CHAPTER 6

Descriptive Statistics

Descriptive statistics simplify our lives by organizing and summarizing data.

INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Describe the nature and uses of descriptive statistics.
2. Identify the characteristics, uses, and limitations of four types of measurement scales—nominal, ordinal, interval, and ratio.
3. Organize research data into frequency distributions, present them as frequency polygons and histograms, and interpret polygons and histograms.
4. Distinguish between the measures of central tendency and the situations in which each should be used. Calculate and interpret the mean, the median, and the mode.
5. Describe appropriate applications of measures of variability and compute variance, standard deviation, and range.
6. Calculate and explain why *z* scores have universal meaning and how this is useful in interpreting the position of a single observation in a distribution.
7. Explain why *z* scores are often transformed into other standard scores.
8. Convert a *z* score to a stanine score and use this to give a verbal description of the score's meaning. Explain why stanine scores are easy to interpret.
9. Transform raw scores into standard scores.
10. Explain advantages and disadvantages of percentile ranks. Calculate percentile rank for a given score.
11. Identify the characteristics of the normal curve. Explain why it is useful in descriptive research.
12. Use the normal curve table to estimate the percentile rank of a given *z* score or estimate the *z* score of a given percentile rank.
13. Identify appropriate applications of Pearson *r* correlation for describing the relationship between variables. Explain why it shows both the direction and the strength of the relationship.
14. Describe the meaning of the coefficient of determination and its application in interpreting the coefficient of correlation.
15. Identify the components of effect size and the factors that increase and decrease effect size.

16 Explain how effect size assesses the strength of relationships between variables.

17 Calculate effect size for a difference between means. Explain why the Pearson *r* is a form of effect size.

18 Perform a meta-analysis and explain the meaning of a meta-analysis outcome.

Statistical procedures are basically methods of handling quantitative information. These procedures have two principal advantages. First, they enable researchers to organize, summarize, and describe observations. Techniques used for these purposes are called **descriptive statistics**. Second, they help determine how reliably researchers can infer that phenomena observed in a limited group—a *sample*—are likely to occur in the unobserved larger population of concern from which the sample was drawn; in other words, how accurately researchers can employ inductive reasoning to infer that what they observe in the part will be observed in the whole. Techniques used for such purposes are called **inferential statistics**.

Knowledge of some basic statistical procedures is essential for researchers proposing to carry out quantitative research. They need statistics to analyze and interpret their data and communicate their findings to others. Researchers also need an understanding of statistics in order to read and evaluate published research in their fields.

SCALES OF MEASUREMENT

A fundamental step in conducting quantitative research is measurement—the process through which observations are translated into numbers. S. S. Stevens (1951) is well remembered for his definition: "In its broadest sense, measurement is the assignment of numerals to objects or events according to rules" (p. 1). Quantitative researchers first identify the variables they want to study; then they use rules to determine how to express these variables numerically. The variable *religious preference* may be measured according to the numbers indicated by students who are asked to select among (1) Catholic, (2) Jewish, (3) Protestant, or (4) Muslim. The variable *weight* may be measured as the numbers observed when subjects step on a scale. The variable self-concept may be operationally defined as scores on the Multidimensional Self-Concept Scale. The nature of the measurement process that produces the numbers determines the interpretation that can be made from them and the statistical procedures that can be used meaningfully with them. The most widely quoted taxonomy of measurement scales is Stevens' classification of measurement as nominal, ordinal, interval, and ratio.

NOMINAL SCALE

The most basic scale of measurement is the **nominal scale**. Nominal measurement involves placing objects or individuals into mutually exclusive categories. Numbers are arbitrarily assigned to the categories for identification

purposes only. The numbers do not indicate any value or amount; thus, one category does not represent "more or less" of a characteristic. School District 231 is not more or less of anything than School District 103. Examples of a nominal scale are using a "0" to represent males and a "1" to represent females, or the religious preference described previously.

Because the numbers in a nominal scale do not represent quantity, they cannot be arithmetically manipulated through addition, subtraction, multiplication, or division. You can only count the number of observations in each category or express the numbers in categories as a percentage of the total number of observations.

ORDINAL SCALE

An **ordinal scale** ranks objects or individuals according to how much of an attribute they possess. Thus, the numbers in an ordinal scale indicate only the order of the categories. Neither the difference between the numbers nor their ratio has meaning. For example, in an untimed sprint we know who came in first, second, and third, but we do not know how much faster one runner was than another. A ranking of students in a music contest is an ordinal scale. We would know who was awarded first place, second place, and so on, but we would not know the extent of difference between them.

The essential requirement for measurement at this level is that the relationship must be such that if object X is greater than object Y and object Y is greater than object Z, then object X is greater than object Z and is written thus: If $(X > Y)$ and $(Y > Z)$, then $(X > Z)$. When appropriate, other wording may be substituted for "greater than," such as "stronger than," "precedes," and "has more of."

The lack of equal intervals in ordinal scales limits the statistical procedures available for analyzing ordinal data. We can use statistics that indicate the points below which certain percentages of the cases fall in a distribution of scores.

INTERVAL SCALE

An **interval scale** not only places objects or events in order but also is marked in equal intervals. Equal differences between the units of measurement represent equal differences in the attribute being measured. Fahrenheit and Celsius thermometers are examples of interval scales. We can say that the difference between 60° and 70° is the same as the distance between 30° and 40°, but we cannot say that 60° is twice as warm as 30° because there is no true zero on an interval scale. Zero on an interval scale is an arbitrary point and does not indicate an absence of the variable being measured. Zero on the Celsius scale is arbitrarily set at the temperature at which water freezes at sea level.

Numbers on an interval scale may be manipulated by addition and subtraction, but because the zero is arbitrary, multiplication and division of the numbers are not feasible. Thus, ratios between the numbers on an interval scale are meaningless. We may report differences between positions on an interval scale, or we may add the numbers to report an average.

It is important to note that in most academic measures, the intervals are equal in terms of the measuring instrument, but not necessarily in terms of the performance being measured. To illustrate, consider a spelling test with the following words: *cat, dish, ball, loquacious, schizophrenia*, and *pneumonia*. Here, the distance between one correct and three correct is the same as the distance between three correct and five correct. However, when considered in terms of spelling performance, the difference between three and five correct suggests a greater difference in ability than does the difference between one and three correct. Unless you can say that the distance between three and five on the spelling test represents the same amount of spelling performance as does the distance between one and three, then these scores indicate nothing more than the rank order of the students.

However, through careful construction it is possible to produce an instrument where the intervals observed between scores on the test give a reasonable approximation of ability intervals. The better intelligence tests are an example of this. The difference in ability between an IQ of 90 and an IQ of 95 may not be precisely the same as the difference between an IQ of 105 and an IQ of 110, but you will not be greatly misled if you assume that the two differences are approximately equal.

It has become common practice to treat many educational variables, such as classroom tests and grades (A = 4, B = 3, and so on) as if they were interval data, even when that assumption is not well justified. It would be difficult to maintain that the difference between F = 0 and D = 1 represents the same difference in academic achievement as the difference between C = 2 and B = 3, or to justify treating scores on our spelling test example as interval data. Be cautious when interpreting statistics derived from such data. The statistics imply interval-level information, but it is actually often somewhere between ordinal and interval.

However, in schools and universities, grade point average is almost always calculated as interval data. Scores on teacher-made tests are usually quasi-interval, somewhere between interval and ordinal data. Although scores on teacher-made tests are rarely as far from interval data as our spelling test example, they almost always vary from true interval data to some extent.

RATIO SCALE

A ratio scale, the highest level of measurement scale, has a true zero point as well as equal intervals. Ratios can be reported between any two given values on the scale. A yardstick used to measure length in units of inches or feet is a ratio scale because the origin on the scale is an absolute zero corresponding to no length at all. Thus, it is possible to state that a stick 6 feet long is twice as long as a stick 3 feet long. Other examples of ratio scales are weight, money, and distance. All types of statistical procedures are appropriate with a ratio scale. Only a few variables in education—largely confined to motor performance and other physiological measures—are ratio in nature. A shot put score of 16 yards is twice as far as a shot put score of 8 yards, but you cannot say that a person who scores 40 on a math test is twice as good at

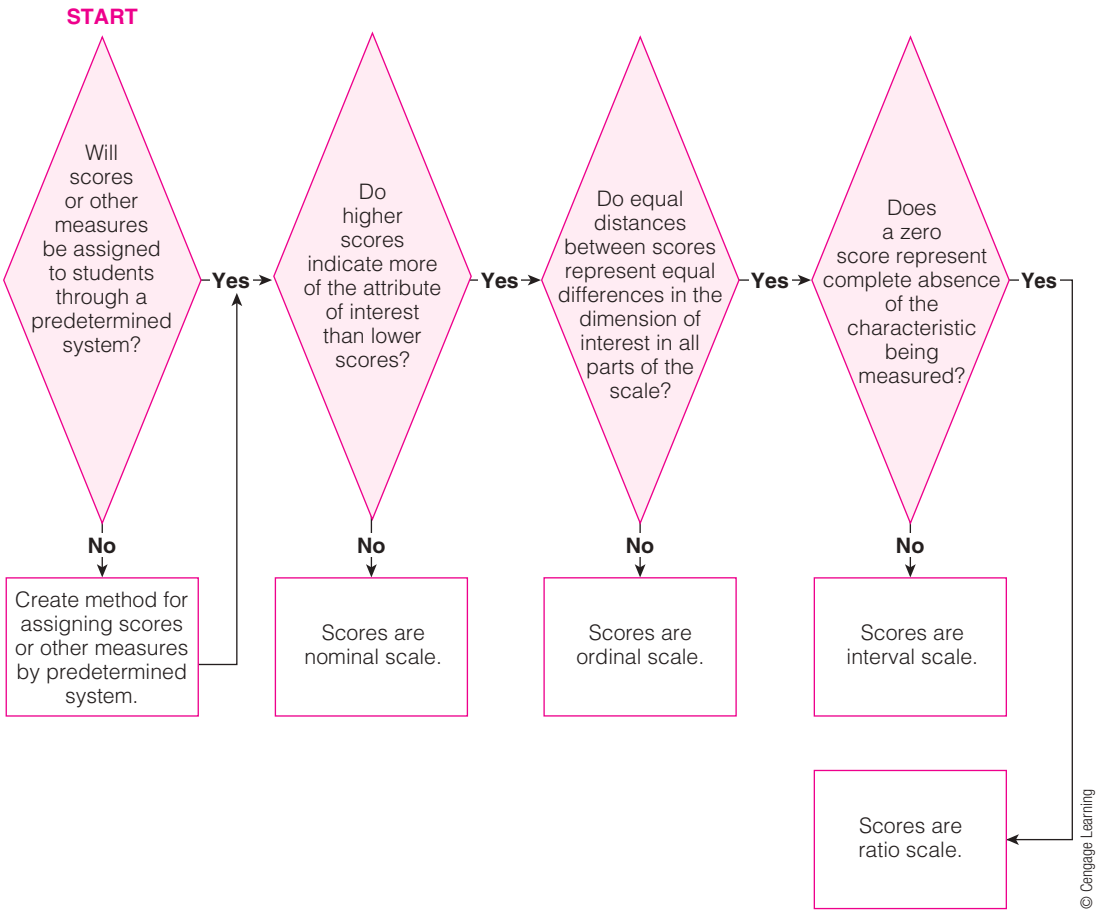


Figure 6.1 Determining Scales of Measurement

math as a person who scores 20, because math test scores are not ratio data. Figure 6.1 shows the decisions made to determine the scale of measurement of an observation.

THINK ABOUT IT 6.1

You are buying a used car. You consider (a) its make (Ford, Toyota, etc.), (b) the miles on the odometer, (c) the year it was made, and (d) its rating in *Consumer Reports*.

1. Which of the above is nominal data?
2. Which of the above is ordinal data?
3. Which of the above is interval data?
4. Which of the above is ratio data?

Answers

1. a; 2. d; 3. c (the year 2012 is not twice the year 1006); 4. b

ORGANIZING RESEARCH DATA

Before applying statistical procedures, researchers must organize large amounts of data into a manageable form. The most familiar ways of organizing data are (1) arranging the measures into frequency distributions and (2) presenting them in graphic form.

FREQUENCY DISTRIBUTIONS

A systematic arrangement of individual measures from highest to lowest is called a **frequency distribution**. The first step in preparing a frequency distribution is to list the scores in a column from highest at top to lowest at bottom. Include all possible intermediate scores even if no one scored them; otherwise, the distribution will appear more compact than it really is. Several identical scores often occur in a distribution. Instead of listing these scores separately, it saves time to add a second column recording the frequency of each measure. Table 6.1 shows the test scores of a group of 105 students in an Ed 101 lecture class. Part A of the table lists the scores in an unorganized form. Part B arranges them in a frequency distribution, with the *f* column showing how many made each score. Now it is possible to examine the general "shape" of the distribution. , You can determine the spread of scores so organized—whether they are distributed evenly or tend to cluster, and where clusters occur in the distribution. For example, looking over the frequency distribution of the scores presented in Part B of Table 6.1, it is easy to see that they range from 21 to 36, that 29 is the most frequent score, and that scores tend to cluster more near the top of the distribution than the bottom. None of this would be apparent had the scores not been organized. Organizing data into frequency distributions also facilitates the computation of various useful statistics.

THINK ABOUT IT 6.2

Here are the scores that Mr. Li's 18 physics class students made on their first exam:
Ali, 21; Ann, 20; Ben, 23; Cal, 20; Dan, 20; Ed, 21; Ima, 22; Jan,19; Kay, 16; Lee, 20; Mel, 18; Mia, 23; Ned, 21; Ona, 21; Sam, 22; Sue, 19; Ted,16; Van, 18. Create a frequency distribution of these scores. (For the answer, see the first two columns in Table 6.2.)

GRAPHIC PRESENTATIONS

It is often helpful and convenient to present research data in graphic form. Among various types of graphs, the most widely used are the **histogram** and the **frequency polygon**. The initial steps in constructing the histogram and the frequency polygon are identical:

1. Lay out the score points on a horizontal dimension (abscissa) from the lowest value on the left to the highest on the right. Leave enough space for an additional score at both ends of the distribution.

Table 6.1 The Test Scores of 105 Students on Ed 101 Test

Part A. Unorganized Scores

33	29	30	30	33	29	33	32	28	24	34	31	27	29	23
25	29	24	27	26	33	33	26	30	28	26	29	32	32	31
28	34	30	31	33	21	29	31	30	32	36	30	31	27	29
26	29	33	32	29	28	28	30	28	27	30	31	34	33	22
30	29	27	29	24	30	21	31	31	33	28	21	31	29	31
31	33	22	29	31	32	32	31	28	29	30	22	33	30	30
32	33	31	33	28	29	27	33	27	21	30	29	28	27	33

Part B. Frequency Distribution

Scores (X)	Tallies	Frequencies (f)	fX	cf
36	/	1	36	105
35		0		104
34	///	3	102	104
33	//// //// ////	15	405	101
32	//// ///	8	256	80
31	//// //// ////	14	434	78
30	//// //// ////	14	420	64
29	//// //// //// /	16	464	50
28	//// ////	10	280	34
27	//// ///	8	216	29
26	////	4	104	16
25	/	1	25	12
24	///	3	72	11
23	/	1	23	8
22	///	3	66	7
21	////	4	84	4
		$N = 105$		

We discuss the fX and cf columns later.

Table 6.2 Mr. Li's Physics Class Exam Scores

(1) X	(2) f	(3) fX	(4) cf
23	2	46	18
22	2	44	16
21	4	84	14
20	4	80	10
19	2	38	6
18	2	36	4
17	0	0	2
16	2	32	2

2. Lay out the frequencies of the scores (or intervals) on the vertical dimension (ordinate).
3. Place a dot above the center of each score at the level of the frequency of that score.

From this point you can construct either a histogram or a polygon. In constructing a histogram, draw through each dot a horizontal line equal to the width representing a score, as shown in Figure 6.2. A score of 26 is thought of as ranging from 25.5 to 26.5, a score of 27 is thought of as ranging from 26.5 to 27.5, and so forth.

To construct a polygon, connect the adjacent dots, and connect the two ends of the resulting figure to the base (zero line) at the points representing 1 less than the lowest score and 1 more than the highest score, as shown in Figure 6.3. Histograms are preferred when you want to indicate the discrete nature of the data, such as when a nominal scale has been used. Polygons are preferred to indicate data of a continuous nature. For additional reading pertaining to the use and merits of other graphing techniques, such as boxplots and error bars, for presenting data or rendering explanations from results, see Walker (2005b).

THINK ABOUT IT 6.3

Construct a histogram and a polygon of the scores of Mr. Li's first physics exam.

Answer
See Figure 6.4.

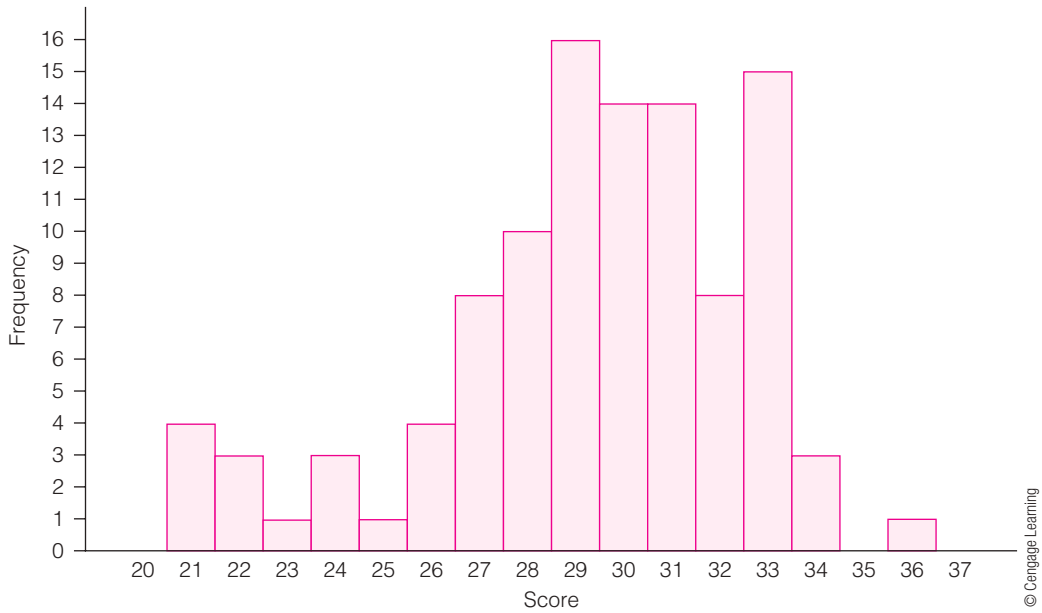


Figure 6.2 Histogram of 105 Test Scores from Table 6.1

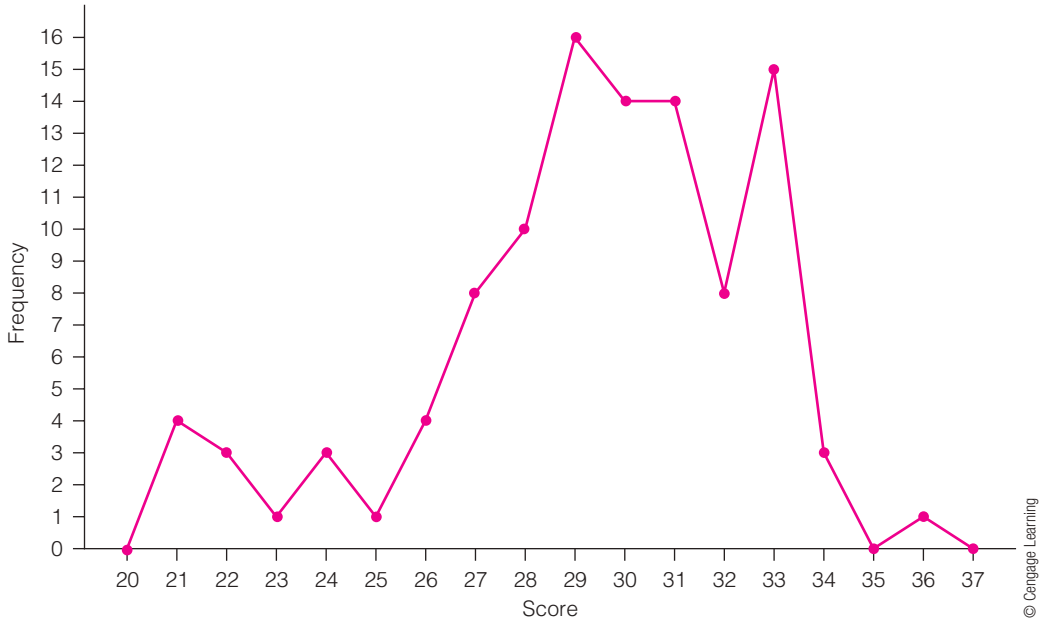


Figure 6.3 Frequency Polygon of 105 Test Scores from Table 6.1

MEASURES OF CENTRAL TENDENCY

A convenient way of summarizing data is to find a single index that can represent a whole set of measures. Finding a single score that can give an indication of the performance of a group of 300 individuals on an aptitude test would be useful for comparative purposes. In statistics, three indexes called **measures of central tendency** are available for such use. To most, the term *average* means the sum of the scores divided by the number of scores. The average can be this measure, known as the mean, or two other popular measures of central tendency known as the *mode* and the *median*. Each of these three can serve as an index to represent a group as a whole.

THE MEAN

The most widely used measure of central tendency is the **mean** or arithmetic average. It is the sum of all the scores in a distribution divided by the number of cases. In terms of a formula, it is

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{N} \tag{6.1}$$

which is usually written as

$$\overline{X} = \frac{\Sigma X}{N} \tag{6.2}$$

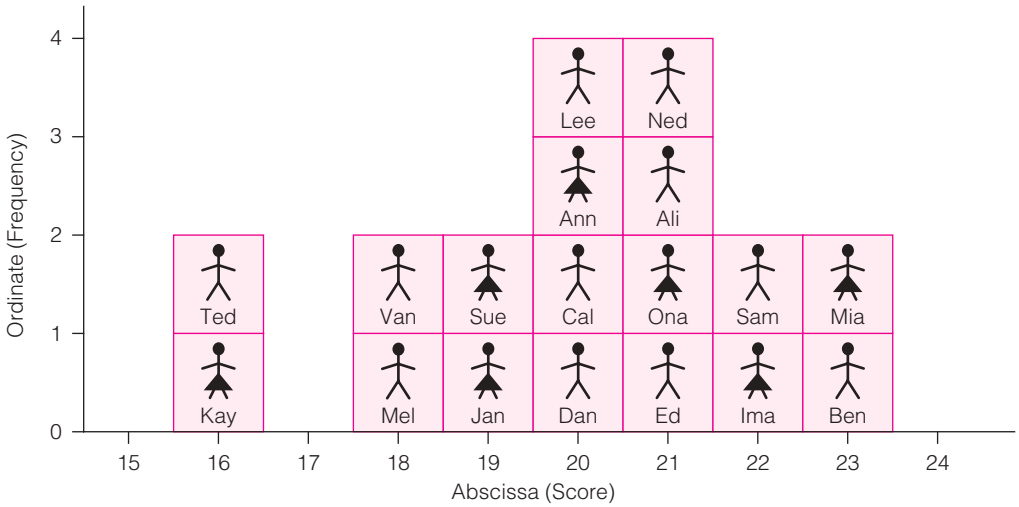
where

$\overline{X}$ = mean
Σ = sum of
X = raw score
N = number of cases

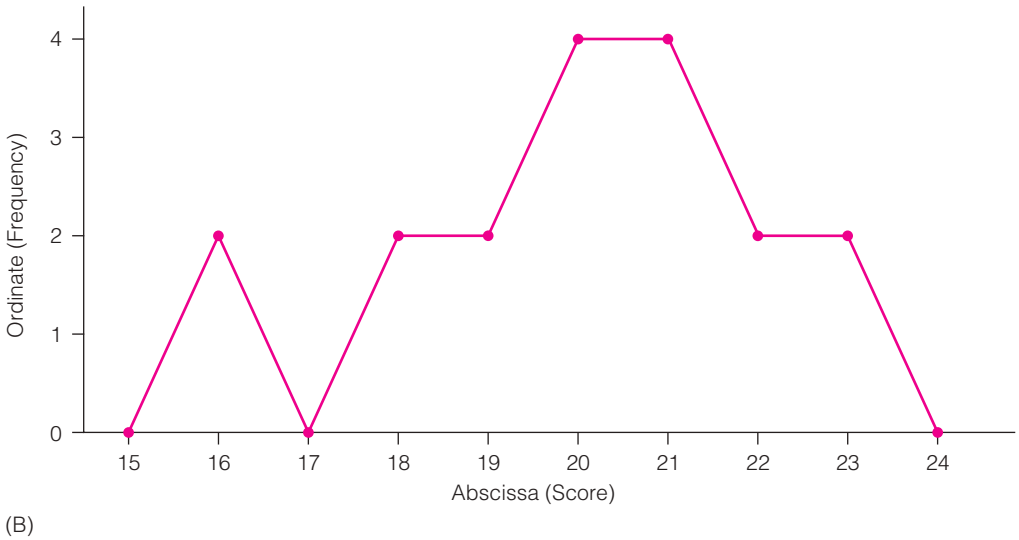
Applying Formula 6.2 to the following IQ scores, we find that the mean is 111:

IQ scores: 112 121 115 101 119 109 100

$$\overline{X} = \frac{112 + 121 + 115 + 101 + 119 + 109 + 100}{7} = \frac{777}{7} = 111$$



(A)



(B)

Figure 6.4 (A) Histogram of Mr. Li's Physics Exam and (B) Polygon of Mr. Li's Physics Exam

Note that in this computation, the scores were not arranged in any particular order. Ordering is unnecessary for calculation of the mean.

Some think of formulas as intimidating incantations. Actually, they are timesavers. It is much easier to write $\overline{X} = \Sigma X/N$ than to write "Add all the scores in a distribution and divide by the number of cases to calculate the mean."

Although it is not necessary to put the scores in order to calculate the mean, with larger sets of numbers it is usually convenient to start with a frequency distribution and multiply each score by its frequency. This is shown in column 3 (fX) in Table 6.2, Mr. Li's physics class exam scores. Adding the numbers in this column will give us the sum of the scores.

$$\Sigma X = 360$$

The mean of the physics exam scores is

$$\overline{X} = \frac{\Sigma X}{N} = \frac{360}{18} = 20$$

THE MEDIAN

The **median** is defined as that point in a distribution of measures below which 50 percent of the cases lie (which means that the other 50 percent will lie above this point). Consider the following distribution of scores, where the median is 18:

14 15 16 17 18 19 20 21 22

In the following 10 scores we seek the point below which 5 scores fall:

14 16 16 17 18 19 20 20 21 22

The point below which 5 scores, or 50 percent of the cases, fall is halfway between 18 and 19. Thus, the median of this distribution is 18.5.

Consider the following scores:

18 20 22 25 25 30

Any point from 22.5 to 24.5 fits the definition of the median. By convention in such cases the median is defined as halfway between these lowest and highest points, in this case 22.5 + 24.5/2 = 23.5.

To find the median of Mr. Li's physics exam scores, we need to find the point below which 18/2 = 9 scores lie. We first create a cumulative frequency column (cf, column 4 in Table 6.2). The cumulative frequency for each interval is the number of scores in that interval plus the total number of scores below it. Since the interval between 15.5 and 16.5 has no scores below it, its cf is equal to its f, which is 2. Since there were no scores of 17, the cf for 17 is still 2. Thus, adding the two scores of 18 yields a cumulative frequency of 4. Continuing up the frequency column, we get cfs of 10, 14, 16, and, finally, 18, which is equal to the number of students.

The point separating the bottom nine scores from the top nine scores, the median, is somewhere in the interval 19.5 to 20.5. Most statistics texts say to partition this interval to locate the median. The cf column tells us that we have six scores below 19.5. We need to add three scores to give us half the

scores (9). Since there are four scores of 20, we go three-fourths of the way from 19.5 to 20.5 to report a median of 20.25. Note that many computer programs, including the Statistical Package for the Social Sciences (SPSS) and the Statistical Analysis System (SAS), simply report the midpoint of the interval—in this case 20—as the median.

Notice that the median does not take into account the size of individual scores. In order to find it, you arrange your data in rank order and find the point that divides the distribution into two equal halves. The median is an ordinal statistic because it is based on rank. You can compute a median from interval or ratio data, but in such cases the interval characteristic of the data is not being used. One circumstance in which the median may be the preferred measure of central tendency arises when there are some extreme scores in the distribution. In this case, the use of a measure of central tendency that takes into account the size of each score results in either overestimation or underestimation of the typical score. The median, because of its insensitivity to extreme scores, is the appropriate index to be applied when you want to find the typical score. For illustration, consider the following distribution:

49 50 51 53 54 55 56 60 89

The score of 54, which is the median of this distribution, is the most typical score. The mean, which takes into account the individual values of the scores 60 and 89, will certainly result in an overestimation of the typical score.

THE MODE

The **mode** is the value in a distribution that occurs most frequently. It is the simplest to find of the three measures of central tendency because it is determined by inspection rather than by computation. Given the distribution of scores

14 16 16 17 18 19 19 19 21 22

you can readily see that the mode of this distribution is 19 because it is the most frequent score. In a histogram or polygon, the mode is the score value of the highest point (the greatest frequency), as you can see in Figures 6.2 and 6.3, where the mode is 29. Sometimes there is more than one mode in a distribution. For example, if the scores had been

14 16 16 16 18 19 19 19 21 22

you would have two modes: 16 and 19. This kind of distribution with two modes is called *bimodal*. Distributions with three or more modes are called *trimodal* or *multimodal*, respectively.

The mode is the least useful indicator of central value in a distribution for two reasons. First, it is unstable. For example, two random samples drawn from the same population may have quite different modes. Second, a distribution may have more than one mode. In published research, the mode is seldom reported as an indicator of central tendency. Its use is largely limited to inspectional purposes. A mode may be reported for any of the scales of measurement, but it is the only measure of central tendency that may legitimately be used with nominal scales.

RESEARCH IN THE PUBLIC EYE

We have all heard the statistics that highly educated people earn more income than do less educated people. According to a September 2011 article by Beckie Supiano, the median lifetime earnings for those holding a bachelor's degree is higher than those with an associate's degree. But the article points out several other details of interest. Of those with an associate's degree, 28.2 percent actually earn more than those with a bachelor's degree as a result of differences in wages by type of occupation. The article points out women at every level earn less than men. Looking at the median, men with no degree or some college earn as much in a lifetime as women with a bachelor's degree, and the median income of women with a Ph.D. or professional degree is equivalent to that of men with bachelor's degrees.

Discuss the implications for researchers of basing general findings only on median data.

COMPARISON OF THE THREE INDEXES OF CENTRAL TENDENCY

Because the mean is an interval or ratio statistic, it is generally a more precise measure than the median (an ordinal statistic) or the mode (a nominal statistic). It takes into account the value of *every* score. It is also the most stable of the three measures of central tendency in that if a number of samples are randomly drawn from a parent population, the means of these samples will vary less from one another than will their medians and their modes. For these reasons, the mean is more frequently used in research than the other two measures.

The mean is the best indicator of the combined performance of an entire group. However, the median is the best indicator of *typical* performance. Consider, for example, a school board whose members have the following annual incomes: $140,000, $60,000, $50,000, $40,000, and $40,000. The mean, $66,000, is the sum of their incomes divided by the number of members, but it is higher than all but one of the board members' incomes. The median, $50,000, gives a better picture of the typical income in the group.

SHAPES OF DISTRIBUTIONS

Frequency distributions can have a variety of shapes. A distribution is symmetrical when the two halves mirror each other. In a **symmetrical distribution**, the values of the mean and the median coincide. If such a distribution has a single mode, rather than two or more modes, the three indexes of central tendency will coincide, as shown in Figure 6.5.

If a distribution is not symmetrical, it is described as **skewed**, pulled out to one end or the other by the presence of extreme scores. In skewed distributions, the values of the measures of central tendency differ. In such distributions, the value of the mean, because it is influenced by the size of extreme scores, is pulled toward the end of the distribution in which the extreme scores lie, as shown in Figures 6.6 and 6.7. The effect of extreme values is less on the median because this index is influenced not by the size of scores, but by their position. Extreme values have no impact on the mode because this index has no relation with either of the ends of the distribution. Skews are labeled according to where the extreme scores lie. A way to remember this is "The tail names the beast." Figure 6.6 shows a **negatively skewed distribution**, whereas Figure 6.7 shows a **positively skewed distribution**.

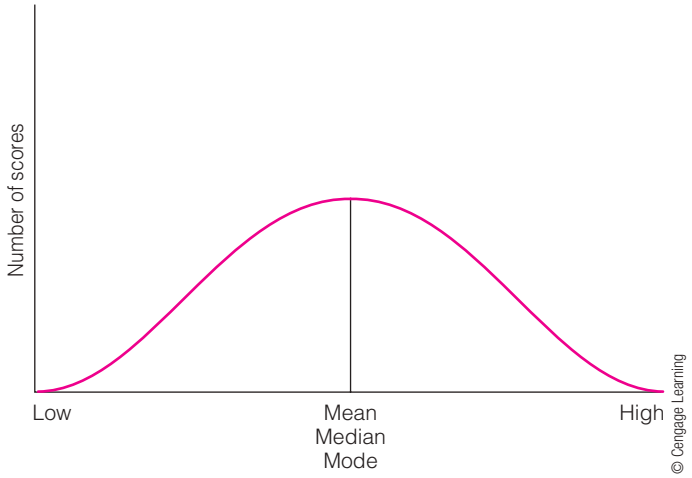


Figure 6.5 Symmetrical Distribution

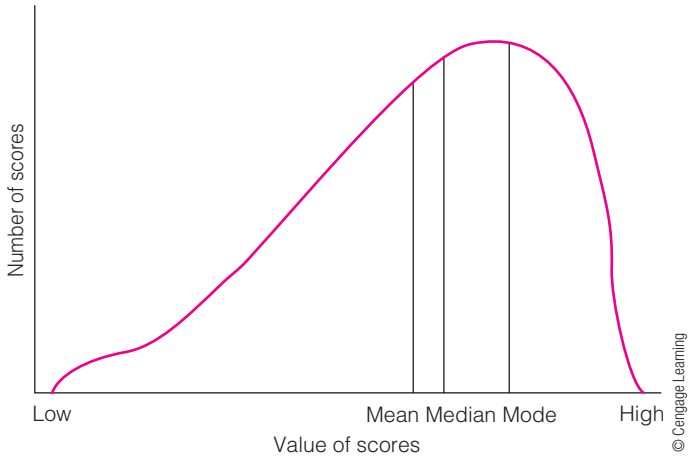


Figure 6.6 Negatively Skewed Distribution

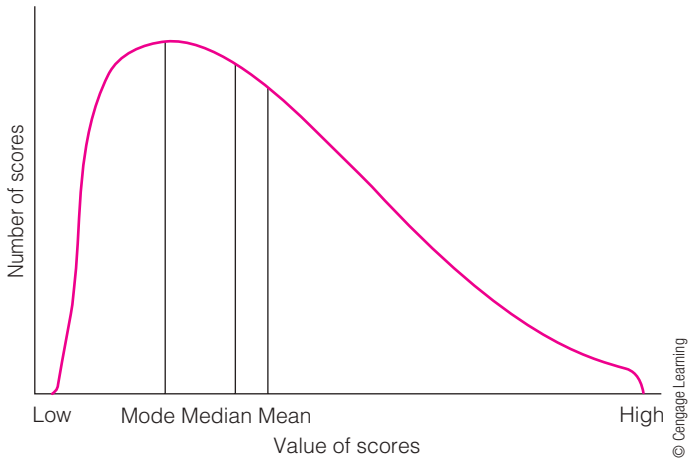


Figure 6.7 Positively Skewed Distribution

MEASURES OF VARIABILITY

Although indexes of central tendency help researchers describe data in terms of average value or typical measure, they do not give the total picture of a distribution. The mean values of two distributions may be identical, whereas the degree of dispersion, spread, or **variability**, of their scores might differ. In one distribution, the scores might cluster around the central value; in the other, they might be scattered. For illustration, consider the following distributions of scores:

(a) 24, 24, 25, 25, 25, 26, 26 $\quad \overline{X} = 175/7 = 25$
(b) 16, 19, 22, 25, 28, 30, 35 $\quad \overline{X} = 175/7 = 25$

The value of the mean in both these distributions is 25, but the degree of scattering of the scores differs considerably. The scores in distribution (a) are obviously much more homogeneous than those in distribution (b). There is clearly a need for indexes that can describe distributions in terms of *variation, spread, dispersion, heterogeneity*, or *scatter* of scores. Three indexes are commonly used for this purpose: range, variance, and standard deviation.

RANGE

The simplest of all indexes of variability is the **range**. It is the difference between the upper real limit of the highest score and the lower real limit of the lowest score. In statistics, any score is thought of as representing an interval width from halfway between that score and the next lowest score (lower real limit) up to halfway between that score and the next highest score (upper real limit). For example, if several children have a recorded score of 12 pull-ups on a physical fitness test, their performances probably range from those who just barely raised their chin over the bar the twelfth time and were finished (lower real limit) to those who completed 12 pull-ups, came up again, and

PICTURE THIS

Joe Rocco

almost raised their chin over the bar, but did not quite make it for pull-up 13 (upper limit). Thus, a score of 12 is considered as representing an interval from halfway between 11 and 12 (11.5) to halfway between 12 and 13 (12.5) or an interval of 1. For example, given the following distribution of scores, you find the range by subtracting 1.5 (the lower limit of the lowest score) from 16.5 (the upper limit of the highest score), which is equal to 15. It is simpler to use Formula 6.3:

2 10 11 12 13 14 16

$$R = (X_h - X_l) + I \tag{6.3}$$

where

R = range
X_h = highest value in a distribution
X_l = lowest value in a distribution
I = interval width

Subtract the lower number from the higher and add 1 (16 − 2 + 1 = 15). In frequency distribution, 1 is the most common interval width.

The range is an unreliable index of variability because it is based on only two values, the highest and the lowest. It is not a stable indicator of the spread of the scores. For this reason, the use of the range is mainly limited to inspectional purposes. Some research reports refer to the range of distributions, but such references are usually used in conjunction with other measures of variability, such as variance and standard deviation.

THINK ABOUT IT 6.4

1. a. What is the range of Mr. Li's physics exam scores?
   b. What is the range of the Ed 101 scores?

Answers

1. a. 23.5 − 15.5 = 8 or, using Formula 6.3, $X_h - X_l + I = 23 - 16 + 1 = 8$.
   b. 36.5 − 20.5 = 16 or $X_h - X_l + I = 36 - 21 + 1 = 15 + 1 = 16$. (Note that the highest occurring score was 36; the lowest occurring score was 21.)

VARIANCE AND STANDARD DEVIATION

Variance and standard deviation are the most frequently used indexes of variability. They are both based on **deviation scores**—scores that show the difference between a raw score and the mean of the distribution. The formula for a deviation score is

$$x = X - \overline{X} \tag{6.4}$$

x = deviation score
X = raw score
$\overline{X}$ = mean

Scores below the mean will have negative deviation scores, and scores above the mean will have positive deviation scores. For example, the mean in Mr. Li's physics exam is 20; thus, Ona's deviation score is $x = 22 - 20 = 2$, whereas Ted's deviation score is $16 - 20 = -4$. By definition, the sum of the deviation scores in a distribution is always 0. Thus, to use deviation scores in calculating measures of variability, you must find a way to work around the fact that $\Sigma x = 0$. The technique used is to square each deviation score so that they all become positive numbers. If you then sum the squared deviations and divide by the number of scores, you have the mean of the squared deviations from the mean, or the **variance**. In mathematical form, variance is

$$\sigma^2 = \frac{\Sigma x^2}{N} \tag{6.5}$$

where

σ^2 = variance
Σ = sum of
x^2 = deviation of each score from the mean $(X - \overline{X})$ squared, otherwise known as the deviation score squared
N = number of cases in the distribution

In column 4 of Table 6.3, we see the deviation scores, differences between each score, and the mean. Column 5 shows each deviation score squared (x^2), and column 6 shows the frequency of each score from column 2 multiplied by x^2. Summing column 6 gives us the sum of the squared deviation scores $\Sigma x^2 = 72$. Dividing this by the number of scores gives us the mean of the squared deviation scores or the variance.

$$\sigma^2 = \frac{\Sigma x^2}{N} = \frac{72}{18} = 4$$

The foregoing procedure is convenient only when the mean is a whole number. This rarely occurs except in textbook examples. We have chosen to do our examples with whole number means so you can understand the concept and not become bogged down with the mathematics.

Formula 6.6 avoids the tedious task of working with squared mixed-number deviation scores such as 7.6667^2. Using Formula 6.6 yields the desired result with much less labor. Thus, we recommend that students always use this formula for computing standard deviation if the computation must be done by hand:

$$\sigma^2 = \frac{\Sigma X^2 - \dfrac{(\Sigma X)^2}{N}}{N} \tag{6.6}$$

where

σ^2 = variance
ΣX^2 = sum of the squares of each score (i.e., each score is first squared, and then these squares are summed)
$(\Sigma X)^2$ = sum of the scores squared (the scores are first summed, and then this total is squared)
N = number of cases

Table 6.3 Variance of Mr. Li's Physics Exam Scores

(1) X	(2) f	(3) fX	(4) x	(5) x^2	(6) fx^2	(7) X^2	(8) fX^2
23	2	46	+3	9	18	529	1058
22	2	44	+2	4	8	484	968
21	4	84	+1	1	4	441	1764
20	4	80	0	0	0	400	1600
19	2	38	−1	1	2	361	722
18	2	36	−2	4	8	324	648
17	0	0					
16	2	32	−4	16	32	256	512

Column 7 in Table 6.3 shows the square of the raw scores. Column 8 shows these raw score squares multiplied by frequency. Summing this fX^2 column gives us the sum of the squared raw scores:

$$\sigma^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N} = \frac{7272 - \frac{360^2}{18}}{18} = \frac{7272 - \frac{129600}{18}}{18} = \frac{7272 - 7200}{18} = \frac{72}{18} = 4$$

Note that this result is the same as that which we obtained with Formula 6.5.

Because each of the deviation scores is squared, the variance is necessarily expressed in units that are squares of the original units of measure. For example, you might find that the variance of the heights of children in a class is 9 square inches. This would tell you that this class is more heterogeneous in height than a class with a variance of 4 square inches and more homogeneous than a class with a variance of 16 square inches.

In most cases, educators prefer an index that summarizes the data in the same unit of measurement as the original data. **Standard deviation (σ)**, the positive square root of variance, provides such an index. By definition, the standard deviation is the square root of the mean of the squared deviation scores. Rewriting this definition using symbols, you obtain

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \tag{6.7}$$

For Mr. Li's physics exam scores, the standard deviation is

$$\sqrt{\frac{72}{18}} = \sqrt{4} = 2$$

The standard deviation belongs to the same statistical family as the mean; that is, like the mean, it is an interval or ratio statistic, and its computation is based on the size of individual scores in the distribution. It is by far the most frequently used measure of variability and is used in conjunction with the mean.

Formulas 6.5, 6.6, and 6.7 are appropriate for calculating the variance and the standard deviation of a population. If scores from a finite group or sample are used to estimate the heterogeneity of a population from which that group was drawn,

research has shown that these formulas more often underestimate the population variance and standard deviation than overestimate them. Mathematically, to derive unbiased estimates, $N-1$ rather than N is used as the denominator.

The formulas for variance and standard deviation based on sample information are

$$s^2 = \frac{\Sigma x^2}{N-1} \tag{6.8}$$

$$s = \sqrt{\frac{\Sigma x^2}{N-1}} \tag{6.9}$$

$$s = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N-1}} \tag{6.10}$$

Following the general custom of using Greek letters for population parameters and Roman letters for sample statistics, the symbols for variance and standard deviation calculated with $N-1$ are s^2 and s, respectively.

With the data in Table 6.3,

$$\frac{72}{18-1} = 4.24 \quad \text{and} \quad s = \sqrt{4.24} = 2.06$$

Formulas 6.8, 6.9, and 6.10 are often used to calculate variance and standard deviation even when there is no intention to estimate population parameters. Many computers and calculators calculate variance and standard deviation this way unless instructed to do otherwise.

Spread, scatter, heterogeneity, dispersion, and volatility are measured by standard deviation, in the same way that volume is measured by bushels and distance is measured by miles. A class with a standard deviation of 1.80 on reading grade level is more heterogeneous than a class with a standard deviation of 0.70. A month when the daily Dow Jones Industrial Average has a standard deviation of 40 is more volatile than a month with a standard deviation of 25. A school where the teachers' monthly salary has a standard deviation of $900 has more disparity than a school where the standard deviation is $500.

MEASURES OF RELATIVE POSITION

Measures of relative position indicate where a score falls in relation to all other scores in the distribution. Researchers often want to assess an individual's relative position in a group, or compare the relative position of one individual on two or more measures or of two or more individuals on the same measure. The most widely used statistics for these purposes are *z* scores, stanines, other standard scores, and percentile rank.

Z SCORE

The most widely used measure of relative position is the ***z* score**, which indicates the positive or negative difference between an individual score and the mean as measured in standard deviation units. It and other indexes derived from it are known as **standard scores**. The *z* score is defined as the distance of

a score from the mean as measured by standard deviation units. The formula for finding a z score is

$$z = \frac{x}{\sigma} = \frac{X - \overline{X}}{\sigma} \tag{6.11}$$

where

X = raw score
$\overline{X}$ = mean of the distribution
σ = standard deviation of the distribution
x = deviation score $(X - \overline{X})$

Applying this formula, a score exactly 1 standard deviation above the mean becomes a z of $+1$, a score exactly 1 standard deviation below the mean becomes a z of -1, and so on. A score equal to the mean will have a z score value of 0. For illustration, suppose a student's score on a psychology test is 72, where the mean of the distribution is 78 and the standard deviation equals 12. Suppose also that the same student has scored 48 on a statistics test, where the mean is 51 and the standard deviation is 6. If you substitute these figures for the appropriate symbols in Formula 6.11, you can derive a z score for each test:

Psychology $$z_1 = \frac{72 - 78}{12} = -0.50$$

Statistics $$z_2 = \frac{48 - 51}{6} = -0.50$$

Both these standard scores belong to the z distribution, where by definition the mean is always 0 and the standard deviation is 1, and therefore they are directly comparable. It is apparent in this example that the score of 72 on the psychology test and the score of 48 on the statistics test are equivalent—that is, both scores indicate the same relative level of performance. In other words, the standing of this student is the same in both tests when compared with the performance of the other students. It would be very difficult to make such a comparison without employing the z score technique.

Let us use another example: Suppose a student who has taken the same tests has obtained a score of 81 on the psychology test and a score of 53 on the statistics test. As before, it is difficult to compare these raw scores to show on which test this student has performed better. Converting the scores to z scores makes the comparison easy. Using Formula 6.11, we find the values of z_1 and z_2 in this case to be as follows:

Psychology $$z_1 = \frac{81 - 78}{12} = +0.25$$

Statistics $$z_2 = \frac{53 - 51}{6} = +0.33$$

This result shows that the score of 53 on the statistics test actually indicates a slightly better relative performance than the score of 81 on the psychology test. Compared with the other students, this student has performed somewhat better in statistics than in psychology.

Because the mean of the z scores in any distribution is 0 and the standard deviation is 1, they have universal meaning. A z score of -0.10 is slightly below

average in a distribution of statistics test scores, a distribution of weights of people in a weight control program, a distribution of pork belly prices, or any other distribution. A z score of +2.40 is very high, whether you are talking about achievement scores, scores on a measure of depression, corn yield per acre, or any other measure.

OTHER STANDARD SCORES

Scores can also be transformed into other standard score scales that do not involve negative numbers or decimals. One of the most common procedures is to convert to ***T* scores** by multiplying the z scores by 10 and adding 50. This results in a scale of positive whole numbers that has a mean of 50 and a standard deviation of 10. The T score formula is

$$T = 10(z) + 50 = 10\left(\frac{X - \overline{X}}{\sigma}\right) + 50 \tag{6.12}$$

A score of 21 on a test for which the mean of the scores is 27 and the standard deviation is 6 would have a z score of −1.00 or a T score of 40:

$$T = 10\left(\frac{21 - 27}{6}\right) + 50 = 40 \tag{6.13}$$

The transformation of z scores into T scores not only enables you to work with whole numbers, but also avoids the adverse psychological implications of describing subjects' performances with negative numbers. In the preceding example, it would be easier to report that the student's score is 40 where the mean score is 50 than to report a score of −1.00 with an average of zero.

In addition to T, there are other transformed standard score distributions. To transform a distribution of scores to a new standardized distribution, multiply the z score by the desired standard deviation and add the desired mean. The general formula is as follows:

$$A = \sigma_A(z) + \mu_A \tag{6.14}$$

where

A = standard score on the new scale
μ_A = mean for the new standard scale
σ_A = standard deviation for the new standard scale

For example, College Entrance Examination Board (CEEB) scores have a mean of 500 and a standard deviation of 100 for its transformed distribution. If you were 1.50 standard deviations above the mean ($z = 1.50$) on the verbal section of the Scholastic Assessment Test (SAT), your score would be reported as 650, which is 500 + (100)(1.50). If your quantitative score were 500, you would have scored exactly at the mean.

The Wechsler Adult Intelligence Test scores are standard scores with a mean of 100 and a standard deviation of 15. A raw score on the mean is reported as 100. A raw score 1 standard deviation below the mean is reported as 85. A raw score 2 standard deviations above the mean is reported as 130.

Transforming a set of scores to standard scores does not alter the shape of the original distribution. If a distribution of scores is skewed, the derived

standard scores also produce a skewed distribution. Only if the original distribution is normal do the standard scores produce a **normal distribution**.

STANINE SCORES

During World War II, the U.S. Army Air Corps developed a standard system of nine scores called **stanine scores** to help its personnel interpret *z* scores. Stanines avoid negative numbers and decimals. A stanine score of 5 represents *z* scores that are average or slightly above or slightly below average—that is, equivalent to *z* scores between −0.25 and +0.25. From there, stanine scores go up to 9 and down to 1 in increments of 0.50, as shown in Table 6.4. Stanines are standardized with the mean of 5 and a standard deviation of 2. The formula for stanines is $2z + 5$. You convert a *z* score to a stanine by multiplying by 2 and adding 5. Stanines are always rounded to the nearest whole number. Whenever this formula yields a result greater than 9, the value 9 is assigned. Whenever the result is less than 1, the value 1 is assigned. Because all *z* scores above 1.75 are assigned a stanine score of 9 and all *z* scores below −1.75 are assigned a score of 1, stanine scores are not useful for comparing extreme scores. Stanines are easy to comprehend. Like all transformations of the *z* score, they have universal meaning. A stanine score of 4 always means below average but not too far below average. Stanines are often used in school systems for reporting students' standardized test scores.

THINK ABOUT IT 6.5

Recall the scores in Mr. Li's physics class in Think About It 6.2.

1. What is the *z* score of a raw score of 21 on Mr. Li's exam?
2. What is the *z* score of a raw score of 18 on Mr. Li's exam?
3. What is the *z* score of a raw score of 20 on Mr. Li's exam?
4. What is the stanine for each of these scores?

Answers

1. +0.50; 2. −1.0; 3. 0
4. Stanines: 1. 6; 2. 3; 3. 5

Table 6.4 Conversion of *z* Scores to Stanines

z Score	Stanine	Interpretation	Percent in Stanine
Above +1.75	9	Among the very highest scores	4
+1.25 to + 1.75	8	Quite well above average	7
+0.75 to +1.25	7	Quite noticeably above average	12
+0.25 to +0.75	6	Above average	17
−0.25 to +0.25	5	Near the average	20
−0.75 to −0.25	4	Below average	17
−1.25 to −0.75	3	Quite noticeably below average	12
−0.75 to −1.25	2	Quite well below average	7
Below −1.75	1	Among the lowest scores	4

PERCENTILE RANK

A measure of relative position that most people find easy to understand and interpret is the **percentile rank** (PR), which indicates the percentage of scores in a distribution that are equal to and fall below a given score point. It is easy to picture a score with a PR of 32 as having 32 percent of the scores in its distribution as equal to this and below it, and a score with a PR of 89 as having 89 percent of the scores equal to it and below it.

The following is a simple way to calculate PR:

1. Arrange the scores in a frequency distribution.
2. Determine the cumulative frequency of scores below the interval containing the score of interest and add one half of the frequency of scores within the interval containing the score of interest.
3. Divide this sum by the total number of scores and multiply by 100.

Consider Mr. Li's physics exam scores (see Table 6.2). To calculate the PR of 21, we start with the cumulative frequency below a score of 21 ($cf = 10$) and add one half of the number scoring 21 ($f_w = 2$). We then divide this number by the total number of students who took the test ($N = 18$). Finally, we multiply the result by 100 and round to the nearest whole number:

$$\text{PR} = \frac{cf_b + \frac{f_w}{2}}{N}(100) \tag{6.15}$$

which is rounded to 67.

A score of 18 is assigned the following PR:

$$\text{PR} = \frac{cf_b + \frac{f_w}{2}}{N}(100) = \frac{2 + \frac{2}{2}}{18}(100) = \frac{3}{18}(100) = 16.67$$

which is rounded to 17.

The major advantages of percentile ranks are as follows:

1. They have universal meaning. A score with a percentile rank of 89 is high in any distribution. A score with a percentile rank of 32 is somewhat low in any distribution.
2. The familiar concept of 0 to 100 percent applies to the interpretation of percentile rank. Schools often report percentile ranks to parents.

The major disadvantages of percentile rank are as follows:

1. As with other ordinal statistics, percentile ranks cannot be added, subtracted, multiplied, or divided.
2. As with all ordinal indexes, equal differences between percentile ranks do not represent equal differences between the scores in the original distribution. If there are many scores near a particular score, a small change in score will produce a major change in percentile rank. If there are few scores near a particular score, a considerable change in raw score will be necessary to produce a change in position relative to other

scores and thus a change in percentile rank. For example, a professor has recorded the weights of the students in his Physical Education 202 class and used them to illustrate the computation of percentile rank. The result is the polygon shown in Figure 6.8.

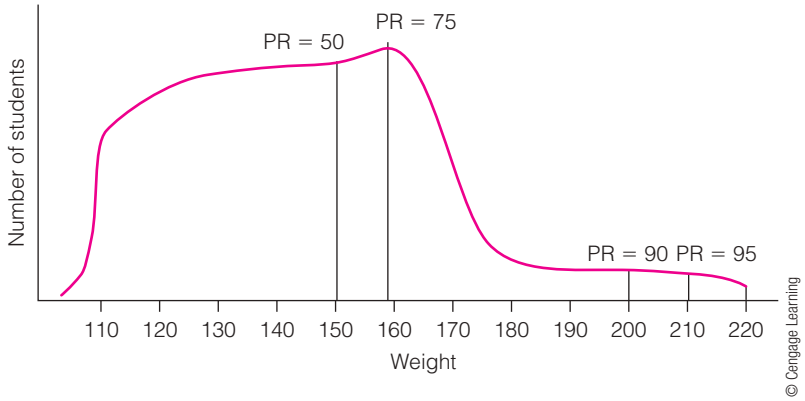


Figure 6.8 Weights of Students in Physical Education 202 Class

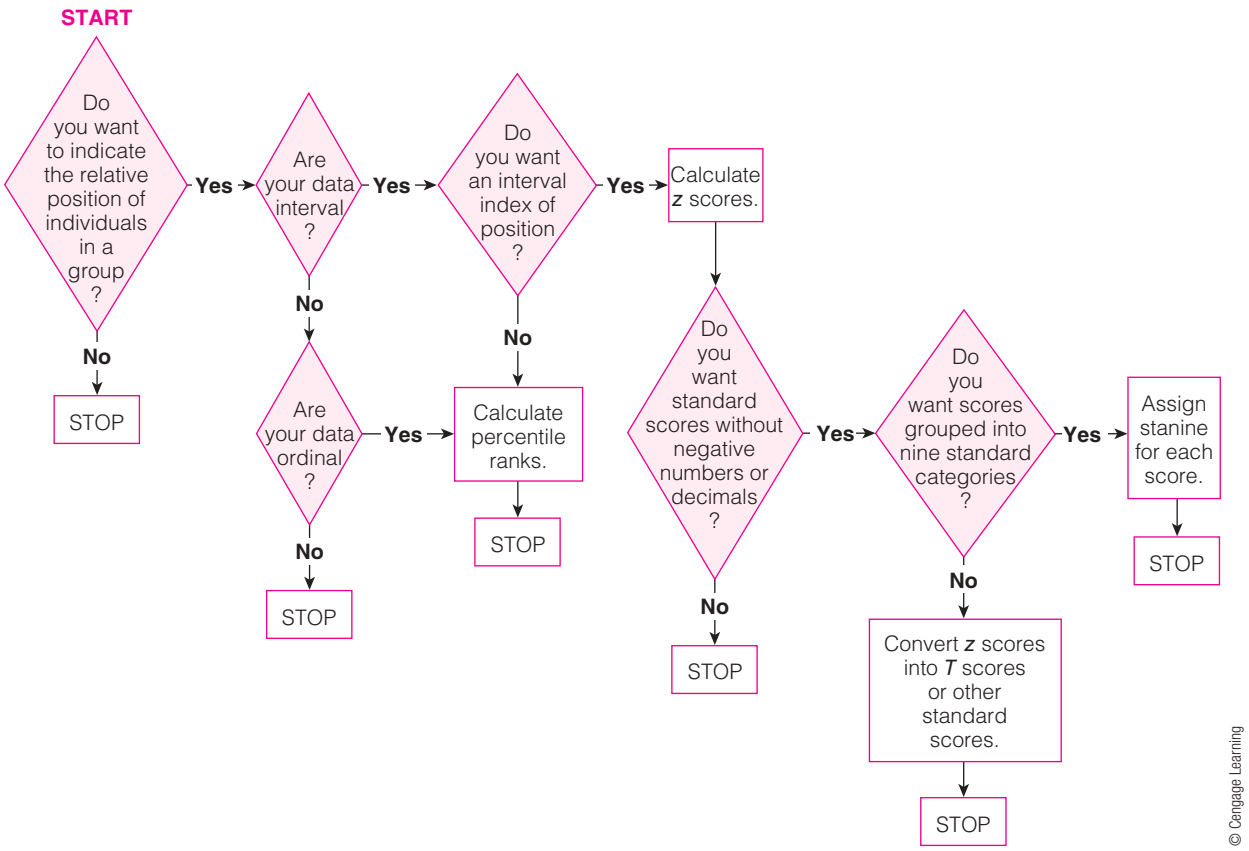


Figure 6.9 Measures of Relative Position

A 160-pound student and a 210-pound student both resolve to lose weight and actually lose 10 pounds each. The 10-pound loss moves the 160-pound student from a percentile rank of 75 to a percentile rank of 50. The same weight loss only changes the heavier student's percentile rank from 95 to 90. Often, many cases concentrate near the middle of a distribution and then taper off with few cases occurring at the extreme ends. In such distributions, minor differences in raw scores will appear as major differences in percentile ranks among those scores that are near the center of the distribution where a large number of the scores typically are located. At the extreme ends of the distribution, where there are few scores, major differences in raw score will have only minor effects on percentile rank. We discuss the phenomenon more closely when we consider the normal curve.

Figure 6.9 shows the process of deciding which index to choose for indicating relative position.

THE NORMAL CURVE

Recall the example of deductive reasoning in Chapter 1, where we concluded that if the probability of having a son is 50 percent and the probability of having a daughter is 50 percent, the probability of two daughters is 25 percent, the probability of two sons is 25 percent, and the probability of one of each is 50 percent. Abraham DeMoivre (1667–1754) pondered the probabilities of various outcomes when the percent of likelihood in each trial is 50/50, as in heads and tails in honest coin flips. He came up with a formula to predict the probabilities of various number of heads (or tails) when a coin is flipped an infinite number of times. The most frequent score is half infinity, the next most frequent scores are half infinity plus one and half infinity minus one, and so forth. When a polygon of the expected proportions of various z scores is made, the outcome is a bell-shaped curve called the **normal curve**. This model proved very useful for gamblers interested in predicting the probability of various gaming outcomes.

Soon after the normal curve was developed, it was noticed that many naturally occurring distributions, such as height, weight, and IQ, formed polygons resembling the normal curve. If in America you measure boys on their 10th birthday, you will find many whose height is near the mean and slightly fewer boys who are a bit above or a bit below the mean. The further you move from the mean, the fewer boys you will find at each height. As in the normal curve probability model, z scores near 0 will be expected to occur more frequently than other z score values, and the farther from 0 a z score is, the less frequently it will be expected to occur.

Inasmuch as so many naturally occurring distributions resemble the normal curve, this theoretical model has proved very useful in research. Whenever actual data are known or believed to resemble the normal curve in distribution, you can deduce many useful estimates from the theoretical properties of the normal curve.

Note in Figure 6.10 that the normal curve is a symmetrical distribution with the same number of cases at specified z score distances below the mean as above the mean. Its mean is the point below which exactly 50 percent of the cases fall and above which the other 50 percent of the cases are located.

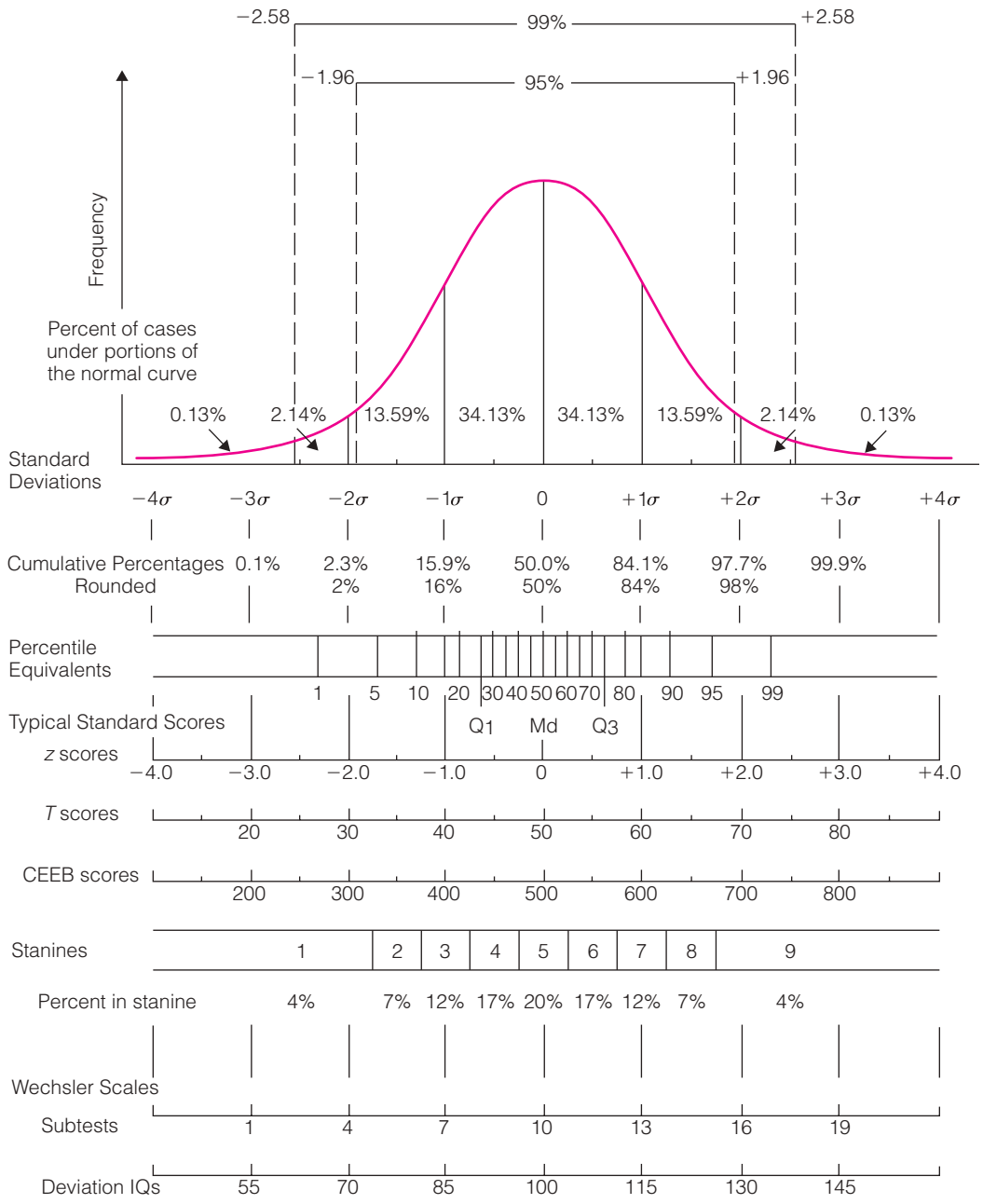


Figure 6.10 Normal Curve and Equivalents

Source: Test Service Notebook 148. Originally published as Test Service Bulletin No. 48, January 1955. Updated to include Normal Curve Equivalents, September 1980 by NCS Pearson, Inc. Reproduced with permission. All rights reserved.

Because the curve is symmetrical, the mean, the median, and the mode are identical. In a normal distribution, most of the cases concentrate near the mean. The frequency of cases decreases as you proceed away from the mean in either direction. Approximately 34 percent of the cases in a normal distribution fall between the mean and 1 standard deviation above the mean, and approximately 34 percent are between the mean and 1 standard deviation

below the mean. Between 1 and 2 standard deviations from the mean on either side of the distribution are approximately 14 percent of the cases. Only approximately 2 percent of the cases fall between 2 and 3 standard deviations from the mean, and only approximately one-tenth of 1 percent of the cases fall above or below 3 standard deviations from the mean.

These characteristics can be seen in Figure 6.10. You can see that approximately one-sixth of the curve falls to the left of 1 standard deviation below the mean. The first line under the curve shows standard deviations from −4 to +4. These are equivalent to *z* scores from −4.00 to +4.00. The cumulative percentage line tells you that 15.90 percent of scores fall below −1 and 97.70 percent falls below +2, and so on. The line following cumulative percentage shows these cumulative percentage scores rounded to the nearest whole percentage. Multiplying each of these numbers by 100 gives you percentile rank.

Percentile equivalents are shown on the next line, which also shows the first quartile (Q_1 sets off the lowest 25 percent of scores), the median (Md), and the third quartile (Q_3 sets off the lower 75 percent or upper 25 percent of scores). Note how slowly the percentile equivalents change below Q_1 and above Q_3 and how rapidly they change between these two points. The next line after percentile equivalents shows *z* scores, which are identical to the scores on the standard deviation line. Following the *z* score line are various standard scores transformed from *z* scores including *T* scores, CEEB scores, stanines, percent in stanine, Wechsler subtest scores, and Weschsler deviation IQs. Note that 95 percent of the normal curve falls between plus and minus $z = 1.96$, and 99 percent falls between plus and minus $z = 2.58$. These boundaries become important when we discuss the use of the normal curve in inferential statistics. To determine the exact percentage of the cases below and above each *z* score in the normal distribution, consult Table A.1 in the Appendix, which gives the areas of the normal curve. Column 1 of Table A.1 contains different *z* values. Column 2 gives the area under the curve between the mean and each *z* value. Column 3 shows the remaining area from each *z* score to the end of the curve. Therefore, the areas in column 2 and column 3 add up to .5000. Take as an example the *z* value of +0.70. The area between this *z* value and the mean can be found in column 2; it is .2580. This figure indicates that 26 percent of the cases fall between this *z* value and the mean of the distribution. Because the mean of the normal distribution coincides with the median, 50 percent of the cases lie below the mean. Add 0.50 to the .2580, and the result tells you that you can expect 75.80 percent of the cases to fall below the *z* value of +0.70. Column 3 indicates that the other 24.20 percent of the cases fall above the *z* value of +0.70.

This procedure is reversed when the *z* value is negative. Suppose you want to find the percentage of cases below the *z* value of −0.70. The area between the mean and a *z* score of −0.70 is .2580 or, in terms of percentage, 25.80 percent of the cases. Subtracting 25.80 from 50, results in 24.20, indicating that only 24.20 percent of the scores lie below a *z* value of −0.70 in a normal distribution. This value can also be found in column 3 of the table, which gives a value of .2420 for a *z* score of 0.70. The percentage of scores above −0.70 is 100 − 24.20, or 75.80 percent. Because the normal

curve is symmetrical, we do not need separate tables for positive and negative z scores. You just have to remember the sign of the z score with which you are working.

Among other applications, the normal curve can be used to help people interpret standard scores. For example, how high is a score of 650 on the SAT? The SAT has a mean of 500 and a standard deviation of 100, so the z score here is 1.50. Consulting Table A.1, column 2, you find .4332 of the normal curve falls between the mean and $z = 1.50$. Adding the 50 percent below the mean, you can say that an SAT score of 600 exceeds the scores of 93 percent of SAT scores.

Because it is known that the population distribution of SAT scores closely resembles the normal curve, PR approximations based on the normal curve will be quite near to the actual PRs. With other scores, as actual distributions become increasingly less like the normal curve, the PR approximations become more distant. Where the shape of a distribution is not known, it is usually reasonable to assume a distribution similar to the normal curve and to use the normal curve table to find reasonable approximations of the PRs of various z scores. The more the actual shape differs from the normal, the less useful the approximations become.

The most common use of the normal curve in descriptive statistics is going from a given z score to a percentile rank, as described in the previous paragraph, but we can also use it to go in the opposite direction, from a given percentile rank to its z score equivalent.

CORRELATION

After completing the second unit in physics class, Mr. Li gave a second exam. Table 6.5 lists his students in column 1. Their z scores on test 1 are shown in column 2, and their z scores on test 2 are shown in column 3. Recall that z scores are a way to indicate the relative positions of scores. They have universal meaning and can be used with any interval or ratio data.

Table 6.5 reveals that there is a tendency for those who had positive z scores on test 1 to have positive z scores on test 2, and for those with negative z scores on test 1 to have negative z scores on test 2. Four students have identical z scores on both tests. The others have z scores with the same sign but different values, except Ali, who had a positive z score on test 1 and a negative z score on test 2.

Figure 6.11 shows a histogram of the first test scores on the abscissa (x) and a histogram of the second test scores turned sideways on the ordinate (y). In the upper right part of Figure 6.11 you find each student's position on both the first and the second test. This gives a picture of how the students tend to have similar z scores on the two tests, but there is some shifting of their relative positions. There is a strong but not perfect positive relationship between the relative positions of each student's scores on the two tests.

Correlations indicate the relationship between paired scores. The correlation indicates whether the relationship between paired scores is positive or negative and the strength of this relationship. The pairs may be two scores

Table 6.5 Mr. Li's First and Second Test *z* Scores

	Test 1 z Scores	Test 2 z Scores	z Score Products
Student	z_x	z_y	$(z_x z_y)$
Ali	+0.50	−1	−0.50
Ann	0	0	0
Ben	+1.50	+1	1.50
Cal	0	−1	0
Dan	0	+0.50	0
Ed	+0.50	+0.50	0.25
Ima	+1	+1.50	1.50
Jan	−0.50	0	0
Kay	−2	−1.50	3
Lee	0	−1	0
Mel	−1	−0.50	0.50
Mia	+1.50	+1	1.50
Ned	+0.50	+1.50	0.75
Ona	+0.50	+1	0.50
Sam	+1	+1	1
Sue	−0.50	−0.50	0.25
Ted	−2	−1	2
Van	−1	–1.50	1.50
			$\Sigma = 13.75$

for the same individual, natural pairs such as husbands and wives, or two individuals matched on some measure such as reading test scores. In addition to looking at correlation through visual means, the researcher can calculate a **correlation coefficient** that represents the correlation.

PEARSON PRODUCT MOMENT CORRELATION COEFFICIENT

A very useful statistic, the **Pearson product moment correlation coefficient** (Pearson *r*), indicates both the direction and the magnitude of the relationship between two variables without needing a scatterplot to show it.

Start with the knowledge that because of the way *z* scores are defined, the sum of the *z* scores squared in any distribution will always equal the number in that distribution. (To test this, square the *z* scores in column 1 or column 2, and find that the sum of the *z* scores squared in each column is 18.) Therefore, the mean of the squared deviations is always 1.

For example, if you measured precisely the square feet in each room in your building (x), and then precisely measured each room in square meters (y), the *z* scores would be identical and the Pearson *r* would be +1.0.

Taking a schedule of Amtrak trains traveling between Chicago and Seattle, showing the miles from Seattle (x) and miles from Chicago (y), the *z* scores of each station in terms of miles from Seattle (x) would be the exact opposite of

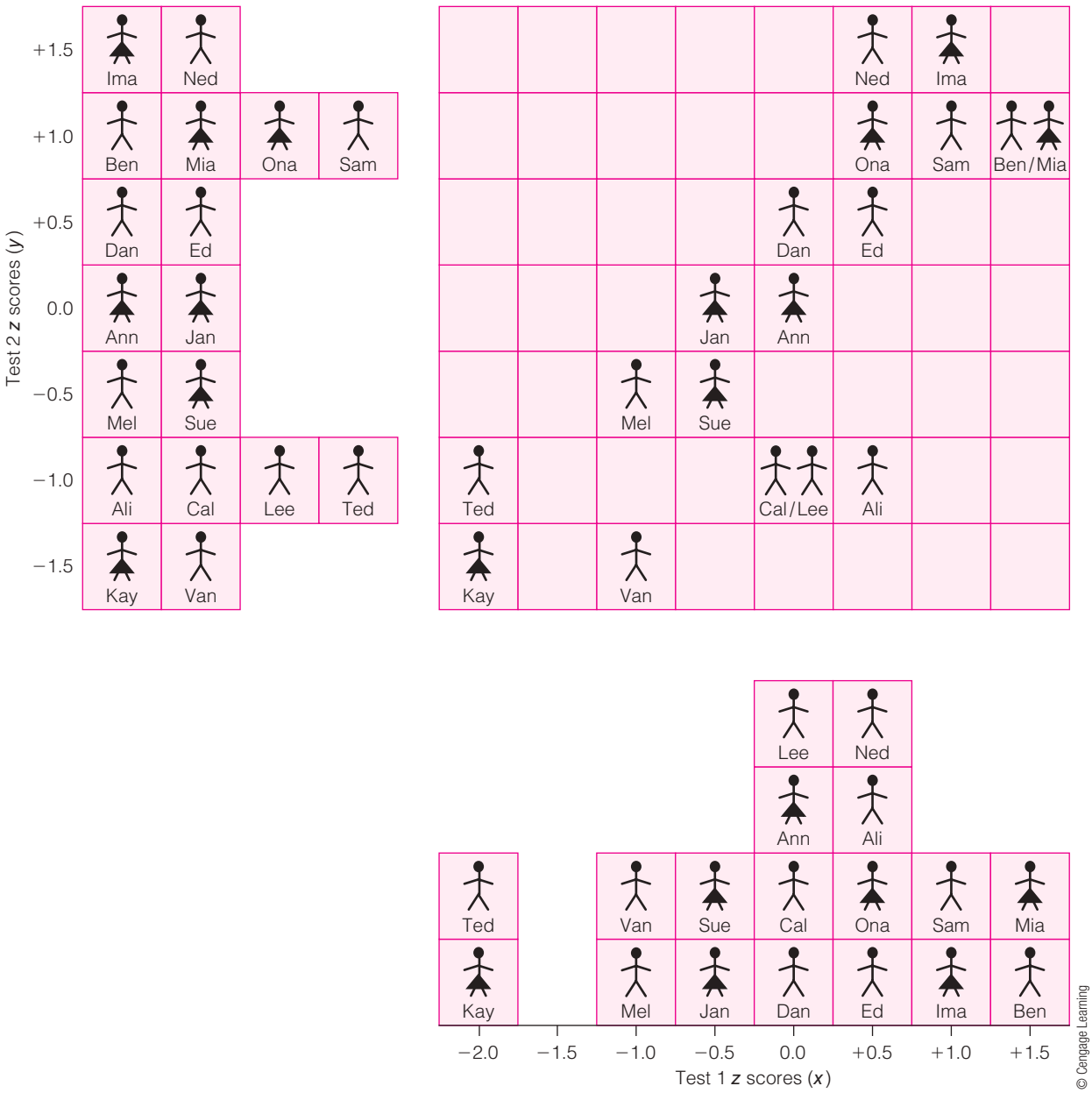


Figure 6.11 Mr. Li's Students' *z* Scores on First and Second Exam Scores

z scores on *y* (miles from Chicago). Each mile farther from Chicago is a mile closer to Seattle. The Pearson *r* would be −1.0.

We have seen that the *z* scores on Mr. Li's two tests in Table 6.5 are similar but not perfectly aligned, so we know that the *z* score product averages for *x* and *y* will be less than +1.0 but approaching +1.0.

The definition of the Pearson *r* is simplicity itself. It is the mean $z_x z_y$ product:

$$r = \frac{\Sigma z_x z_y}{N} \tag{6.16}$$

where

r = Pearson product moment coefficient of correlation
$\Sigma z_x z_y$ = sum of the z score products
N = number of paired scores

The sum of the z score products in Table 6.5, column 4, is 13.75. Therefore, $r = 13.75/18 = .76$. This confirms what we had already concluded: The z scores on the two tests *are positively related, and this relationship is strong* but not perfect.

Whenever individuals tend to have z scores of the same sign but do not have exactly the same z score on X that they have on Y, the sum of the z scores will be positive but less than N. Therefore, the mean z score product, the Pearson r, will be less than +1. If positive z_x scores tend to be paired with negative z_y scores, but they are not perfect mirror images of each other, the sum of the $z_x z_y$ products will be a negative number, nearer to zero than to negative N. Therefore, the mean will be between −1 and 0. If there is a strong but not perfectly negative relationship, the r will be near to −1.00. If there is no overall relationship between the paired z scores, their product will be zero and their mean will be zero.

Here we have an index that indicates not only the direction of relationships between variables, but also the strength of the relationships. This index is never greater than +1.00 or less than −1.00. It has universal meaning.

Means and standard deviations that are whole numbers almost never occur except in textbooks. Typical z scores are awkward decimal or mixed numbers, such as .4716 or 1.6667. Formula 6.16 is useful for understanding what the Pearson r means, but it is hopeless for calculation. Formula 6.17 avoids the need for calculating z scores and multiplying awkward decimals and mixed numbers. It also avoids rounding errors. Its result is the same as that of Formula 6.16:

$$r = \frac{\Sigma XY - \dfrac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\left(\Sigma X^2 - \dfrac{(\Sigma X)^2}{N}\right)\left(\Sigma Y^2 - \dfrac{(\Sigma Y)^2}{N}\right)}} \tag{6.17}$$

where

r = Pearson r
ΣX = sum of scores in X distribution
ΣY = sum of scores in Y distribution
ΣX^2 = sum of the squared scores in X distribution
ΣY^2 = sum of the squared scores in Y distribution
ΣXY = sum of products of paired X and Y scores
N = number of paired X and Y scores (subjects)

Table 6.6 provides the data needed to calculate the Pearson r for Mr. Li's tests 1 and 2 using Formula 6.17. Column 1 lists the students. Column 2 shows each student's raw score on test 1 (X), and column 3 shows these raw scores squared (X^2). Column 4 shows each student's raw score on test 2 (Y), and column 5 shows these scores squared (Y^2). Column 6 shows the product of each student's X raw score multiplied by his or her Y raw score (XY).

Table 6.6 Mr. Li's Physics Class Raw Scores Illustrating the Calculation of the Pearson *r*

(1)	(2)	(3)	(4)	(5)	(6)
Name	X	X^2	Y	Y^2	XY
Ali	21	441	22	484	462
Ann	20	400	26	676	520
Ben	23	529	30	900	690
Cal	20	400	22	484	440
Dan	20	400	28	784	560
Ed	21	441	28	784	588
Ima	22	484	32	1024	704
Jan	19	361	26	676	494
Kay	16	256	20	400	320
Lee	20	400	22	484	440
Mel	18	824	24	576	432
Mia	23	529	30	900	690
Ned	21	361	32	1024	672
Ona	21	331	30	900	630
Sam	22	484	30	900	660
Sue	19	361	32	1024	672
Ted	16	256	22	484	352
Van	18	324	20	400	360
	$\Sigma X = 360$	$\Sigma X^2 = 7272$	$\Sigma Y = 468$	$\Sigma Y^2 = 12456$	$\Sigma XY = 9470$

Using Formula 6.17, we obtain

$$r = \frac{9470 - \frac{(360)(468)}{18}}{\sqrt{\left(7272 - \frac{(360)^2}{18}\right)\left(12456 - \frac{(468)^2}{18}\right)}}$$

$$= \frac{9470 - \frac{168480}{18}}{\sqrt{\left(7272 - \frac{129600}{18}\right)\left(12456 - \frac{219024}{18}\right)}}$$

$$= \frac{9470 - 9360}{\sqrt{(7272 - 7200)(12456 - 12168)}}$$

$$= \frac{9470 - 9360}{\sqrt{(72)\,(288)}} = \frac{9470 - 9360}{\sqrt{20736}} = \frac{9470 - 9360}{144} = \frac{110}{144} = .76$$

SCATTERPLOTS

The upper right part of Figure 6.11 shows icons with names to illustrate the concept that each icon represents an individual's z scores on both dimensions, test 1 and test 2. It is easier in practice to use a dot to represent individuals' pairs of scores and plot them on a graph called a **scatterplot**, as shown in Figure 6.12.

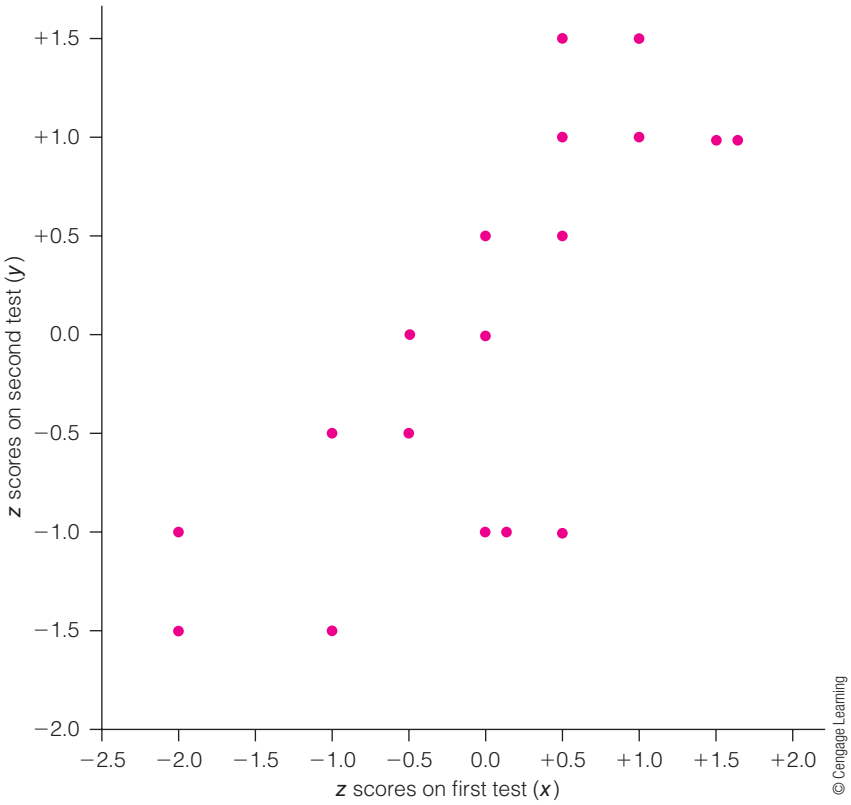


Figure 6.12 Scatterplot of Mr. Li's Students' First and Second Test Scores

In a research situation, the z scores on the horizontal axis will be those of the independent variable, with the lowest z score on the left and the highest z score on the right. The z scores on the vertical axis will be those of the dependent variable (y), with the lowest z score at the bottom and the highest z score at the top.

A scatterplot illustrates the direction of the relationship between the variables. A scatterplot with dots going from lower left to upper right indicates a **positive correlation** (as variable x increases, variable y also increases). One with dots going from upper left to lower right indicates a **negative correlation** (as variable x increases, variable y decreases).

A scatterplot of z scores also reveals the strength of the relationship between variables. If the dots in the scatterplot form a narrow band such that when a straight line is drawn through the band the dots will be near the line, there is a strong **linear relationship** between the variables. However, if the dots in the z score scatterplot scatter widely, the relationship between variables is relatively weak. The scatterplots in Figure 6.13 show various positive and negative and strong and weak relationships, expressed mathematically by r.

You can see in comparing these scatterplots that the tilt of the "cloud" of dots becomes increasingly less as r moves from a 45° angle for $r = +1.00$ or $r = -1.00$ until it reaches $r = 0$, where it is flat and matches the line for the mean of the xy scores.

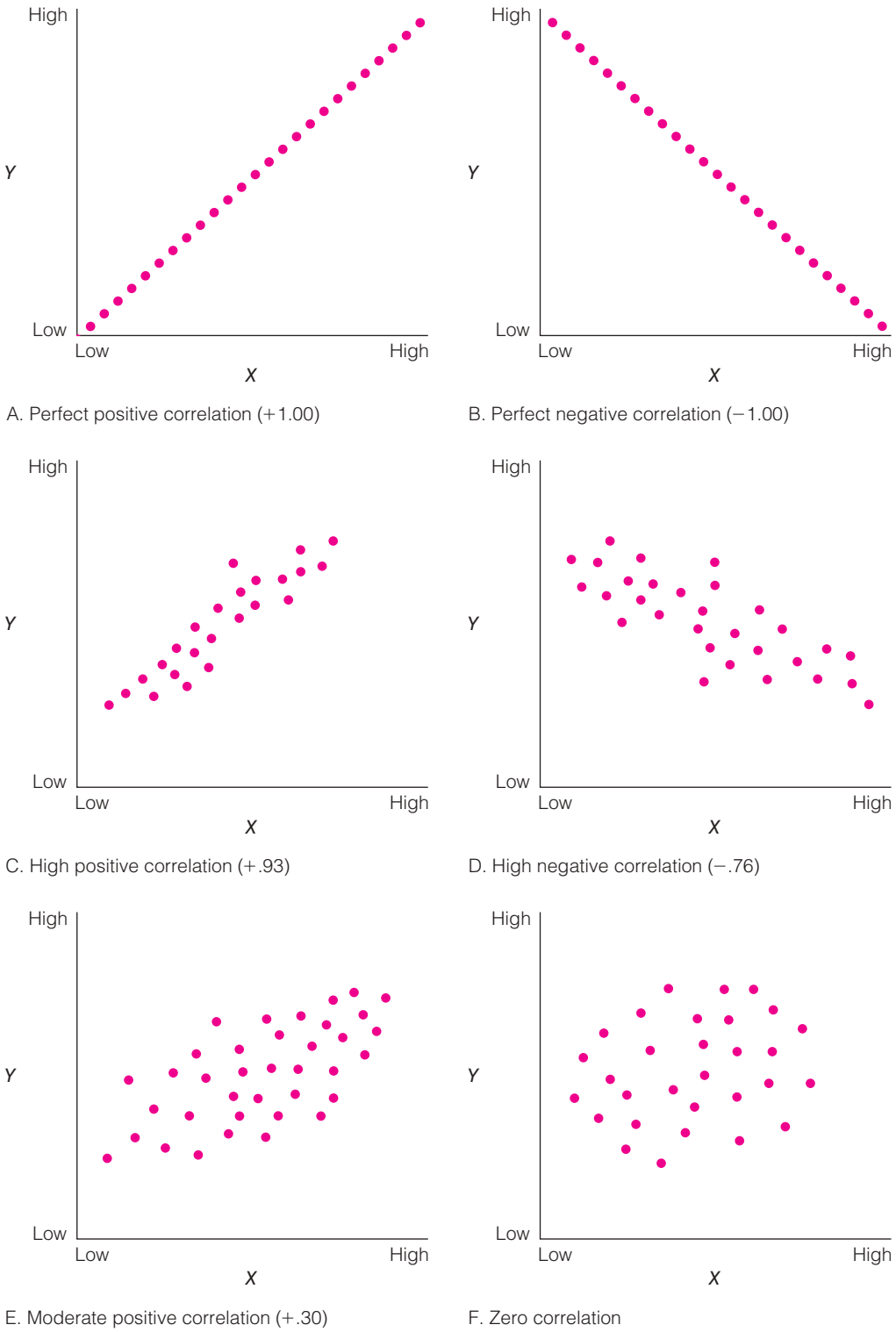


Figure 6.13 Scatterplots of Selected Values of *r*

Once you are used to correlations, you will be able to picture a scatterplot for any Pearson correlation you encounter.

Like the mean and standard deviation, the Pearson *r* is an interval statistic that can also be used with ratio data. An assumption underlying the product moment coefficient of correlation is that the relationship between the two variables (*X* and *Y*) is linear—that is, that a straight line provides a reasonable expression of the relationship of one variable to the other. If a curved line expresses this relationship, it is said to be a **curvilinear relationship**. In a curvilinear relationship, as the values of *X* increase, the values of *Y* increase up to a point, at which further increases in *X* are associated with decreases in *Y*. An example is the relationship between anxiety and performance. As individuals' anxiety level increases, so does their performance, but only up to a point. With further increases in anxiety, performance decreases. Another example is the amount of care people require during the course of a lifetime: It is high in the early and late years and usually relatively low in the middle years. A scatterplot of this relationship would produce a curve.

A practical way of determining whether the relationship between two variables is linear or curvilinear is to examine a scatterplot of the data. Figure 6.14 shows two diagrams, one of which (A) indicates a linear relationship and the other (B) a curvilinear one.

THINK ABOUT IT 6.6

After scoring test 2 and entering these exam scores in his computer, Mr. Li also entered each student's number of days absent during the second unit. He instructed the computer to calculate the Pearson *r* for these two variables (days absent and test 2 scores). The *r* was −.40. What would he conclude?

Answer

There was a moderate tendency for those with high days absent to have lower test 2 scores, and those with low days absent to have higher scores on test 2.

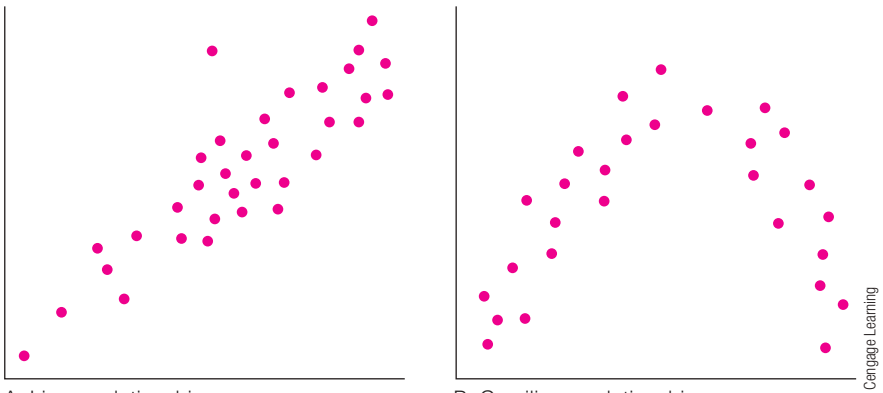


Figure 6.14 Linear and Curvilinear Relationships

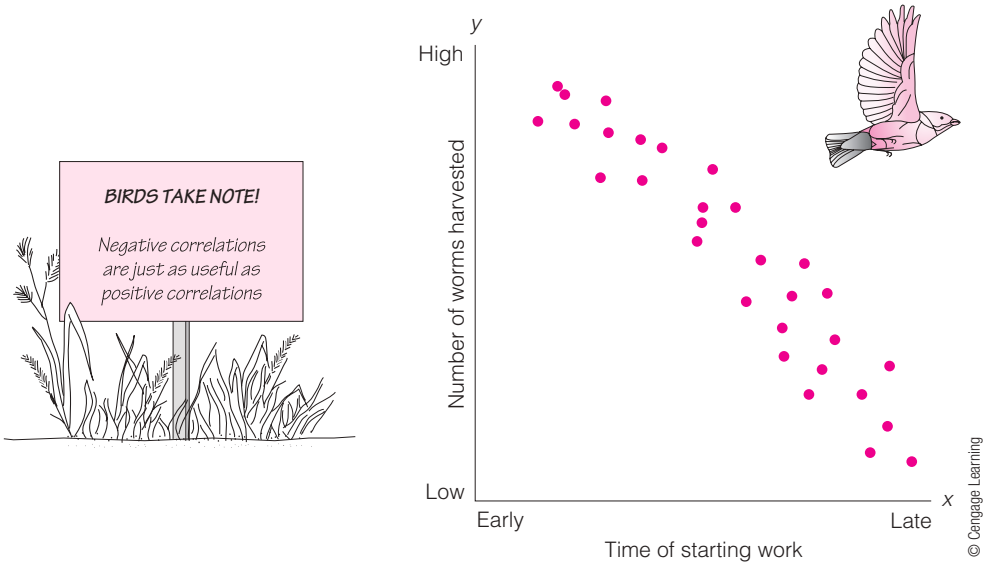


Figure 6.15 A Useful Negative Correlation

If the relationship between variables is curvilinear, the computation of the Pearson *r* will result in a misleading underestimation of the degree of relationship. In this case, another index, such as the correlation ratio (△), should be applied. A discussion of the correlation ratio is presented in Chapter 13.

THINK ABOUT IT 6.7

1. What is the best estimate of the Pearson *r* in Figure 6.15? (a) .80, (b) .60, (c) 0, (d) −.60, (e) −.90.

Answers

1. e

INTERPRETATION OF PEARSON *r*

You have seen that when two variables are highly related in a positive way, the correlation between them approaches +1.00. When they are highly related in a negative way, the correlation approaches −1.00. When there is little relation between variables, the correlation will be near 0. The Pearson *r* provides a meaningful index for indicating relationship, with the sign of the coefficient indicating the direction of the relationship, and the difference between the coefficient and 0 indicating the degree of the relationship.

However, in interpreting the correlation coefficient, keep the following points in mind:

1. *Correlation does not necessarily indicate causation*. When two variables are found to be correlated, this indicates that relative positions in one

variable are *associated* with relative positions in the other variable. It does not necessarily mean that changes in one variable are *caused* by changes in the other variable.

We may find a correlation between two variables not because there is an intrinsic relationship between these variables, but because they are both related to a third variable. For example, if we correlate the average teachers' salary for each of the past 20 years and the dollar value of coffee sold during each of these years, we derive a high correlation. This does not mean that as soon as teachers' salaries are raised, they spend the extra money on coffee. We observe a correlation between the two variables because each of them is highly correlated with a third variable, general inflation.

2. *The size of a correlation is in part a function of the variability of the two distributions to be correlated*. Restricting the range of the scores to be correlated reduces the observed degree of relationship between two variables. For example, people have observed that success in playing basketball is related to height: The taller an individual is, the more probable that he or she will do well in this sport. This statement is true about the population at large, where there is a wide range of heights. However, within a basketball team whose members are all tall, there may be little or no correlation between height and success because the range of heights is restricted. For a college that accepts students with a wide range of scores on a scholastic aptitude test, you would expect a correlation between the test scores and college grades. For a college that accepts only students with very high scholastic aptitude scores, you would expect very little correlation between the test scores and grades because of the restricted range of the test scores in this situation.

   If we correlate shoe size and reading vocabulary scores for a single grade level, we would expect a correlation of approximately zero. However, if we correlated this variable for all elementary students, we would obtain a high correlation because as children mature their feet become larger and their vocabulary increases.

3. *Correlation coefficients should not be interpreted in terms of percentage of perfect correlations*. Because correlation coefficients are expressed as decimal fractions, sometimes they are interpreted mistakenly as a percentage of perfect correlation. For example, an r of .80 does not indicate 80 percent of a perfect relationship between two variables. This interpretation is erroneous because, for example, an r of .80 does not express a relationship that is twice as great as an r of .40. A way of determining the degree to which you can predict one variable from the other is to calculate an index called the **coefficient of determination**. The coefficient of determination is the square of the correlation coefficient. It gives the percentage of variance in one variable that is associated with the variance in the other. For example, if you find a correlation of +.80 between achievement and intelligence, 64 percent of the variance in achievement is associated with variance in intelligence

test scores. Probably the best way to give meaning to the size of the correlation coefficient is to picture the degree of scatter implied by correlations of different sizes (as illustrated in Figure 6.12) and to become familiar with the size of correlations commonly observed between variables of interest.

4. *Avoid interpreting the coefficients of correlation in an absolute sense.* In interpreting the degree of correlation, keep in mind the purpose for which it is being used. For example, it may not be wise to use a correlation of .50 for predicting the future performance of an individual. However, if you could develop a measure that you could administer to high school seniors that correlated with their subsequent college freshman grade point average (GPA), your services could be in high demand because both ACT and SAT scores correlate approximately .40 with subsequent freshman GPAs. Correlations and their use in research are discussed further in Chapter 13.

EFFECT SIZE

We have seen that the Pearson *r* indicates both the direction and the strength of a relationship between variables. The Pearson *r* has universal meaning in that an *r* near +1.00 always indicates a strong positive relationship no matter the variables under consideration. An *r* near −1.00 always means a strong negative relationship and an *r* near 0 always means a weak relationship. Smith and Glass (1977) were among the first contemporary researchers to write about the concept of **effect size**, a statistic that has universal meaning to assess both the direction and the strength of a difference between two means. In this sense, an effect size subtracts the mean of the control group from the mean of the experimental group and then divides this difference by the standard deviation of the control group, as seen in Formula 6.18:

$$\triangle = \frac{\overline{X}_e - \overline{X}_c}{s_c} \tag{6.18}$$

where

$\triangle$ = effect size for a difference between means
$\overline{X}_e$ = mean of the experimental group
$\overline{X}_c$ = mean of the control group
s_c = standard deviation of the control group

In experimental studies, effect size can be used to compare the direction and the relative strength of different independent variables (i.e., intervention) on the same dependent variable.

Consider an experiment in which on-task behavior is the dependent variable. The experimental group receives contingent reinforcement, whereas the control group does not. (We explain this contrast more fully in Chapter 11.) The control group has a mean of 90 and a standard deviation of 10. The experimental group has a mean of 96. The effect size is 96 − 90/10 = .60. Consider

another experiment with the same dependent variable (i.e., on-task behavior), but with a drug treatment versus a placebo as the independent variable. The control group, which received the placebo, has a mean of 95 and a standard deviation of 8. The experimental group, which received a drug, has a mean of 97. The effect size is 97 − 95/8 .25. The evidence suggests that contingent reinforcement has had a greater effect related to on-task behavior than has the drug.

Effect sizes are interpreted in the same way that *z* scores are interpreted. Effect size can be used to compare the direction and the relative magnitude of the relationships that various independent variables have with a common dependent variable. In addition, it can be used to help decide whether the difference an independent variable makes on the dependent variable is strong enough to recommend its implementation in practice.

One approach is to ask if a given effect size is larger or smaller than effect sizes found in other studies with the same dependent variable. Also, you can assess the utility of an effect size by relating the cost in time, money, and other resources needed to implement the independent variable in relation to the importance of the dependent variable. For example, a school of nursing would be interested in a brief inexpensive procedure that produced an effect size of .20 on state nursing licensure exam scores. Effect size is a useful statistic for assessing the strength and utility of a treatment or other independent variable. Cohen (1988), with approximate verification from Lipsey and Wilson (1993) and Sawilowsky (2009), suggested the following interpretations, but as with guidelines these should be applied cautiously within the context of a research study:

An effect size of .20 is small.
An effect size of .50 is medium.
An effect size of .80 is large.

Cohen also developed an alternate index of effect size symbolized by lowercase *d*. **Cohen's *d*** is computed by subtracting the mean of the control group from the mean of the experimental group and dividing by the pooled standard deviation of the two groups:

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\Sigma x_1^2 + \Sigma x_2^2}{n_1 + n_2 - 2}}} \tag{6.19}$$

where

d = effect size
X_1 = mean of one group
X_2 = mean of the other group
Σx_1 = sum of deviation scores squared in the first group
Σx_2 = sum of deviation scores squared in the second group
n_1 = number in first group
n_2 = number in second group

For example, we compare the scores of 28 students taught method A (group 1) and 22 students taught method B (group 2) with the following statistics. The *d*

is .56 indicating that the group 1 mean is higher than the group 2 mean by an effect size of .56 or a medium effect size.

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{n_1 + n_2 - 2}}} = \frac{82 - 79}{\sqrt{\frac{1390}{28 + 22 - 2}}}$$

$$= \frac{3}{\sqrt{\frac{1390}{48}}} = \frac{3}{\sqrt{28.9583}} = \frac{3}{5.3513} = .56$$

Often Cohen's *d* is employed more readily than Smith and Glass's △ because *d* does not require designating one group as the control group in the numerator. Also, the denominator is an estimate of the population standard deviation based on the variance within both groups and the number in both groups.

As a form of *z* scores, effect sizes have universal meaning. An effect size of −.50 always means that the group 1 mean was half a standard deviation below the group 2 mean. Effect sizes are important for evaluating the results of a quantitative study that has produced statistically significant results. The concept of effect size that has been promoted in the field of educational research is now widely used in other social science disciplines. The *Publication Manual of the American Psychological Association* (2010, p. 34) asserts: "For the reader to appreciate the magnitude or importance of a study's findings, it is almost always necessary to include some measure of effect size in the Results section."

Effect size can be calculated for various statistics other than $\overline{X}_1 - \overline{X}_2$. The coefficient of determination, referred to previously, is considered an effect size for correlations. For further discussion of effect size, see Walker (2004; 2005a; 2005c), who analyzes various effect sizes encountered in sundry situations within social science research, particularly effect sizes affiliated with Cohen's *d* and its related indices, and also r-based functions of variance accounted for effect sizes. Additionally, Kline (2004), Levin and Robinson (2000), McGrath and Meyer (2006), Thompson (2002), and Vacha-Haase and Thompson (2004) all have extended the discussion and potential possibilities via the use and analysis of effect sizes with statistically significant results. Lastly, Onwuegbuzie (2003) has proposed ways to apply effect size in qualitative research. Thus, the reporting of effect sizes is becoming more frequent in education and other social sciences fields.

A note of caution: Effect size is independent of sample size. Therefore, large effect sizes can easily be observed through chance alone with very small samples. A rule of thumb is that samples of less than 30 are considered small. (Most of our examples are less than 30, so you are not bogged down by the math.) In Chapter 7, we present ways of taking into account size of sample as well as effect size in evaluating results.

META-ANALYSIS

Smith and Glass (1977) developed the concept of **meta-analysis**, which is a statistical technique that combines the effect sizes reported in the results of studies with the same (or similar) independent and dependent variables. In 1981, Glass, McGaw, and Smith wrote *Meta-Analysis in Social Research*, which is

considered the classic work on meta-analysis. The result of a meta-analysis provides an overall summary of the outcomes of a number of studies by calculating a weighted average of their effect sizes. Meta-analysis gives a better estimate of the relationship among variables than do single studies alone. It is important that the group of studies included in a meta-analysis focus on the same hypothesis or research questions with the same variables. You would not conduct a meta-analysis on school achievement in general, but rather would focus on a particular area such as the effect of specific science teaching strategies on student achievement in high school chemistry. The average effect size as a formula is

$$\overline{\triangle} = \frac{\triangle_1 n_1 + \triangle_1 n_2 + \cdots + \triangle_k n_k}{N} \tag{6.20}$$

where

$\overline{\triangle}$ = average effect size
$\triangle_1$ = effect size for group 1
$\triangle_k$ = effect size of the last group
n_1 = number in first group
n_k = number in last group
N = total number of subjects

Suppose we have four studies investigating the effect of phonics instruction on reading proficiency. Their statistics are as follows:

Study 1: effect size = .90, n = 60
Study 2: effect size = .40, n = 40
Study 3: effect size = −.20, n = 30
Study 4: effect size = .10, n = 70

The average effect size is

$$\overline{\triangle} = .9(60) + .4(40) + -.2(30) + .1(70) = \frac{54 + 16 + -6 + 7}{200} = \frac{71}{200} = .36$$

This is approximately halfway between what Cohen describes as a small effect size and a medium effect size. Another way of looking at it is to use the normal curve. Consulting Table A.1 in the Appendix, you see that a z score of .36 has a percentile rank of 64. The mean in a treatment group is equivalent to a score with a percentile rank of approximately 64 in the control group. If the treatment is relatively inexpensive, you would be inclined to recommend it in practice. If the treatment is expensive and/or the dependent variable is relatively unimportant, you would not be inclined to recommend it. The following is an abstract of a meta-analysis by Graham and Perin (2007):

> There is considerable concern that the majority of adolescents do not develop the competence in writing they need to be successful in school, the workplace, or their personal lives. A common explanation for why youngsters do not write well is that schools do not do a good job of teaching this complex skill. In an effort to identify effective instructional practices for teaching writing to adolescents, the authors conducted a meta-analysis of the writing intervention literature (Grades 4–12), focusing their efforts on experimental and quasi-experimental

studies. They located 123 studies that yielded 154 effect sizes for quality of writing. The authors calculated an average weighted effect size (presented in parentheses) for the following 11 interventions: strategy instruction (.82), summarization (.82), peer assistance (.75), setting product goals (.70), word processing (.55), sentence combining (.50), inquiry (.32), prewriting activities (.32), process writing approach (.32), study of models (.25), grammar instruction (–.32).

Based on the average weighted effect size of the interventions, Graham and Perin (2007, p. 445) made a number of recommendations:

1. Teach adolescents strategies for planning, revision, and editing their compositions.
2. Teach strategies and procedures for summarizing reading material.
3. Develop instructional arrangements in which adolescents work together to plan, draft, revise, and edit their work.
4. Set clear and specific goals for what students are to accomplish with their writing.
5. Have students use word processing as a primary tool for writing.
6. Teach them how to combine sentences into increasingly more complex sentences.
7. Provide teachers with professional development in how to use the process writing approach.
8. Involve students with writing activities designed to sharpen their skills of inquiry.
9. Engage students in prewriting activities that help them to gather and organize ideas.
10. Provide students with good models for each type of writing that is the focus of instruction.

Meta-analysis is also used to integrate the findings of nonexperimental studies. For example, you might conduct a meta-analysis of studies that have examined gender differences in mathematics performance on standardized tests, or a meta-analysis might involve studies investigating the correlation of certain noncognitive variables to achievement in graduate education courses.

Meta-analysis has sometimes been criticized for including the results of poorly designed and conducted studies along with the results of more credible studies (criteria for evaluating research designs are presented in Chapter 10). This problem can be resolved by first calculating the average effect size of all studies and then calculating the average effect size of well-designed studies to determine if the latter agree with the former. Qin, Johnson, and Johnson (1995) did this in a meta-analysis of studies comparing problem-solving performance of subjects under cooperative versus competitive conditions. In 55 cases, cooperation outperformed competition; whereas in 8, competition outperformed cooperation. The average effect size was .55 in favor of cooperation. The average effect size of the 33 studies judged as being of high methodological quality was .68.

The statistical computations involved in a meta-analysis are beyond the scope of this text. A number of software programs are available for performing the two main computations: (1) calculating the effect size estimates and (2) analyzing the estimates obtained. The reader is referred to Cummings (2012), Hedges (1998), Hunter and Schmidt (2001), Lipsey and Wilson (2000), Rosenthal (1991), and Walker (2003) for further discussion of meta-analysis approaches and applications.

SUMMARY

Descriptive statistics serve to describe and summarize observations. The descriptive technique to be employed is selected according to the purpose the statistic is to serve and the scale of measurement used.

Scales of measurement are means of quantifying observations. There are four types:

1. Nominal scales classify observations into mutually exclusive categories.
2. Ordinal scales sort objects or classes of objects on the basis of their relative standing.
3. Interval scales use equal intervals for measurement and indicate the degree to which a person or an object possesses a certain quality.
4. Ratio scales use equal intervals for measurement and measure from an absolute zero point.

Once observations are quantified, the data can be arranged into frequency distributions and shown graphically.

Measures of central tendency—the mode, the median, and the mean—provide a single index to represent the average value of a whole set of measures. The mode, which is a nominal statistic, is the least stable and least useful measure in educational research. The median is an ordinal statistic that takes into account the ranks of scores within a distribution but not the size of the individual scores. The mean, which is an interval (or ratio) statistic, is the most stable and most widely used index of central tendency. Another way of describing observations is to indicate the variation, or spread, of the values within a distribution. The range, the variance, and the standard deviation are three indexes used for this purpose. The range, a nominal statistic, is the distance between the highest and the lowest values in a distribution, plus 1. Variance is the mean of the squared deviations of scores from the mean. It is an interval (or ratio) statistic. Standard deviation—the square root of the variance—is the most widely used index of variability.

Standard scores are used to indicate the position of a single score in a distribution. The most widely used is the *z* score, which converts values into standard deviation units. The *z* scores are often converted into stanines, *T* scores, or other standard scores. An ordinal index of location shows a score's position in percentile rank, which indicates what percentage of scores fall equal to or below the midpoint of the score's interval. Using the characteristics and the areas of the normal curve, you can approximate the percentage of cases below and above each *z* score in a normal distribution.

Correlation techniques enable researchers to describe the relationship between two sets of measures. Product moment correlation (Pearson *r*), the most widely used index of relationships, is used with interval or ratio data. Table 6.7 summarizes correlation indexes appropriate for interval, ordinal, and nominal data.

Effect size—the difference between the means of the experimental and control groups divided by the standard deviation of the control group—is a useful measure of the strength or magnitude of their relationship.

Table 6.7 Summary of Descriptive Statistics Presented in This Chapter

	Nominal	Ordinal	Interval
Indexes of central tendency	Mode	Median	Mean
Indexes of variability	Range		Variance and standard deviation
Indexes of location	Label or classification	Percentile rank	*z score, T scores, and other standard scores*
Correlation indexes			Pearson *r*

Meta-analysis is a widely used method for combining the statistical results of a group of studies on the same problem. It enables researchers to succinctly summarize results of many studies on a particular question. The most widely used index is the average effect size.

KEY CONCEPTS

coefficient of determination
Cohen's *d*
correlation
correlation coefficient
curvilinear relationship
descriptive statistics
deviation scores
effect size
frequency distribution
frequency polygon
histogram
inferential statistics
interval scale
linear relationship
mean
measures of central tendency
median
meta-analysis
mode
negative correlation
negatively skewed distribution
nominal scale
normal curve
normal distribution
ordinal scale
Pearson product moment correlation coefficient
percentile rank
positive correlation
positively skewed distribution
range
scatterplot
skewed distribution ratio scale
standard deviation scatterplot
standard score
stanine score
symmetrical distribution
T score
variability
variance
z score

EXERCISES

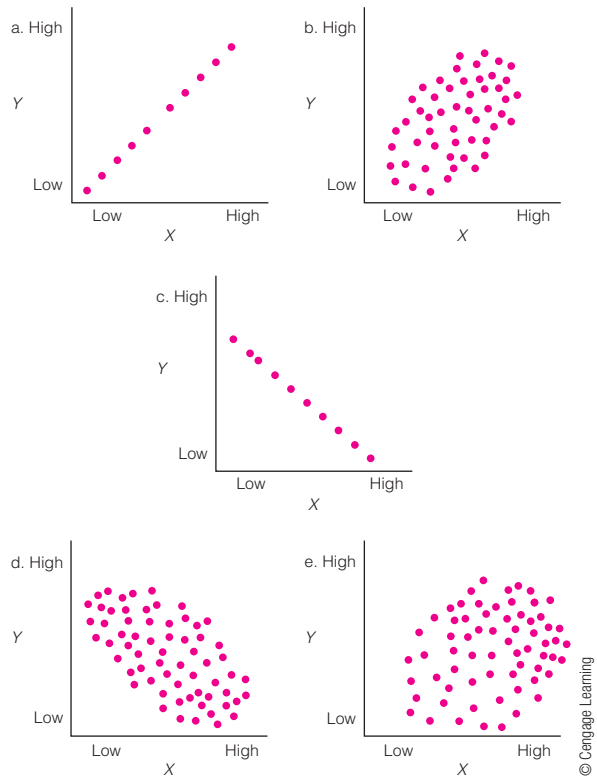
1. Identify the type of measurement scale—nominal, ordinal, interval, or ratio—suggested by each statement:
   a. Lili finished the math test in 35 minutes, whereas Tsavo finished the same test in 25 minutes.
   b. Tsavo speaks French, but Lili does not.
   c. Tsavo is taller than Lili.
   d. Lili is 6 feet 2 inches tall.
   e. Lili's IQ is 120, whereas Tsavo's IQ is 110.

2. Draw a histogram and a frequency polygon for the following frequency distribution:

x	*f*	*x*	*f*	*x*	*f*	*x*	*f*
80	1	76	6	73	20	70	7
79	2	75	15	72	17	69	3
78	3	74	22	71	9		
77	10						

3. Provide answers as requested, given the following distribution: 15, 14, 14, 13, 11, 10, 10, 10, 8, 5.
   a. Calculate the mean
   b. Determine the value of the median
   c. Determine the value of the mode

4. Briefly explain the relationship between the skewness of a distribution of scores and the resulting values of the mean, median, and mode.
5. Identify the measure—mode, mean, or median—that best suits each type of scale:
   a. Ordinal
   b. Nominal
   c. Interval
6. Identify the measure—mode, mean, or median—that each term defines:
   a. The middle score
   b. The arithmetic average
   c. The most frequently occurring score
7. Discuss the advantages and disadvantages of range and standard deviation as measures of variability of the scores.
8. a. Calculate the z score for a score of 5 in a distribution with a mean of 7 and a standard deviation of 0.50.
   b. What would be the stanine score for this score?
9. Using Table A.1, what is the estimated percentile rank for a z score of +1.20?
10. Why do you think the U.S. Census Bureau reports median income instead of mean income?
11. On an analysis of achievement test, male scores were more heterogeneous, more spread from low to high, than female scores.
    a. Which statistics would be greater for males?
    b. Which gender had more stanine scores of 9?
    c. Which gender had more stanine scores of 1?
    d. Which gender had more stanine scores of 5?
12. The mean score on a test is 40, and the standard deviation is 4. Express each of the following raw scores as a z score:
    a. 41
    b. 30
    c. 48
    d. 36
    e. 46
13. a. What would be the T score for the raw score of 46 in Exercise 12?
    b. What would be the stanine score for the raw score of 46?
14. In a normal distribution, what percentage of the scores would fall below a z score of −1? A z score of 0? A z score of +.67?
15. Describe the relationship shown by these scatterplots. Then estimate the correlation coefficients.



16. Each dot in a scatterplot represents _____.
17. Five girls took a history test and a geography test, with the following results:

	History		Geography	
	Raw Score	*z Score*	*Raw Score*	*z Score*
Ann	28	.5	85	1.50
Nesa	32	1.5	65	.50
Nanner	26	0	55	0
Benazir	20	−1.5	45	−0.50
Yoko	24	−.5	25	−1.50

History	Geography
$\Sigma = 310$	$\Sigma X = 275$
$\sigma = 4$	$\sigma = 20$

a. What is the mean of the history test?
b. Whose performance in history is most in agreement with her performance in geography?
c. What is the correlation between the history and geography scores?

18. Given that the history test mean in Exercise 17 is lower than the geography test mean, which of the following conclusions would be correct?
a. These girls are better in history than in geography.
b. These girls are better in geography than in history.
c. Their teacher has probably spent twice as much time on geography as on history.
d. Their teacher knows more geography than history.
e. None of the above.

19. If the coefficient of correlation between variable X and variable Y is found to be $-.98$, which of the following would be indicated?
a. Variable X and variable Y are closely related.
b. Variable X and variable Y are unrelated.
c. Variable X and variable Y are perfectly related.
d. Variable Y is a result of variable X.

20. For each of the following cases, indicate which statistic should be used—mean, standard deviation, z score, or Pearson r.
a. We want to know the extent to which the scores of a class are spread out or heterogeneous.
b. We want to determine how Zumbiwa's score compares with the scores of the rest of the class.
c. We want to know how well the class as a whole did on an examination.
d. We want to predict the future achievement of students from their IQ.

21. Interpret the following: "On the College Board exam ($\Sigma = 500$, $\sigma = 100$), the mean of this year's Central High School seniors was 490, and the standard deviation was 110. The correlation between the exam scores and high school grade point average was $+.40$."

22. Interpret the following: "Parents of Central High students were asked to rank 10 problems from 10 = most serious to 1 = less serious. The median for the problem 'cyberbullying' was 5.21."

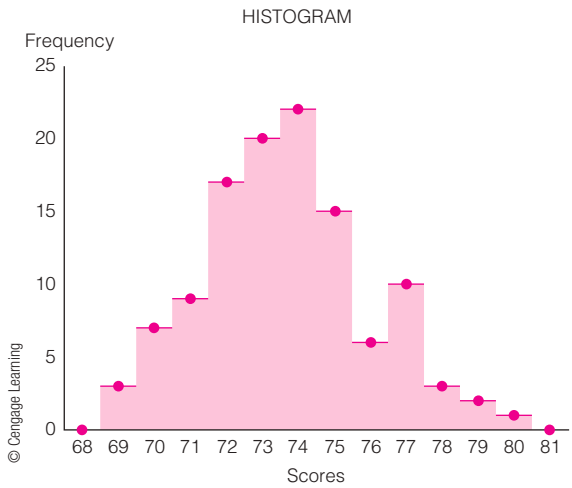
23. Smith and Glass's Δ and Cohen's d are two ways of defining what?

24. Define *effect size* and explain how it is used.

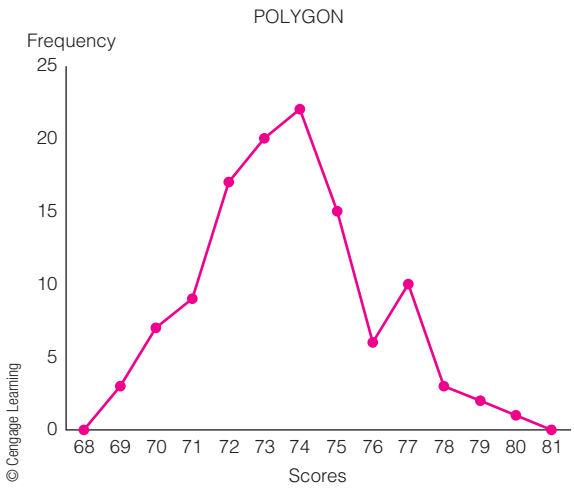
25. What is the purpose of meta-analysis?

ANSWERS

1. a. Ratio
b. Nominal
c. Ordinal
d. Ratio
e. Interval



2. Figures may be expanded or contracted vertically or horizontally, or both, and be correct if the relationships between scores and frequencies are maintained.



3. **a.** Mean = 11
   b. Median = 10.50
   c. Mode = 10
4. The three measures are not equal in a skewed distribution. The mean is pulled in the direction of the skewed side. Thus, in a positively skewed distribution the mean is always higher than the median, and the mode is usually lowest in value. In a negatively skewed distribution, the mean is always lower than the median, and the mode is usually highest in value.
5. **a.** Median
   b. Mode
   c. Mean
6. **a.** Median
   b. Mean
   c. Mode
7. The range is easy to calculate and to explain. The standard deviation takes into account all the scores and is more stable.
8. **a.** $z = (5 - 7)/.5 = .40$
   b. Stanine = 1
9. 88 (rounded from 88.49)
10. Because the median is more typical of the incomes in the United States. It is not influenced by the extreme salaries of the billionaires and millionaires.
11. **a.** Male scores had higher variance, standard deviation, and range.
    b. Males
    c. Males
    d. Females
12. **a.** .25
    b. −2.5
    c. 2
    d. −1
    e. 1.50
13. **a.** $T = 10z + 50 = 10(1.5) + 50 = 65$
    b. Stanine score $2z + 5$ rounded = $(2)(1.5) + 5 = 8$
14. 16 percent; 50 percent; 75 percent
15. **a.** Perfect positive, +1 correlation
    b. Positive, +75
    c. Perfect negative, –1
    d. Negative, −.75
    e. No correlation, 0
16. An individual's score on two dimensions or other paired *z* scores
17. **a.** 26
    b. Nanner; she had the same *z* score on both test.
    c. $r = \frac{\Sigma z_x z_y}{N} = \frac{3}{5} = .6$
18. e
19. a
20. **a.** Standard deviation
    b. *z* score
    c. Mean
    d. Pearson *r*
21. As a group, the Central High seniors were slightly below the national average. Their scores were slightly more heterogeneous than usual. Those with high scores tended to have high GPAs. Those with low scores tended to have low GPAs. The relationship between scores and GPAs was moderate positive.
22. Parents ranked "cyberbullying" about average.
23. Effect size
24. Effect size (a form of *z* score) is the difference between experimental and control groups divided by the standard deviation of the control group. It indicates the strength of the relationship between the independent and dependent variables.
25. It combines the result of studies with similar independent and dependent variables to produce an average effect size, a mathematical summary of the results.

REFERENCES

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cummings, G. (2012). *Understanding the new statistics: Effect sizes, confidence Intervals, and meta-analysis*. New York: Routledge.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*(3), 445–476.

Hedges, L. V. (1998). *Statistical methods for meta-analysis*. San Diego: Academic Press.

Hunter, J. E., & Schmidt, F. C. (2001). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Kline, R. B. (2004). *Beyond significance testing; Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Levin, J. R., & Robinson, D. H. (2000). Statistical hypothesis testing, effect size estimation, and the conclusion coherence of primary research studies. *Educational Researcher, 29*(1), 34–36.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181–1209

Lipsey, M. W., & Wilson, D. B. (2000). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods, 11*, 386–401.

Onwuegbuzie, A. D. (2003). Effect sizes in qualitative research: A prolegomenon. *Quality and Quantity*, *37*, 393–409.

Qin, Z., Johnson, D. W., & Johnson, R. T. (1995). Cooperative versus competitive efforts and problem solving. *Review of Educational Research*, *65*(2), 129–143.

Rosenthal, R. (1991). (Series Ed.), *Meta-analytic procedures for social research*. Newbury Park, CA: Sage Publications.

Sawilowsky, S. S. (2009). Very large and huge effect sizes. *Journal of Modern Applied Statistical Methods, 8*, 597–599.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752–760.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (p. 1). New York: Wiley.

Supiano, B. (September 2, 2011). In lifetime earnings, education matters, but so do occupation, gender, and race. *Chronicle of Higher Education*, LVIII (2), p. A28.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25–32.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, *51*(4), 473–481.

Walker, D. A. (2003). Converting Kendall's tau for correlational or meta-analytic analyses. *Journal of Modern Applied Statistical Methods, 2*(2), 525–530.

Walker, D. A. (2004). The importance of drawing meaningful conclusions from data: A review of the literature with meta-analytic inquiry. *NASPA Journal, 41*(3), 452–469.

Walker, D. A. (2005a). Bias affiliated with two variants of Cohen's d when deter mining U_1 as a measure of the percent of non-overlap. *Journal of Modern Applied Statistical Methods, 4*(1), 100–105.

Walker, D. A. (2005b). A graph is worth a thousand words? The publication rate of graphs in the *JCSD*, 1999 to 2004. *Journal of College Student Development, 46*(6), 689–698.

Walker, D. A. (2005c). An SPSS matrix for determining effect sizes from three categories: *r* and functions of *r*, differences between proportions, and standardized differences between means. *Journal of Modern Applied Statistical Methods, 4*(1), 333–342.