
5. Posada	.338	.268	Yes
6. Jones	.337	.364	No
7. H. Ramirez	.332	.301	Yes
8. Ortiz	.332	.264	Yes
9. Renteria	.332	.270	Yes
10. Utley	.332	.292	Yes

** Proportion of hits per official number of times at bat.*
Source: http://sports.espn.go.com/mlb/stats/batting.

who made very good landings were lucky, while some who made very bad landings were unlucky. Therefore, there would be a tendency, attributable to chance, that good landings would be followed by less good landings and poor landings would be followed by less poor landings—even if trainees had not been praised after very good landings or reprimanded after very bad landings.

Avoiding the Regression Fallacy

The regression fallacy can be avoided by splitting the subset of extreme observations into two groups. In the above example, one group of trainees would continue to be praised after very good landings and reprimanded after very poor landings. A second group of trainees would receive no feedback whatsoever after very good and very bad landings. In effect, the second group would serve as a control for regression toward the mean, since any shift toward the mean on their second landings would be due to chance. Most important, any observed difference between the two groups (that survives a statistical analysis described in Part 2) would be viewed as a real difference not attributable to the regression effect.

Watch out for the regression fallacy in educational research involving groups of underachievers. For example, a group of fourth graders, selected to attend a special program for underachieving readers, might show an improvement. Whether this improvement can be attributed to the special program or to a regression effect requires information from a control group of similarly underachieving fourth graders who did not attend the special program. It is crucial, therefore, that research with underachievers always includes a control group for regression toward the mean.

Progress Check *7.4 After a group of college students attended a stress-reduction clinic, declines were observed in the anxiety scores of those who, prior to attending the clinic, had scored high on a test for anxiety.

(a) Can this decline be attributed to the stress-reduction clinic? Explain your answer.

(b) What type of study, if any, would permit valid conclusions about the effect of the stress-reduction clinic?

Answers on page 498.

Summary

If a linear relationship exists between two variables, then one variable can be predicted from the other by using the least squares regression equation, as described in Formulas 7.1, 7.2, and 7.3.

The least squares equation minimizes the total of all squared predictive errors that would have occurred if the equation had been used to predict known Y scores from the original correlation analysis.

An estimate of predictive error can be obtained from Formula 7.5. Known as the *standard error of estimate,* this estimate is a special kind of standard deviation that roughly reflects the average amount of predictive error. The value of the standard error of estimate depends mainly on the size of the correlation coefficient. The larger the correlation coefficient, in either the positive or negative direction, the smaller the standard error of estimate and the more favorable the prognosis for predictions.

The regression equation assumes a linear relationship between variables, and the standard error of estimate assumes homoscedasticity—approximately equal dispersion of data points about all segments of the regression line.

Serious predictive efforts usually involve multiple regression equations composed of more than one predictor, or X, variable. These multiple regression equations share many common features with the simple regression equations discussed in this chapter.

Regression toward the mean refers to a tendency for scores, particularly extreme scores, to shrink toward the mean. The regression fallacy is committed whenever regression toward the mean is interpreted as a real, rather than a chance effect. To guard against the regression fallacy, control groups should be used to estimate the regression effect.

Important Terms

Least squares regression equation
Standard error of estimate ($s_{y|x}$)
Regression toward the mean
Multiple regression equation
Regression fallacy

Key Equations

PREDICTION EQUATION

$$Y' = bX + a$$

$$\text{where } b = \sqrt{\frac{SS_y}{SS_x}}r$$

$$\text{and } a = \bar{Y} - b\bar{X}$$

REVIEW QUESTIONS

7.5 Assume that an r of $-.80$ describes the strong negative relationship between years of heavy smoking (X) and life expectancy (Y). Assume, furthermore, that the distributions of heavy smoking and life expectancy each have the following means and sums of squares:

$$\bar{X} = 5 \qquad \bar{Y} = 60$$
$$SS_x = 35 \qquad SS_y = 70$$

(a) Determine the least squares regression equation for predicting life expectancy from years of heavy smoking.

(b) Determine the standard error of estimate, $s_{y|x}$, assuming that the correlation of $-.80$ was based on $n = 50$ pairs of observations.

(c) Supply a rough interpretation of $s_{y|x}$.

(d) Predict the life expectancy for Sara, who has smoked heavily for eight years.

(e) Predict the life expectancy for Katie, who has never smoked heavily.

7.6 Each of the following pairs represents the number of licensed drivers (X) and the number of cars (Y) for seven houses in my neighborhood:

DRIVERS (X)	CARS (Y)
5	4
5	3
2	2
2	2
3	2
1	1
2	2

(a) Construct a scatterplot to verify a lack of pronounced curvilinearity.

(b) Determine the least squares equation for these data. (Remember, you will first have to calculate r, SS_y, and SS_x.)

(c) Determine the standard error of estimate, $s_{y|x}$, given that $n = 7$.

(d) Predict the number of cars for each of two new families with two and five drivers.

7.7 At a large bank, length of service is the best single predictor of employees' salaries. Can we conclude, therefore, that there is a cause-effect relationship between length of service and salary?

***7.8** In studies dating back over 100 years, it's well established that regression toward the mean occurs between the heights of fathers and the heights of their *adult* sons. Indicate whether the following statements are true or false.

(a) Sons of tall fathers will tend to be shorter than their fathers.

(b) Sons of short fathers will tend to be taller than the mean for all sons.

(c) Every son of a tall father will be shorter than his father.

(d) Taken as a group, adult sons are shorter than their fathers.

(e) Fathers of tall sons will tend to be taller than their sons.

(f) Fathers of short sons will tend to be taller than their sons but shorter than the mean for all fathers.

Answers on page 498.

7.9 Someone suggests that it would be a good investment strategy to buy the five poorest-performing stocks on the New York Stock Exchange and capitalize on regression toward the mean. Comments?

7.10 In the original study of regression toward the mean, Sir Francis Galton noted a tendency for offspring of both tall and short parents to drift toward the mean height for offspring and referred to this tendency as "regression toward mediocrity." What is wrong with the conclusion that eventually all heights will be close to their mean?

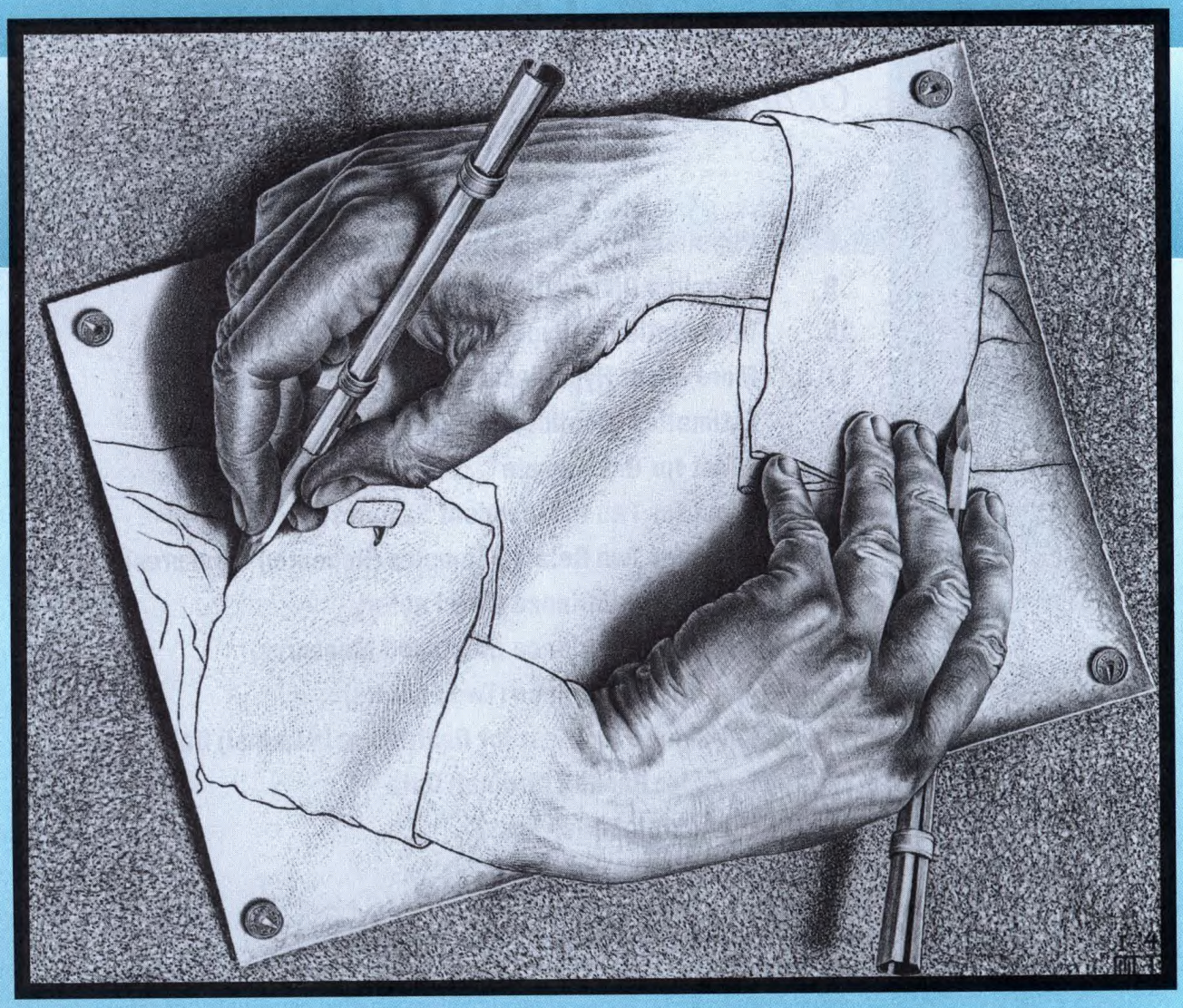
PART 2

Inferential Statistics:

Generalizing beyond Data

Preview

The remaining chapters deal with the problem of generalizing beyond sets of actual observations. The next two chapters develop essential concepts and tools for inferential statistics, while subsequent chapters introduce a series of statistical tests or procedures, all of which permit us to generalize beyond an observed result, whether from a survey or an experiment, by considering the effects of chance.



CHAPTER 8

Populations, Samples, and Probability

POPULATIONS AND SAMPLES

PROBABILITY

Preview

In everyday life, we regularly generalize from limited sets of observations. One sip indicates that the batch of soup is too salty; dipping a toe in the swimming pool reassures us before taking the first plunge; a test drive triggers suspicions that the used car is not what it was advertised to be; and a casual encounter with a stranger stimulates fantasies about a deeper relationship. Valid generalizations in inferential statistics require either random sampling in the case of surveys or random assignment in the case of experiments. Introduced in this chapter, tables of random numbers can be used as aids to random sampling or random assignment.

Conclusions that we'll encounter in inferential statistics, such as "95 percent confident" or "significant at the. 05 level," are statements based on probabilities. We'll define probability for a simple event and then discuss two rules for finding probabilities of more complex outcomes, including (in Review Question 8.14 on page 190) the probability of the catastrophic failure of the Challenger shuttle in 1986, which took the lives of seven astronauts.

POPULATIONS AND SAMPLES

Generalizations can backfire if a sample misrepresents the population. Faced with the possibility of erroneous generalizations, you might prefer to bypass the uncertainties of inferential statistics by surveying an entire population. This is often done if the size of the population is small. For instance, you calculate your GPA from all of your course grades, not just from a sample. If the size of the population is large, however, complete surveys are often prohibitively expensive and sometimes impossible. Under these circumstances, you might have to use samples and risk the possibility of erroneous generalizations. For instance, you might have to use a sample to estimate the mean annual income for parents of all students at a large university.

8.1 POPULATIONS

Reminder:
Population *refers to any complete set of observations (or potential observations).*

Any complete set of observations (or potential observations) may be characterized as a **population.** Accurate descriptions of populations specify the nature of the observations to be taken. For example, a population might be described as "attitudes toward abortion of currently enrolled students at Bucknell University" or as "SAT critical reading scores of currently enrolled students at Rutgers University."

Real Populations

Pollsters, such as the Gallup Organization, deal with real populations. A *real* population is one in which all potential observations are accessible at the time of sampling. Examples of real populations include the two described in the previous paragraph, as well as the ages of all visitors to Disneyland on a given day, the ethnic backgrounds of all current employees of the U.S. Postal Department, and presidential preferences of all currently registered voters in the United States. Incidentally, federal law requires that a complete survey be taken every ten years of the real population of all U.S. households—at considerable expense, involving thousands of data collectors—as a means of revising election districts for the House of Representatives. (An estimated undercount of millions of people in both the 1990 and 2000 censuses has revived a suggestion, long endorsed by statisticians, that the entire U.S. population could be estimated more accurately if a highly trained group of data collectors focused only on a random sample of households.)

WWW

INTERNET SITE
Go to the Web site for this book (http://www.wiley.com\college\witte). Click on the *Student Companion Site,* then *Internet Sites,* and finally ***U.S. Census Bureau*** to view its Web site, including links to its many reports and to population clocks that show current population estimates for the United States and the world.

Hypothetical Populations

Insofar as research workers concern themselves with populations, they often invoke the notion of a hypothetical population. A *hypothetical* population is one in which all potential observations are not accessible at the time of sampling. In most experiments, subjects are selected from very small, uninspiring real populations: the lab rats housed in the local animal colony or student volunteers from general psychology classes. Experimental subjects often are viewed, nevertheless, as a sample from a much larger hypothetical population, loosely described as "the scores of all similar animal subjects (or student volunteers) who could conceivably undergo the present experiment."

According to the rules of inferential statistics, generalizations should be made only to real populations that, in fact, have been sampled. Generalizations to hypothetical populations should be viewed, therefore, as provisional conclusions based on the wisdom of the researcher rather than on any logical or statistical necessity. In effect, it's an open question—answered only by additional experimentation—whether or not a given experimental finding merits the generality assigned to it by the researcher.

8.2 SAMPLES

Reminder:
Sample *refers to any subset of observations from a population.*

Any subset of observations from a population may be characterized as a **sample.** In typical applications of inferential statistics, the sample size is small relative to the population size. Less than 1 percent of all U.S. households are included in the Bureau of Labor Statistics' monthly survey to estimate the rate of unemployment. Although, at most, only about 4,000 voters have been sampled in recent presidential election polls by Gallup, predictions have been amazingly accurate since 1952—missing the actual percentage of votes for the winning candidate by an average of only 2%, according to the subscription-only service maintained by the Gallup Organization at http://www.gallup.com/poll/topics/ptaccuracy.asp.

Optimal Sample Size

There is no simple rule of thumb for determining the best or optimal sample size for any particular situation. Often sample sizes are in the hundreds or even the thousands for surveys, but they are less than 100 for most experiments. Optimal sample size depends on the answers to a number of questions, including "What is the estimated variability among observations?" and "What is an acceptable amount of error in our conclusion?" Once these types of questions have been answered, with the aid of guidelines such as those discussed in Section 11.11, specific procedures can be followed to determine the optimal sample size for any situation.

Progress Check *8.1 For each of the following pairs, indicate with a Yes or No whether the relationship between the first and second expressions could describe that between a sample and its population, respectively.

(a) students in the last row; students in class

(b) citizens of Wyoming; citizens of New York