SEVENTH EDITION

Psychological Testing

Principles, Applications, and Issues

Robert M. Kaplan
University of California, Los Angeles

Dennis P. Saccuzzo
San Diego State University

WADSWORTH
CENGAGE Learning™

Australia • Brazil • Japan • Korea • Mexico • Singapore • Spain • United Kingdom • United States

WADSWORTH
CENGAGE Learning

Psychological Testing: Principles, Applications, and Issues, Seventh Edition

Robert M. Kaplan, Dennis P. Saccuzzo

Editor: Jaime Perkins

Editorial Assistant: Wilson Co

Technology Project Manager: Amy Cohen

Marketing Manager: Kim Russell

Marketing Assistant: Molly Felz

Marketing Communications Manager: Talia Wise

Project Manager, Editorial Production: Charlene M. Carpentier

Creative Director: Rob Hugel

Art Director: Vernon Boes

Print Buyer: Linda Hsu

Permissions Editor: Bob Kauser

Production Service: Newgen–Austin

Text Designer: Lisa Henry

Photo Researcher: Darren Wright

Copy Editor: Mary Ann Grobbel

Cover Designer: Larry Didona

Cover Image: "Geometric shapes below clouds" ©Pete Turner/The Image Bank/Getty Images

Compositor: Newgen

Library of Congress Control Number: 2008927883
Student Edition:
ISBN-13: 978-0-495-09555-2
ISBN-10: 0-495-09555-9

Wadsworth
10 Davis Drive
Belmont, CA 94002-3098
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at **international.cengage.com/region.**

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

For your course and learning solutions, visit **academic.cengage.com.**

Purchase any of our products at your local college store or at our preferred online store **www.ichapters.com.**

Printed in the United States of America
2 3 4 5 6 7 12 11 10 09

CHAPTER 2

Norms and Basic Statistics for Testing

LEARNING OBJECTIVES

When you have completed this chapter, you should be able to:

- Discuss three properties of scales of measurement
- Determine why properties of scales are important in the field of measurement
- Identify methods for displaying distributions of scores
- Calculate the mean and the standard deviation for a set of scores
- Define a *Z* score and explain how it is used
- Relate the concepts of mean, standard deviation, and *Z* score to the concept of a standard normal distribution
- Define quartiles, deciles, and stanines and explain how they are used
- Tell how norms are created
- Relate the notion of tracking to the establishment of norms

We all use numbers as a basic way of communicating: Our money system requires us to understand and manipulate numbers, we estimate how long it will take to do things, we count, we express evaluations on scales, and so on. Think about how many times you use numbers in an average day. There is no way to avoid them.

One advantage of number systems is that they allow us to manipulate information. Through sets of well-defined rules, we can use numbers to learn more about the world. *Tests* are devices used to translate observations into numbers. Because the outcome of a test is almost always represented as a score, much of this book is about what scores mean. This chapter reviews some of the basic rules used to evaluate number systems. These rules and number systems are the psychologist's partners in learning about human behavior.

If you have had a course in psychological statistics, then this chapter will reinforce the basic concepts you have already learned. If you need additional review, reread your introductory statistics book. Most such books cover the information in this chapter. If you have not had a course in statistics, then this chapter will provide some of the information needed for understanding other chapters in this book.

WHY WE NEED STATISTICS

Through its commitment to the scientific method, modern psychology has advanced beyond centuries of speculation about human nature. Scientific study requires systematic observations and an estimation of the extent to which observations could have been influenced by chance alone (Salkind, 2007). Statistical methods serve two important purposes in the quest for scientific understanding.

First, statistics are used for purposes of description. Numbers provide convenient summaries and allow us to evaluate some observations relative to others (Cohen & Lea, 2004; Pagano, 2004; Thompson, 2006). For example, if you get a score of 54 on a psychology examination, you probably want to know what the 54 means. Is it lower than the average score, or is it about the same? Knowing the answer can make the feedback you get from your examination more meaningful. If you discover that the 54 puts you in the top 5% of the class, then you might assume you have a good chance for an A. If it puts you in the bottom 5%, then you will feel differently.

Second, we can use statistics to make **inferences**, which are logical deductions about events that cannot be observed directly. For example, you do not know how many people watched a particular television movie unless you ask everyone. However, by using scientific sample surveys, you can infer the percentage of people who saw the film. Data gathering and analysis might be considered analogous to criminal investigation and prosecution (Cox, 2006; Regenwetter, 2006; Tukey, 1977). First comes the detective work of gathering and displaying clues, or what the statistician John Tukey calls *exploratory data analysis*. Then comes a period of *confirmatory data analysis*, when the clues are evaluated against rigid statistical rules. This latter phase is like the work done by judges and juries.

Some students have an aversion to numbers and anything mathematical. If you find yourself among them, you are not alone. Not only students but also professional psychologists can feel uneasy about statistics. However, statistics and the basic

principles of measurement lie at the center of the modern science of psychology. Scientific statements are usually based on careful study, and such systematic study requires some numerical analysis.

This chapter reviews both descriptive and inferential statistics. **Descriptive statistics** are methods used to provide a concise description of a collection of quantitative information. **Inferential statistics** are methods used to make inferences from observations of a small group of people known as a *sample* to a larger group of individuals known as a *population*. Typically, the psychologist wants to make statements about the larger group but cannot possibly make all the necessary observations. Instead, he or she observes a relatively small group of subjects (sample) and uses inferential statistics to estimate the characteristics of the larger group (Salkind, 2007).

SCALES OF MEASUREMENT

One may define *measurement* as the application of rules for assigning numbers to objects. The rules are the specific procedures used to transform qualities of attributes into numbers (Camilli, Cizek, & Lugg, 2001; Nunnally & Bernstein, 1994; Yanai, 2003). For example, to rate the quality of wines, wine tasters must use a specific set of rules. They might rate the wine on a 10-point scale where 1 means extremely bad and 10 means extremely good. For a taster to assign the numbers, the system of rules must be clearly defined. The basic feature of these types of systems is the scale of measurement. For example, to measure the height of your classmates, you might use the scale of inches; to measure their weight, you might use the scale of pounds.

There are numerous systems by which we assign numbers in psychology. Indeed, the study of measurement systems is what this book is about. Before we consider any specific scale of measurement, however, we should consider the general properties of measurement scales.

Properties of Scales

Three important properties make scales of measurement different from one another: magnitude, equal intervals, and an absolute 0.

Magnitude

Magnitude is the property of "moreness." A scale has the property of magnitude if we can say that a particular instance of the attribute represents more, less, or equal amounts of the given quantity than does another instance (Aron & Aron, 2003; Hurlburt, 2003; McCall, 2001; Howell, 2008). On a scale of height, for example, if we can say that John is taller than Fred, then the scale has the property of magnitude. A scale that does not have this property arises, for example, when a gym coach assigns identification numbers to teams in a league (team 1, team 2, and so forth). Because the numbers only label the teams, they do not have the property of magnitude. If the coach were to rank the teams by the number of games they have won, then the new numbering system (games won) would have the property of magnitude.

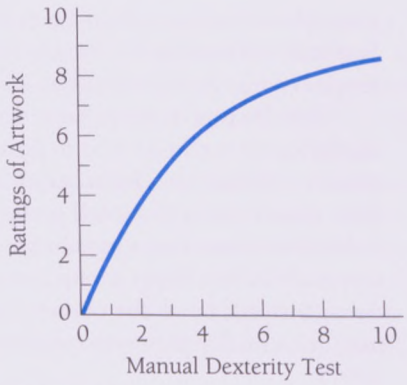


FIGURE **2.1** Hypothetical relationship between ratings of artwork and manual dexterity. In some ranges of the scale, the relationship is more direct than it is in others.

Equal Intervals

The concept of equal intervals is a little more complex than that of magnitude. A scale has the property of equal intervals if the difference between two points at any place on the scale has the same meaning as the difference between two other points that differ by the same number of scale units. For example, the difference between inch 2 and inch 4 on a ruler represents the same quantity as the difference between inch 10 and inch 12: exactly 2 inches.

As simple as this concept seems, a psychological test rarely has the property of equal intervals. For example, the difference between Intelligence Quotients (IQs) of 45 and 50 does not mean the same thing as the difference between IQs of 105 and 110. Although each of these differences is 5 points ($50 - 45 = 5$ and $110 - 105 = 5$), the 5 points at the first level do not mean the same thing as 5 points at the second. We know that IQ predicts classroom performance. However, the difference in classroom performance associated with differences between IQ scores of 45 and 50 is not the same as the differences in classroom performance associated with IQ score differences of 105 and 110. In later chapters we will discuss this problem in more detail.

When a scale has the property of *equal intervals*, the relationship between the measured units and some outcome can be described by a straight line or a linear equation in the form $Y = a + bX$. This equation shows that an increase in equal units on a given scale reflects equal increases in the meaningful correlates of units. For example, Figure 2.1 shows the hypothetical relationship between scores on a test of manual dexterity and ratings of artwork. Notice that the relationship is not a straight line. By examining the points on the figure, you can see that at first the relationship is nearly linear: Increases in manual dexterity are associated with increases in ratings of artwork. Then the relationship becomes nonlinear. The figure shows that after a manual dexterity score of approximately 5, increases in dexterity produce relatively smaller increases in quality of artwork.

TABLE **2.1**
Scales of Measurement and Their Properties

	Property		
Type of scale	**Magnitude**	**Equal intervals**	**Absolute 0**
Nominal	No	No	No
Ordinal	Yes	No	No
Interval	Yes	Yes	No
Ratio	Yes	Yes	Yes

Absolute 0

An absolute 0 is obtained when nothing of the property being measured exists. For example, if you are measuring heart rate and observe that your patient has a rate of 0 and has died, then you would conclude that there is no heart rate at all. For many psychological qualities, it is extremely difficult, if not impossible, to define an absolute 0 point. For example, if one measures shyness on a scale from 0 through 10, then it is hard to define what it means for a person to have absolutely no shyness (McCall, 2001).

Types of Scales

Table 2.1 defines four scales of measurement based on the properties we have just discussed. You can see that a nominal scale does not have the property of magnitude, equal intervals, or an absolute 0. **Nominal scales** are really not scales at all; their only purpose is to name objects. For example, the numbers on the backs of football players' uniforms are nominal. Nominal scales are used when the information is qualitative rather than quantitative. Social science researchers commonly label groups in sample surveys with numbers (such as 1 = African American, 2 = white, and 3 = Mexican American). When these numbers have been attached to categories, most statistical procedures are not meaningful. On the scale for ethnic groups, for instance, what would a mean of 1.87 signify? This is not to say that the sophisticated statistical analysis of nominal data is impossible. Indeed, several new and exciting developments in data analysis allow extensive and detailed use of nominal data (Chen, 2002; Miller, Scurfield, Drga, Galvin, & Whitmore, 2002; Stout, 2002).

A scale with the property of magnitude but not equal intervals or an absolute 0 is an **ordinal scale**. This scale allows you to rank individuals or objects but not to say anything about the meaning of the differences between the ranks. If you were to rank the members of your class by height, then you would have an ordinal scale. For example, if Fred was the tallest, Susan the second tallest, and George the third tallest, you would assign them the ranks 1, 2, and 3, respectively. You would not give any consideration to the fact that Fred is 8 inches taller than Susan, but Susan is only 2 inches taller than George.

For most problems in psychology, the precision to measure the exact differences between intervals does not exist. So, most often one must use ordinal scales of

measurement. For example, IQ tests do not have the property of equal intervals or an absolute 0, but they do have the property of magnitude. If they had the property of equal intervals, then the difference between an IQ of 70 and one of 90 should have the same meaning as the difference between an IQ of 125 and one of 145. Because it does not, the scale can only be considered ordinal. Furthermore, there is no point on the scale that represents no intelligence at all—that is, the scale does not have an absolute 0.

When a scale has the properties of magnitude and equal intervals but not absolute 0, we refer to it as an **interval scale**. The most common example of an interval scale is the measurement of temperature in degrees Fahrenheit. This temperature scale clearly has the property of magnitude, because 35°F is warmer than 32°F, 65°F is warmer than 64°F, and so on. Also, the difference between 90°F and 80°F is equal to a similar difference of 10° at any point on the scale. However, on the Fahrenheit scale, temperature does not have the property of absolute 0. If it did, then the 0 point would be more meaningful. As it is, 0 on the Fahrenheit scale does not have a particular meaning. Water freezes at 32°F and boils at 212°F. Because the scale does not have an absolute 0, we cannot make statements in terms of ratios. A temperature of 22°F is not twice as hot as 11°F, and 70°F is not twice as hot as 35°F.

The Celsius scale of temperature is also an interval rather than a ratio scale. Although 0 represents freezing on the Celsius scale, it is not an absolute 0. Remember that an absolute 0 is a point at which nothing of the property being measured exists. Even on the Celsius scale of temperature, there is still plenty of room on the thermometer below 0. When the temperature goes below freezing, some aspect of heat is still being measured.

A scale that has all three properties (magnitude, equal intervals, and an absolute 0) is called a **ratio scale**. To continue our example, a ratio scale of temperature would have the properties of the Fahrenheit and Celsius scales but also include a meaningful 0 point. There is a point at which all molecular activity ceases, a point of absolute 0 on a temperature scale. Because the Kelvin scale is based on the absolute 0 point, it is a ratio scale: 22°K is twice as cold as 44°K. Examples of ratio scales also appear in the numbers we see on a regular basis. For example, consider the number of yards gained by running backs on football teams. Zero yards actually means that the player has gained no yards at all. If one player has gained 1000 yards and another has gained only 500, then we can say that the first athlete has gained twice as many yards as the second.

Another example is the speed of travel. For instance, 0 miles per hour (mph) is the point at which there is no speed at all. If you are driving onto a highway at 30 mph and increase your speed to 60 when you merge, then you have doubled your speed.

Permissible Operations

Level of measurement is important because it defines which mathematical operations we can apply to numerical data. For nominal data, each observation can be placed in only one mutually exclusive category. For example, you are a member of only one gender. One can use nominal data to create frequency distributions (see the next section), but no mathematical manipulations of the data are permissible. Ordinal measurements can be manipulated using arithmetic; however, the result is

often difficult to interpret because it reflects neither the magnitudes of the manipulated observations nor the true amounts of the property that have been measured. For example, if the heights of 15 children are rank ordered, knowing a given child's rank does not reveal how tall he or she stands. Averages of these ranks are equally uninformative about height.

With interval data, one can apply any arithmetic operation to the differences between scores. The results can be interpreted in relation to the magnitudes of the underlying property. However, interval data cannot be used to make statements about ratios. For example, if IQ is measured on an interval scale, one cannot say that an IQ of 160 is twice as high as an IQ of 80. This mathematical operation is reserved for ratio scales, for which any mathematical operation is permissible.

FREQUENCY DISTRIBUTIONS

A single test score means more if one relates it to other test scores. A *distribution* of scores summarizes the scores for a group of individuals. In testing, there are many ways to record a distribution of scores.

The **frequency distribution** displays scores on a variable or a measure to reflect how frequently each value was obtained. With a frequency distribution, one defines all the possible scores and determines how many people obtained each of those scores. Usually, scores are arranged on the horizontal axis from the lowest to the highest value. The vertical axis reflects how many times each of the values on the horizontal axis was observed. For most distributions of test scores, the frequency distribution is bell-shaped, with the greatest frequency of scores toward the center of the distribution and decreasing scores as the values become greater or less than the value in the center of the distribution.

Figure 2.2 shows a frequency distribution of 1000 observations that takes on values between 61 and 90. Notice that the most frequent observations fall toward

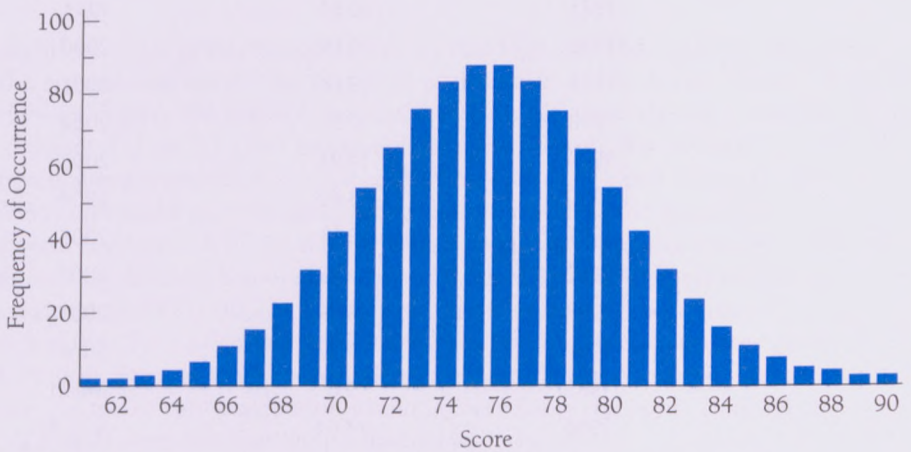


FIGURE **2.2** Frequency distribution approximating a normal distribution of 1000 observations.

the center of the distribution, around 75 and 76. As you look toward the extremes of the distribution, you will find a systematic decline in the frequency with which the scores occur. For example, the score of 71 is observed less frequently than 72, which is observed less frequently than 73, and so on. Similarly, 78 is observed more frequently than 79, which is noted more often than 80, and so forth.

Though this neat symmetric relationship does not characterize all sets of scores, it occurs frequently enough in practice for us to devote special attention to it. In the section on the normal distribution, we explain this concept in greater detail.

Table 2.2 lists the rainfall amounts in San Diego, California, between 1964 and 2007. Figure 2.3 is a histogram based on the observations. The distribution is

TABLE **2.2**
Inches of Rainfall in San Diego, 1964–2007

Year	Rainfall (Inches)	Year	Rainfall (Inches)
1964	5.15	1988	12.44
1965	8.81	1989	5.88
1966	14.76	1990	7.62
1967	10.86	1991	12.31
1968	7.86	1992	12.48
1969	11.48	1993	18.26
1970	6.23	1994	9.93
1971	8.03	1995	17.13
1972	6.12	1996	5.18
1973	10.99	1997	8.74
1974	6.59	1998	20.89
1975	10.64	1999	6.51
1976	10.14	2000	5.77
1977	9.18	2001	8.82
1978	17.3	2002	3.44
1979	14.93	2003	10.24
1980	15.62	2004	5.31
1981	8.13	2005	22.81
1982	11.85	2006	5.35
1983	18.49	2007	3.62
1984	5.37	Sum	454.8
1985	9.6	Mean	10.34
1986	14.64	Standard Deviation	4.71
1987	9.3		

Data from http://cdec.water.ca.gov.

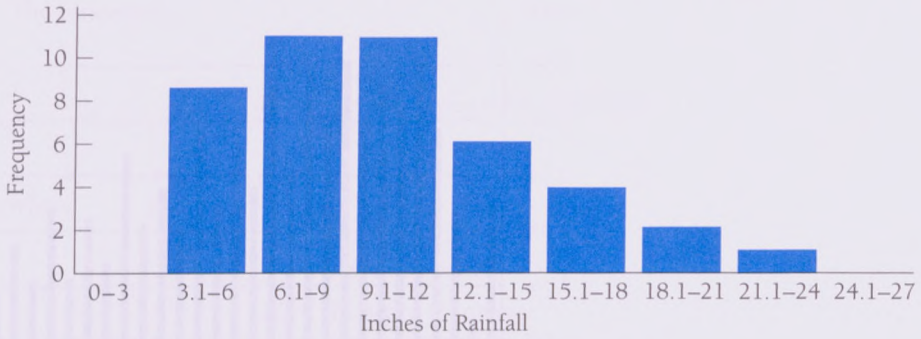


FIGURE **2.3** Histogram for San Diego rainfall, 1964–2007.

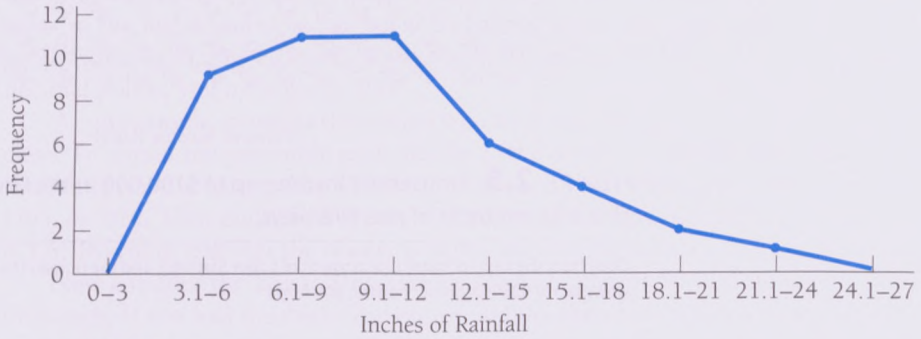


FIGURE **2.4** Frequency polygon for rainfall in San Diego, 1964–2007.

slightly skewed, or asymmetrical. We say that Figure 2.3 has a *positive* skew because the tail goes off toward the higher or positive side of the X axis. There is a slight skew in Figures 2.3 and 2.4, but the asymmetry in these figures is relatively hard to detect. Figure 2.5 gives an example of a distribution that is clearly skewed. The figure summarizes annual household income in the United States in 2007. Very few people make high incomes, while the great bulk of the population is bunched toward the low end of the income distribution. Of particular interest is that this figure only includes household incomes less than $100,000. For household incomes greater than $100,000, the government only reports incomes using class intervals of $50,000. In 2007, about 16% of the U.S. households had incomes greater than $100,000. Because some households have extremely high incomes, you can imagine that the tail of this distribution would go very far to the right. Thus, income is an example of a variable that has positive skew.

One can also present this same set of data as a frequency polygon (see Figure 2.4). Here the amount of rainfall is placed on the graph as a point that represents the

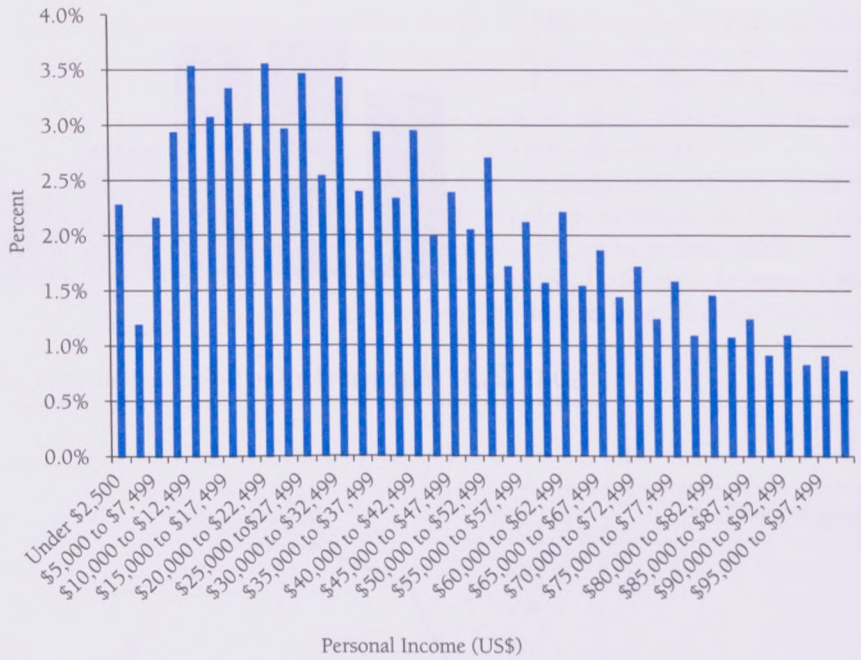


FIGURE **2.5** Household income up to $100,000 in the United States for 2007. This is an example of positive skew.

(Data from the United States Department of Labor Statistics and the Bureau the Census. http://ferret.bls. census .gov/macro/ 032003/hhinc/ new06_000.htm.)

frequencies with which each interval occurs. Lines are then drawn to connect these points.

Whenever you draw a frequency distribution or a frequency polygon, you must decide on the width of the class interval. The **class interval** for inches of rainfall is the unit on the horizontal axis. For example, in Figures 2.3 and 2.4, the class interval is 3 inches—that is, the demarcations along the X axis increase in 3-inch intervals. This interval is used here for convenience; the choice of 3 inches is otherwise arbitrary.

PERCENTILE RANKS

Percentile ranks replace simple ranks when we want to adjust for the number of scores in a group. A **percentile rank** answers the question "What percent of the scores fall below a particular score (X_i)?" To calculate a percentile rank, you need only follow these simple steps: (1) determine how many cases fall below the score of interest, (2) determine how many cases are in the group, (3) divide the number of cases below the score of interest (Step 1) by the total number of cases in the group (Step 2), and (4) multiply the result of Step 3 by 100.

The formula is

$$P_r = \frac{B}{N} \times 100 = \text{percentile rank of } X_i$$

where

P_r = percentile rank
X_i = the score of interest
B = the number of scores below X_i
N = the total number of scores

This means that you form a ratio of the number of cases below the score of interest and the total number of scores. Because there will always be either the same or fewer cases in the numerator (top half) of the equation than there are in the denominator (bottom half), this ratio will always be less than or equal to 1. To get rid of the decimal points, you multiply by 100.

As an example, consider the runner who finishes 62nd out of 63 racers in a gym class. To obtain the percentile rank, divide 1 (the number of people who finish behind the person of interest) by 63 (the number of scores in the group). This gives you 1/63, or .016. Then multiply this result by 100 to obtain the percentile rank, which is 1.6. This rank tells you the runner is below the 2nd percentile.

Now consider the Bay to Breakers race, which attracts 50,000 runners to San Francisco. If you had finished 62nd out of 50,000, then the number of people who were behind you would be 49,938. Dividing this by the number of entrants gives .9988. When you multiply by 100, you get a percentile rank of 99.88. This tells you that finishing 62nd in the Bay to Breakers race is exceptionally good because it places you in the 99.88th percentile.

Psychological Testing in Everyday Life 2.1 presents the calculation of percentile ranks of the infant mortality rates of selected countries as reported by the World Health Organization in 2007. Infant mortality is defined as the number of babies out of 1000 who are born alive but die before their first birthday. Before proceeding, we should point out that the meaning of this calculation depends on which countries are used in the comparison.

In this example, the calculation of the percentile rank is broken into five steps and uses the raw data in the table. In Step 1, we arrange the data points in ascending order. Singapore has the lowest infant mortality rate (2.3), Japan is next (2.8), and Afghanistan has the highest rate (157.0).

In Step 2, we determine the number of cases with worse rates than that of the case of interest. In this example, the case of interest is the United States. Therefore, we count the number of cases with a worse rate than that of the United States. Eleven countries—Israel, Saudi Arabia, Colombia, China, Turkey, Morocco, Bolivia, Laos, Ethiopia, Mozambique, and Afghanistan—have infant mortality rates greater than 6.4.

PSYCHOLOGICAL TESTING IN EVERYDAY LIFE 2.1

Infant Mortality in Selected Countries, 2007

Country	Infant Mortality per 1000 Live Births
Afghanistan	157.0
Australia	4.6
Bolivia	45.6
China	23.0
Colombia	19.1
Ethiopia	86.9
France	3.4
Israel	6.8
Italy	5.7
Japan	2.8
Laos	51.4
Morocco	30.6
Mozambique	95.9
Saudi Arabia	18.8
Singapore	2.3
Spain	4.3
Turkey	27.5
United States	6.4
Mean	32.9
SD	41.9

To calculate the percentile rank of infant mortality in the United States in comparison to that in selected countries, use the following formula:

$$P_r = \frac{B}{N} \times 100$$

where

P_r = the percentile rank
B = the number of cases with worse rates than the case of interest
N = the total number of cases

Country	Infant Mortality per 1000 Live Births
Singapore	2.3
Japan	2.8
France	3.4
Spain	4.3
Australia	4.6
Italy	5.7
United States	6.4
Israel	6.8
Saudi Arabia	18.8
Colombia	19.1
China	23.0
Turkey	27.5
Morocco	30.6
Bolivia	45.6
Laos	51.4
Ethiopia	86.9
Mozambique	95.9
Afghanistan	157.0

STEPS

1. Arrange data in ascending order—that is, the lowest score first, the second lowest score second, and so on.

$$N = 18, \text{ mean} = 32.9, \text{ standard deviation} = 41.9$$

2. Determine the number of cases with worse rates than the score of interest. There are 11 countries in this sample with infant mortality rates greater than that in the United States.
3. Determine the number of cases in the sample (18).
4. Divide the number of scores worse than the score of interest (Step 2) by the total number of scores (Step 3):

$$\frac{11}{18} = .61$$

5. Multiply by 100:

$$.61 \times 100 = 61\text{st percentile rank}$$

In Step 3, we determine the total number of cases (18).

In Step 4, we divide the number of scores worse than the score of interest by the total number of scores:

$$\frac{11}{18} = .61$$

Technically, the percentile rank is a percentage. Step 4 gives a proportion. Therefore, in Step 5 you transform this into a whole number by multiplying by 100:

$$.61 \times 100 = 61$$

Thus, the United States is in the 61st percentile.

The percentile rank depends absolutely on the cases used for comparison. In this example, you calculated that the United States is in the 61st percentile for infant mortality within this group of countries. If all countries in the world had been included, then the ranking of the United States might have been different.

Using this procedure, try to calculate the percentile rank for Bolivia. The calculation is the same except that there are four countries with worse rates than Bolivia (as opposed to 11 worse than the United States). Thus, the percentile rank for Bolivia is

$$\frac{4}{18} = .22 \times 100 = 22$$

or the 22nd percentile. Now try France. You should get a percentile rank of 83.

PERCENTILES

Percentiles are the specific scores or points within a distribution. Percentiles divide the total frequency for a set of observations into hundredths. Instead of indicating what percentage of scores fall below a particular score, as percentile ranks do, percentiles indicate the particular score, below which a defined percentage of scores falls.

Try to calculate the percentile and percentile rank for some of the data in Psychological Testing in Everyday Life 2.1. As an example, look at Italy. The infant mortality rate in Italy is 5.72/1000. When calculating the percentile rank, you exclude the score of interest and count those below (in other words, Italy is not included in the count). There are 12 countries in this sample with infant mortality rates worse than Italy's. To calculate the percentile rank, divide this number of countries by the total number of cases and multiply by 100:

$$P_r = \frac{B}{N} \times 100 = \frac{12}{18} \times 100 = .67 \times 100 = 67$$

Thus, Italy is in the 67th percentile rank, or the 67th percentile in this example is 5.72/1000 or 5.72 deaths per 1000 live births.

Now take the example of Israel. The calculation of percentile rank requires looking at the number of cases below the case of interest. In this example, 10 countries

in this group have infant mortality rates worse than Israel's. Thus, the percentile rank for Israel is $10/18 \times 100 = 56$. The 56th percentile corresponds with the point or score of 6.75 (6.75/1000 live births).

In summary, the percentile and the percentile rank are similar. The percentile gives the point in a distribution below which a specified percentage of cases fall (6.75/1000 for Israel). The percentile is in raw score units. The percentile rank gives the percentage of cases below the percentile; in this example, the percentile rank is 56.

When reporting percentiles and percentile ranks, you must carefully specify the population you are working with. Remember that a percentile rank is a measure of relative performance. When interpreting a percentile rank, you should always ask the question "Relative to what?" Suppose, for instance, that you finished in the 17th percentile in a swimming race (or fifth in a heat of six competitors). Does this mean that you are a slow swimmer? Not necessarily. It may be that this was a heat in the Olympic games, and the participants were the fastest swimmers in the world. An Olympic swimmer competing against a random sample of all people in the world would probably finish in the 99.99th percentile. The example for infant mortality rates depends on which countries in the world were selected for comparison. The United States actually does quite poorly when compared with European countries and the advanced economies in Asia (Singapore and Japan). However, the U.S. infant mortality rate looks much better compared with countries in the developing world.[1]

DESCRIBING DISTRIBUTIONS

Mean

Statistics are used to summarize data. If you consider a set of scores, the mass of information may be too much to interpret all at once. That is why we need numerical conveniences to help summarize the information. An example of a set of scores that can be summarized is shown in Table 2.2 (see page 32), amounts of rainfall in San Diego. We signify the variable as X. A *variable* is a score that can have different values. The amount of rain is a variable because different amounts of rain fell in different years.

The arithmetic average score in a distribution is called the **mean**. To calculate the mean, we total the scores and divide the sum by the number of cases, or N. The capital Greek letter sigma (Σ) means summation. Thus, the formula for the mean, which we signify as $\bar{X}$, is

$$\bar{X} = \frac{\Sigma X}{N}$$

[1]We used a similar example in the last edition based on data from 2003. By 2007, there were significant improvements in the infant mortality rates in developing countries. The rate for Mozambique declined from 148.6 down to 95.9 per thousand live births. Ethiopia reduced its infant mortality rate from 142.6 to 86.9. However, the rates worsened slightly for several developed countries, including Israel, Italy, and Spain.

In words, this formula says to total the scores and divide the sum by the number of cases. Using the information in Table 2.2, we find the mean by following these steps:

1. Obtain ΣX, or the sum of the scores: $5.15 + 8.81 + 14.76 + 10.86 + 7.86 + \cdots + 3.62 = 454.80$
2. Find N, or the number of scores:

$$N = 44$$

3. Divide ΣX by N: $454.80/44 = 10.34$

Psychological Testing in Everyday Life 2.2 summarizes common symbols used in basic statistics.

Standard Deviation

The standard deviation is an approximation of the average deviation around the mean. The standard deviation for the amount of rainfall in San Diego is 4.71. To understand rainfall in San Diego, you need to consider at least two dimensions: first, the amount of rain that falls in a particular year; second, the degree of variation from year to year in the amount of rain that falls. The calculation suggests that, on the average, the variation around the mean is approximately 4.71 inches.

However informative, knowing the mean of a group of scores does not give you that much information. As an illustration, look at the following sets of numbers.

Set 1	Set 2	Set 3
4	5	8
4	5	8
4	4	6
4	4	2
4	3	0
4	3	0

PSYCHOLOGICAL TESTING IN EVERYDAY LIFE 2.2

Common Symbols

You need to understand and recognize the symbols used throughout this book. $\bar{X}$ is the mean; it is pronounced "X bar." Σ is the summation sign. It means sum, or add, scores together and is the capital Greek letter sigma. X is a variable that takes on different values. Each value of X_i represents a raw score, also called an *obtained score*.

Calculate the mean of the first set. You should get 4. What is the mean of the second set? If you calculate correctly, you should get 4 again. Next find the mean for Set 3. It is also 4. The three distributions of scores appear quite different but have the same mean, so it is important to consider other characteristics of the distribution of scores besides the mean. The difference between the three sets lies in *variability*. There is no variability in Set 1, a small amount in Set 2, and a lot in Set 3.

Measuring this variation is similar to finding the average deviation around the mean. One way to measure variability is to subtract the mean from each score $(X - \bar{X})$ and then total the deviations. Statisticians often signify this with a lowercase x, as in $x = (X - \bar{X})$. Try this for the data in Table 2.2. Did you get 0? You should have, and this is not an unusual example. In fact, the sum of the deviations around the mean will always equal 0. However, you do have an alternative: You can square all the deviations around the mean in order to get rid of any negative signs. Then you can obtain the average squared deviation around the mean, known as the **variance**. The formula for the variance is

$$\sigma^2 = \frac{\Sigma(X - \bar{X})^2}{N}$$

where $(X - \bar{X})$ is the deviation of a score from the mean. The symbol σ is the lowercase Greek sigma; σ^2 is used as a standard description of the variance.

Though the variance is a useful statistic commonly used in data analysis, it shows the variable in squared deviations around the mean rather than in deviations around the mean. In other words, the variance is the average squared deviation around the mean. To get it back into the units that will make sense to us, we need to take the square root of the variance. The square root of the variance is the standard deviation (σ), and it is represented by the following formula

$$\sigma = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}}$$

The **standard deviation** is thus the square root of the average squared deviation around the mean. Although the standard deviation is not an average deviation, it gives a useful approximation of how much a typical score is above or below the average score.

Because of their mathematical properties, the variance and the standard deviation have many advantages. For example, knowing the standard deviation of a normally distributed batch of data allows us to make precise statements about the distribution. The formulas just presented are for computing the variance and the standard deviation of a population. That is why we use the lowercase Greek sigma (σ and σ^2). Psychological Testing in Everyday Life 2.3 summarizes when you should use Greek and Roman letters. Most often we use the standard deviation for a sample to estimate the standard deviation for a population. When we talk about a sample, we replace the Greek σ with a Roman letter S. Also, we divide by

$N - 1$ rather than N to recognize that S of a sample is only an estimate of the variance of the population.

$$S = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N - 1}}$$

In calculating the standard deviation, it is often easier to use the raw score equivalent formula, which is

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

This calculation can also be done automatically by some minicalculators.

In reading the formula, you may be confused by a few points. In particular, be careful not to confuse ΣX^2 and $(\Sigma X)^2$. To get ΣX^2, each individual score is squared and the values are summed. For the scores 3, 5, 7, and 8, ΣX^2 would be $3^2 + 5^2 + 7^2 + 8^2 = 9 + 25 + 49 + 64 = 147$. To obtain $(\Sigma X)^2$, the scores are first summed and the total is squared. Using the example, $(\Sigma X)^2 = (3 + 5 + 7 + 8)^2 = 23^2 = 529$.

Z Score

One problem with means and standard deviations is that they do not convey enough information for us to make meaningful assessments or accurate interpretations of data. Other metrics are designed for more exact interpretations. The Z score

PSYCHOLOGICAL TESTING IN EVERYDAY LIFE 2.3

Terms and Symbols Used to Describe Populations and Samples

	Population	Sample
Definition	All elements with the same definition	A subset of the population, usually drawn to represent it in an unbiased fashion
Descriptive characteristics	Parameters	Statistics
Symbols used to describe	Greek	Roman
Symbol for mean	μ	$\bar{X}$
Symbol for standard deviation	σ	S

transforms data into standardized units that are easier to interpret. A Z score is the difference between a score and the mean, divided by the standard deviation:

$$Z = \frac{X_i - \bar{X}}{S}$$

In other words, a Z score is the deviation of a score X_i from the mean $\bar{X}$ in standard deviation units. If a score is equal to the mean, then its Z score is 0. For example, suppose the score and the mean are both 6; then $6 - 6 = 0$. Zero divided by anything is still 0. If the score is greater than the mean, then the Z score is positive; if the score is less than the mean, then the Z score is negative.

Let's try an example. Suppose that $X_i = 6$, the mean $\bar{X} = 3$, and the standard deviation $S = 3$. Plugging these values into the formula, we get

$$Z = \frac{6-3}{3} = \frac{3}{3} = 1$$

Let's try another example. Suppose $X_i = 4$, $\bar{X} = 5.75$, and $S = 2.11$. What is the Z score? It is $-.83$:

$$Z = \frac{4-5.74}{2.11} = \frac{-1.74}{2.11} = -.83$$

This means that the score we observed (4) is .83 standard deviation below the average score, or that the score is below the mean but its difference from the mean is slightly less than the average deviation.

Example of Depression in Medical Students: Center for Epidemiologic Studies Depression Scale (CES-D)

The CES-D is a general measure of depression that has been used extensively in epidemiological studies. The scale includes 20 items and taps dimensions of depressed mood, hopelessness, appetite loss, sleep disturbance, and energy level. Each year, students at the University of California, San Diego, School of Medicine are asked to report how often they experienced a particular symptom during the first week of school on a 4-point scale ranging from rarely or none of the time [0 to 1 days (0)] to most or all of the time [5 to 7 days (3)]. Items 4, 8, 12, and 16 on the CES-D are reverse scored. For these items, 0 is scored as 3, 1 is scored as 2, 2 as 1, and 3 as 0. The CES-D score is obtained by summing the circled numbers. Scores on the CES-D range from 0 to 60, with scores greater than 16 indicating clinically significant levels of depressive symptomatology in adults.

Feel free to take the CES-D measure yourself. Calculate your score by summing the numbers you have circled. However, you must first reverse the scores on items 4, 8, 12, and 16. As you will see in Chapter 5, the CES-D does not have high validity for determining clinical depression. If your score is less than 16, the evidence suggests that you are not clinically depressed. If your score is high, it raises suspicions about depression—though this does not mean you have a problem. (Of course, you may want to talk with your college counselor if you are feeling depressed.)

Center for Epidemiologic Studies Depression Scale (CES-D)

Instructions: *Circle the number for each statement that best describes how often you felt or behaved this way DURING THE PAST WEEK.*

	Rarely or none of the time (less than 1 day)	Some or a little of the time (1–2 days)	Occasionally or a moderate amount of the time (3–4 days)	Most or all of the time (5–7 days)
1. I was bothered by things that usually don't bother me	0	1	2	3
2. I did not feel like eating	0	1	2	3
3. I felt that I could not shake off the blues even with help from my family or friends	0	1	2	3
R 4. I felt that I was just as good as other people	0	1	2	3
5. I had trouble keeping my mind on what I was doing	0	1	2	3
6. I felt depressed	0	1	2	3
7. I felt that everything I did was an effort.	0	1	2	3
R 8. I felt hopeful about the future	0	1	2	3
9. I thought my life had been a failure	0	1	2	3
10. I felt fearful	0	1	2	3
11. My sleep was restless	0	1	2	3
R 12. I was happy	0	1	2	3
13. I talked less than usual	0	1	2	3
14. I felt lonely	0	1	2	3
15. People were unfriendly	0	1	2	3
R 16. I enjoyed life	0	1	2	3
17. I had crying spells	0	1	2	3
18. I felt sad	0	1	2	3
19. I felt that people disliked me	0	1	2	3
20. I could not get "going."	0	1	2	3

Table 2.3 shows CES-D scores for a selected sample of medical students. You can use these data to practice calculating means, standard deviations, and *Z* scores.

In creating the frequency distribution for the CES-D scores of medical students we used an arbitrary class interval of 5.

TABLE **2.3**
The Calculation of Mean, Standard Deviation, and *Z* Scores for CES-D Scores

Name	Test score (*X*)	X^2	*Z* score
John	14	196	.42
Carla	10	100	−.15
Fred	8	64	−.44
Monica	8	64	−.44
Eng	26	676	2.13
Fritz	0	0	−1.58
Mary	14	196	.42
Susan	3	9	−1.15
Debbie	9	81	−.29
Elizabeth	10	100	−.15
Sarah	7	49	−.58
Marcel	12	144	.14
Robin	10	100	−.15
Mike	25	625	1.99
Carl	9	81	.29
Phyllis	12	144	.14
Jennie	23	529	1.70
Richard	7	49	−.58
Tyler	13	169	.28
Frank	1	1	−1.43
	$\Sigma X = 221$	$\Sigma X^2 = 3377$	

$$\bar{X} = \frac{\Sigma X}{N} = \frac{221}{20} = 11.05$$

$$S = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N-1}} = \sqrt{\frac{3377 - \frac{(221)^2}{20}}{20-1}} = 7.01$$

$$\text{Monica's } Z \text{ score} = \frac{X - \bar{X}}{S} = \frac{8 - 11.05}{7.01} = -.44$$

$$\text{Marcel's } Z \text{ score} = \frac{X - \bar{X}}{S} = \frac{12 - 11.05}{7.01} = .14$$

$$\text{Jennie's } Z \text{ score} = \frac{X - \bar{X}}{S} = \frac{23 - 11.05}{7.01} = 1.70$$

Standard Normal Distribution

Now we consider the standard normal distribution because it is central to statistics and psychological testing. First, though, you should participate in a short exercise. Take any coin and flip it 10 times. Now repeat this exercise of 10 coin flips 25 times. Record the number of heads you observe in each group of 10 flips. When you are

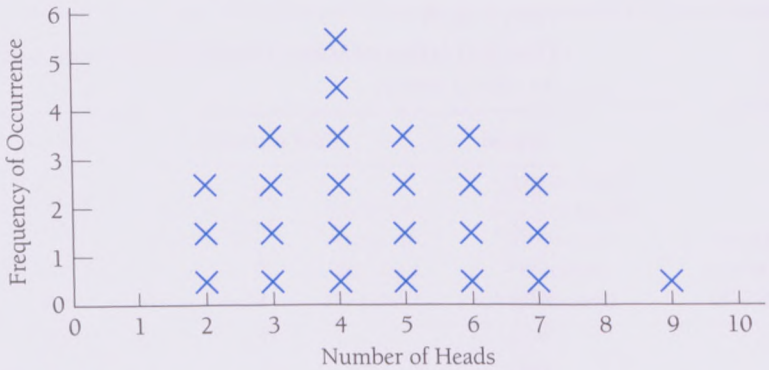


FIGURE **2.6** Frequency distribution of the number of heads in 25 sets of 10 flips.

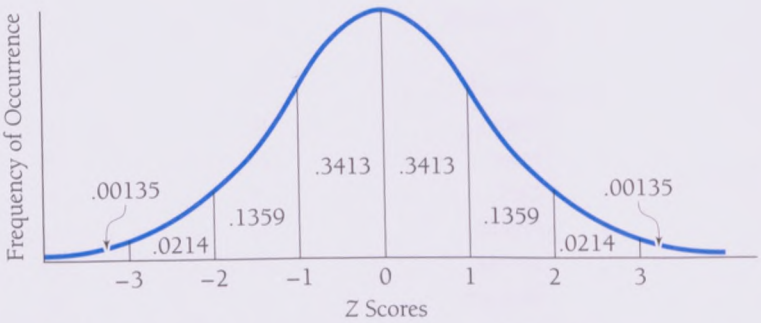


FIGURE **2.7** The theoretical distribution of the number of heads in an infinite number of coin flips.

done, make a frequency distribution showing how many times you observed 1 head in your 10 flips, 2 heads, 3 heads, and so on.

Your frequency distribution might look like the example shown in Figure 2.6. The most frequently observed events are approximately equal numbers of heads and tails. Toward the extremes of 10 heads and 0 tails or 10 tails and 0 heads, events are observed with decreasing frequency. For example, there were no occasions in which fewer than 2 heads were observed and only one occasion in which more than 8 heads were observed. This is what we would expect from the laws of probability. On the average, we would expect half of the flips to show heads and half to show tails if heads and tails are equally probable events. Although observing a long string of heads or tails is possible, it is improbable. In other words, we sometimes see the coin come up heads in 9 out of 10 flips. The likelihood that this will happen, however, is quite small.

Figure 2.7 shows the theoretical distribution of heads in an infinite number of flips of the coin. This figure might look a little like the distribution from your coin-flipping exercise or the distribution shown in Figure 2.6. Actually, this is a normal distribution, or what is known as a *symmetrical binomial probability distribution*.

On most occasions, we refer to units on the X (or horizontal) axis of the normal distribution in Z score units. Any variable transformed into Z score units takes on special properties. First, Z scores have a mean of 0 and a standard deviation of 1.0. If you think about this for a minute, you should be able to figure out why this is true. Recall that the sum of the deviations around the mean is always equal to 0. The numerator of the Z score equation is the deviation around the mean, while the denominator is a constant. Thus, the mean of Z scores can be expressed as

$$\frac{\Sigma(X_i - \bar{X})/S}{N} \quad \text{or} \quad \frac{\Sigma Z}{N}$$

Because $\Sigma(X_i - \bar{X})$ will always equal 0, the mean of Z scores will always be 0. In Figure 2.7, the standardized, or Z score, units are marked on the X axis. The numbers under the curve are the proportions of cases (in decimal form) that we would expect to observe in each area. Multiplying these proportions by 100 yields percentages. For example, we see that 34.13% or .3413 of the cases fall between the mean and one standard deviation above the mean. Do not forget that 50% of the cases fall below the mean. Putting these two bits of information together, we can conclude that if a score is one standard deviation above the mean, then it is at about the 84th percentile rank (50 + 34.13 = 84.13 to be exact). A score that is one standard deviation below the mean would be about the 16th percentile rank (50 − 34.13 = 15.87). Thus, you can use what you have learned about means, standard deviations, Z scores, and the normal curve to transform raw scores, which have little meaning, into percentile scores, which are easier to interpret. These methods can be used only when the distribution of scores is normal or approximately normal. Methods for nonnormal distributions are discussed in most statistics books under "nonparametric statistics."

Percentiles and Z Scores

These percentile ranks are the percentage of scores that fall below the observed Z score. For example, the Z score −1.6 is associated with the percentile rank of 5.48. The Z score 1.0 (third column) is associated with the percentile rank of 84.13.

Part I of Appendix 1 is a simplified version of Part II, which you need for more advanced use of Z scores. Part II gives the areas between the mean and various Z scores. Standard scored values are listed in the "Z" column. To find the proportion of the distribution between the mean of the distribution and a given Z score, you must locate the entry indicated by a specific Z score. Z scores are carried to a second decimal place in the columns that go across the table. First, consider the second column of the table because it is similar to Part I of Appendix 1. Take the Z score of 1.0. The second column is labeled .00, which means that the second decimal place is also 0. The number listed in the table is .3413. Because this is a positive number, it is above the mean. Because the area below the mean is .5, the total area below a Z score of 1.0 is .5 + .3413 = .8413. To make this into a percentile (as shown in Part I of the appendix), multiply by 100 to get 84.13. Now try the example of a Z score of 1.64. To locate this value, find 1.6 in the first column. Then move your hand across the row until you get to the number below the heading .04. The number is .4495. Again, this is a positive Z score, so you must add the observed proportion to the .5

that falls below the mean. The proportion below 1.64 is .9495. Stated another way, 94.95% of the cases fall below a Z score of 1.64. Now try to find the percentile rank of cases that fall below a Z score of 1.10. If you are using the table correctly, you should obtain 86.43.

Now try −.75. Because this is a negative Z score, the percentage of cases falling below the mean should be less than 50. But there are no negative values in Part II of Appendix 1. For a negative Z score, there are several ways to obtain the appropriate area under the curve. The tables in Appendix 1 give the area from the mean to the Z score. For a Z score of −.75, the area between the mean and the Z score is .2734. You can find this by entering the table in the row labeled .7 and then moving across the row until you get to the figure in that row below the heading .05. There you should find the number .2734. We know that .5 of the cases fall below the mean. Thus, for a negative Z score, we can obtain the proportion of cases falling below the score by subtracting .2734, the tabled value listed in the appendix, from .5. In this case, the result is

$$.5 - .2734 = .2266$$

Because finding the percentile ranks associated with negative Z scores can be tricky, you might want to use Part I of Appendix 1 to see if you are in the right ballpark. This table gives both negative and positive Z scores but does not give the detail associated with the second decimal place. Look up −.7 in Part I. The percentile rank is 24.20. Now consider a Z score of −.8. That percentile rank is 21.19. Thus, you know that a Z score of −.75 should be associated with a percentile rank between 21.19 and 24.20. In fact, we have calculated that the actual percentile rank is 22.66.

Practice with Appendix 1 until you are confident you understand how it works. Do not hesitate to ask your professor or teaching assistant if you are confused. This is an important concept that you will need throughout the rest of the book. After you have mastered using the Tables in Appendix 1, you might try a nifty website (http://davidmlane.com/hyperstat/z_table.html) that can find the probabilities for you.

Look at one more example from Table 2.2 (rainfall in San Diego, page 32). California had a dry year in 1999 and in 2007. In both years, the newscasters frequently commented that this was highly unusual. They described it as the "La Nina" effect, and some even claimed that it signaled global warming. The question is whether or not the amount of rainfall received in 1999 and 2007 was unusual given what we know about rainfall in general. To evaluate this, calculate the Z score for rainfall. According to Table 2.2, there were 6.51 inches of rainfall in 1999 and 3.62 inches in 2007. The mean for rainfall is 10.33 inches and the standard deviation is 4.71. Thus, the Z score for 1999 is

$$\frac{6.51 - 10.33}{4.71} = -.81$$

Next determine where a Z score of −.81 falls within the Z distribution. According to Appendix 1, a Z score of −.81 is equal to the 20.9th percentile (50 − 29.1). Thus, the low rainfall year in 1999 was unusual—given all years, it was in about the 21st percentile. However, it was not *that* unusual. You can estimate that there would be less

rainfall in approximately 17% of all years. 2007 was a different case. The *Z* score for 2007 was −1.43. Rainfall in 2007 was in the 7.64th percentile. (Using Appendix 1, you can look up *Z* score of −1.43 and find an area below the mean of 0.4236. Then you can estimate the percentile as 50 − 42.36 = 7.64.)

You can also turn the process around. Instead of using *Z* scores to find the percentile ranks, you can use the percentile ranks to find the corresponding *Z* scores. To do this, look in Part II of Appendix 1 under percentiles and find the corresponding *Z* score. For example, suppose you wish to find the *Z* score associated with the 90th percentile. When you enter Part II of Appendix 1, look for the value closest to the 90th percentile. This can be a little tricky because of the way the table is structured. Because the 90th percentile is associated with a positive *Z* score, you are actually looking for the area above the 50th percentile. So you should look for the entry closest to .4000 (.5000 + .4000 = .9000). The closest value to .4000 is .3997, which is found in the row labeled 1.2 and the column labeled .08. This tells you that a person who obtains a *Z* score of 1.28 is at approximately the 90th percentile in the distribution.

Now return to the example of CES-D scores for medical students (Table 2.3). Monica had a *Z* score on the CES-D of −.44. Using Appendix 1, you can see that she was in the 33rd percentile (obtained as .50 − .1700 = .33 × 100 = 33). Marcel, with his *Z* score of .14, was in the 56th percentile; and Jennie, with a *Z* score of 1.70, was in the 96th percentile. You might have few worries about Monica and Marcel. However, it appears that Jennie is more depressed than 96% of her classmates and may need to talk to someone.

An Example Close to Home

One of the difficulties in grading students is that performance is usually rated in terms of raw scores, such as the number of items a person correctly answers on an examination. You are probably familiar with the experience of having a test returned to you with some number that makes little sense to you. For instance, the professor comes into class and hands you your test with a 72 on it. You must then wait patiently while he or she draws the distribution on the board and tries to put your 72 into some category that you understand, such as B+.

An alternative way of doing things would be to give you a *Z* score as feedback on your performance. To do this, your professor would subtract the average score (mean) from your score and divide by the standard deviation. If your *Z* score was positive, you would immediately know that your score was above average; if it was negative, you would know your performance was below average.

Suppose your professor tells you in advance that you will be graded on a curve according to the following rigid criteria. If you are in the top 15% of the class, you will get an A (85th percentile or above); between the 60th and the 84th percentiles, a B; between the 20th and the 59th percentiles, a C; between the 6th and the 19th percentiles, a D; and in the 5th percentile or below, an F. Using Appendix 1, you should be able to find the *Z* scores associated with each of these cutoff points for normal distributions of scores. Try it on your own and then consult Table 2.4 to see if you are correct. Looking at Table 2.4, you should be able to determine what your grade would be in this class on the basis of your *Z* score. If your *Z* score is 1.04 or greater, you would receive an A; if it were greater than .25 but less than 1.04,

TABLE **2.4**
Z Score Cutoffs for a Grading System

Grade	Percentiles	*Z* score cutoff
A	85–100	1.04
B	60–84	.25
C	20–59	−.84
D	6–19	−1.56
F	0–5	<−1.56

you would get a B; and so on. This system assumes that the scores are distributed normally.

Now try an example that puts a few of these concepts together. Suppose you get a 60 on a social psychology examination. You learned in class that the mean for the test was 55.70 and that the standard deviation was 6.08. If your professor uses the grading system that was just described, what would your grade be?

To solve this problem, first find your *Z* score. Recall the formula for a *Z* score:

$$Z = \frac{X_i - \bar{X}}{S}$$

So your *Z* score would be

$$Z = \frac{60 - 55.70}{6.08} = \frac{4.30}{6.08} = .707$$

Looking at Table 2.4, you see that .707 is greater than .25 (the cutoff for a B) but less than 1.04 (the cutoff for an A). Now find your exact standing in the class. To do this, look again at Appendix 1. Because the table gives *Z* scores only to the second decimal, round .707 to .71. You will find that 76.11% of the scores fall below a *Z* score of .71. This means that you would be in approximately the 76th percentile, or you would have performed better on this examination than approximately 76 out of every 100 students.

McCall's *T*

There are many other systems by which one can transform raw scores to give them more intuitive meaning. One system was established in 1939 by W. A. McCall, who originally intended to develop a system to derive equal units on mental quantities. He suggested that a random sample of 12-year-olds be tested and that the distribution of their scores be obtained. Then percentile equivalents were to be assigned to each raw score, showing the percentile rank in the group for the people who had obtained that raw score. After this had been accomplished, the mean of the distribution would be set at 50 to correspond with the 50th percentile. In McCall's system, called **McCall's *T***, the standard deviation was set at 10.

In effect, McCall generated a system that is exactly the same as standard scores (*Z* scores), except that the mean in McCall's system is 50 rather than 0 and the

standard deviation is 10 rather than 1. Indeed, a Z score can be transformed to a ***T* score** by applying the linear transformation

$$T = 10Z + 50$$

You can thus get from a Z score to McCall's T by multiplying the Z score by 10 and adding 50. It should be noted that McCall did not originally intend to create an alternative to the Z score. He wanted to obtain one set of scores that could then be applied to other situations without standardizing the entire set of numbers.

There is nothing magical about the mean of 50 and the standard deviation of 10. It is a simple matter to create systems such as standard scores with any mean and standard deviation you like. If you want to say that you got a score 1000 points higher than a person who was one standard deviation below you, then you could devise a system with a mean of 100,000 and a standard deviation of 1000. If you had calculated Z scores for this distribution, then you would obtain this with the transformation

$$NS \text{ (for new score)} = 1000Z + 100{,}000$$

In fact, you can create any system you desire. To do so, just multiply the Z score by whatever you would like the standard deviation of your distribution to be and then add the number you would like the mean of your new distribution to be.

An example of a test developed using standardized scores is the SAT Reasoning Test. When this test was created in 1941, the developers decided to make the mean score 500 and the standard deviation 100. Thus, they multiplied the Z scores for those who took the test by 100 and added 500. For a long time, the basic scoring system was used and the 1941 norms were applied. In other words, if the average score of test takers was below the 1941 reference point, the mean for any year could be less than or more than 500. However, in 1995, the test was changed so that the mean each year would be 500 and the standard deviation would be 100. In other words, the test is recalibrated each year. However, drifts continue. For example, in 2007 the average scores on the SAT were 494 for writing, 502 for critical reading and 515 for math (data from www.collegeboard.com).

It is important to make the distinction between standardization and normalization. McCall's T and the other methods described in this section standardize scores by applying a linear transformation. These transformations do not change the characteristics of the distributions. If a distribution of scores is skewed before the transformation is applied, it will also be skewed after the transformation has been used. In other words, transformations standardize but do not normalize.

Quartiles and Deciles

The terms *quartiles* and *deciles* are frequently used when tests and test results are discussed. The two terms refer to divisions of the percentile scale into groups. The quartile system divides the percentage scale into four groups, whereas the decile system divides the scale into 10 groups.

Quartiles are points that divide the frequency distribution into equal fourths. The first quartile is the 25th percentile; the second quartile is the **median**, or 50th, percentile; and the third quartile is the 75th percentile. These are abbreviated Q1, Q2, and Q3, respectively. One fourth of the cases will fall below Q1, one half will

TABLE **2.5**
Transformation of Percentile Scores into Stanines

Percentage of cases	Percentiles	Stanines
4	1–4	1 Bottom 4 percent
7	5–11	2
12	12–23	3
17	24–40	4
20	41–60	5
17	61–77	6
12	78–89	7
7	90–96	8
4	97–100	9 Top 4 percent

fall below Q2, and three fourths will fall below Q3. The interquartile range is the interval of scores bounded by the 25th and 75th percentiles. In other words, the **interquartile range** is bounded by the range of scores that represents the middle 50% of the distribution.

Deciles are similar to quartiles except that they use points that mark 10% rather than 25% intervals. Thus, the top decile, or D9, is the point below which 90% of the cases fall. The next decile (D8) marks the 80th percentile, and so forth.

Another system developed in the U.S. Air Force during World War II is known as the **stanine system**. This system converts any set of scores into a transformed scale, which ranges from 1 to 9. Actually the term *stanine* comes from "standard nine." The scale is standardized to have a mean of 5 and a standard deviation of approximately 2. It has been suggested that stanines had computational advantages because they required only one column on a computer card (Anastasi & Urbina, 1997). Because computer cards are no longer used, this advantage is now questionable.

Table 2.5 shows how percentile scores are converted into stanines. As you can see, for every 100 scores, the lowest 4 (or bottom 4% of the cases) fall into the first stanine. The next 7 (or 7% of the cases) fall into the second stanine, and so on. Finally, the top 4 cases fall into the top stanine. Using what you have learned about Z scores and the standard normal distribution, you should be able to figure out the stanine for a score if you know the mean and the standard deviation of the distribution that the score comes from. For example, suppose that Igor received a 48 on his normally distributed chemistry midterm. The mean in Igor's class was 42.6, and the standard deviation was 3.6. First you must find Igor's Z score. Do this by using the formula

$$Z = \frac{X_i - \bar{X}}{S} \quad \text{so} \quad Z = \frac{48 - 42.6}{3.6} = 1.5$$

Now you need to transform Igor's Z score into his percentile rank. To do this, use Appendix 1. Part I shows that a Z score of 1.5 is in approximately the 93rd percentile. Thus, it falls into the 8th stanine.

Actually, you would rarely go through all these steps to find a stanine. There are easier ways of doing this, including computer programs that do it automatically. However, working out stanines the long way will help you become familiar with a variety of concepts covered in this chapter, including standard scores, means, standard deviations, and percentiles. First, review the five steps to go from raw scores to stanines:

1. Find the mean of the raw scores.
2. Find the standard deviation of the raw scores.
3. Change the raw scores to *Z* scores.
4. Change the *Z* scores to percentiles (using Appendix 1).
5. Use Table 2.5 to convert percentiles into stanines.

An alternative method is to calculate the percentile rank for each score and use Table 2.5 to obtain the stanines. Remember: In practice, you would probably use a computer program to obtain the stanines. Although stanines are not used much in the modern computer era, you can still find them in popular educational tests such as the Stanford Achievement Test.

NORMS

Norms refer to the performances by defined groups on particular tests. There are many ways to express norms, and we have discussed some of these under the headings of *Z* scores, percentiles, and means. The norms for a test are based on the distribution of scores obtained by some defined sample of individuals. The mean is a norm, and the 50th percentile is a norm. Norms are used to give information about performance relative to what has been observed in a standardization sample.

Much has been written about norms and their inadequacies. In later chapters, we shall discuss this material in relation to particular tests. We cover only the highlights here. Whenever you see a norm for a test, you should ask how it was established. Norms are obtained by administering the test to a sample of people and obtaining the distribution of scores for that group.

For example, say you develop a measure of anxiety associated with taking tests in college. After establishing some psychometric properties for the test, you administer the test to normative groups of college students. The scores of these groups of students might then serve as the norms. Say that, for the normative groups of students, the average score is 19. When your friend Alice comes to take the test and obtains a score of 24, the psychologist using the test might conclude that Alice is above average in test anxiety.

The SAT, as indicated earlier, has norms. The test was administered to millions of high-school seniors from all over the United States. With distributions of scores for this normative group, one could obtain a distribution to provide meaning for particular categories of scores. For example, in the 1941 national sample, a person who scored 650 on the verbal portion of the SAT was at the 93rd percentile of high-school seniors. However, if you took the test before 1995 and scored 650, it did not mean that you were in the 93rd percentile of the people who took the test when you did. Rather, it meant that you would have been at the 93rd percentile if you had been in the group the test had been standardized on. However, if the normative

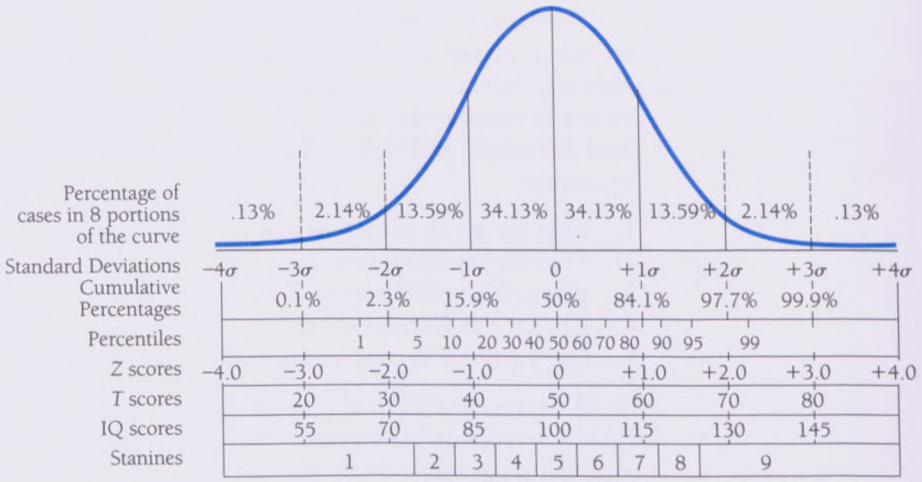


FIGURE **2.8** Standard normal distribution with cumulative percentages, percentiles, *Z* scores, *T* scores, IQ scores, and stanines.

group was a representative sample of the group to which you belonged (and there is every reason to believe it was), then you could reasonably assume that you were in approximately the 93rd percentile of your own group.[2] After 1995, an SAT score of 650 would place you in the 93rd percentile of the people who took the test during the year you completed it. Some controversies surrounding norms are discussed in Psychological Testing in Everyday Life 2.4.

In Chapters 9 and 10 we will review intelligence tests. Most intelligence tests are transformed to have a mean of 100 and a standard deviation of 15. Thus, an IQ score of 115 is one standard deviation above the mean and an IQ score of 130 is two standard deviations above the mean. Using the information we have reviewed, you can determine that an IQ score of 115 is approximately in the 84th percentile, while an IQ score of 85 is approximately in the 16th percentile. Only some 0.13% of the population obtains an IQ score of 145, which is three standard deviations above the mean. Figure 2.8 shows the standard normal distribution with the *Z* scores, *T* scores, IQ scores, and stanines. Examining the figure, locate the point that is one standard deviation above the mean. That point is associated with a *Z* score of 1.0, a *T* score of 60, an IQ score of 115, and the seventh stanine. Using the figure, try to find the score on each scale for an observation that falls two standard deviations below the mean. You should get a *Z* score of −2.0, a *T* score of 30, an IQ score of 70, and a stanine of 1.

Age-Related Norms

Certain tests have different normative groups for particular age groups. Most IQ tests are of this sort. When the Stanford-Binet IQ test was originally created, distributions of the performance of random samples of children were obtained for various

[2]Based on the *American Testing Program Guide for 1989–1991*, College Board of the Educational Testing Service, Princeton, New Jersey.

age groups. When applying an IQ test, the tester's task is to determine the mental age of the person being tested. This is accomplished through various exercises that help locate the age-level norm at which a child is performing.

Tracking

One of the most common uses of age-related norms is for growth charts used by pediatricians. Consider the question "Is my son tall or short?" The answer will usually depend on a comparison of your son to other boys of the same age. Your son would be quite tall if he were 5 feet at age 8 but quite short if he were only 5 feet at age 18. Thus, the comparison is usually with people of the same age.

PSYCHOLOGICAL TESTING IN EVERYDAY LIFE 2.4

Within-Group Norming Controversy

One of the most troubling issues in psychological testing is that different racial and ethnic groups do not have the same average level of performance on many tests (see Chapter 19). When tests are used to select employees, a higher percentage of majority applicants are typically selected than their representation in the general population would indicate. For example, employers who use general aptitude tests consistently overselect white applicants and underselect African Americans and Latinos or Latinas. *Overselection* is defined as selecting a higher percentage from a particular group than would be expected on the basis of the representation of that group in the applicant pool. If 60% of the applicants are white and 75% of those hired are white, then overselection has occurred.

The U.S. Department of Labor uses the General Aptitude Test Battery (GATB) to refer job applicants to employers. At one point, however, studies demonstrated that the GATB adversely affected the hiring of African Americans and Latinos and Latinas. To remedy this problem, a few years ago the department created separate norms for different groups. In other words, to obtain a standardized score, each applicant was compared only with members of his or her own racial or ethnic group. As a result, overselection based on test scores was eliminated. However, this provoked other problems. For example, consider two applicants, one white and one African American, who are in the 70th percentile on the GATB. Although they have the same score, they are compared with different normative groups. The raw score for the white applicant would be 327, while that for the African American would be 283 (Brown, 1994). This was seen as a problem because an African American applicant might be selected for a job even though she had a lower raw score, or got fewer items correct, than did a white applicant.

The problem of within-group norming is highlighted by opposing opinions from different prestigious groups. The National Academy of Sciences, the most elite group of scholars in the United States, reviewed the issue and concluded that separate norms were appropriate. Specifically, they argued that minority workers at a given level of expected job performance are less likely to be hired

(continues)

PSYCHOLOGICAL TESTING IN EVERYDAY LIFE 2.4 *(continued)*

than are majority group members. The use of separate norms was therefore required in order to avoid adverse impact in hiring decisions (Gottfredson, 1994; Hartigan & Wigdor, 1989).

In contrast to this conclusion, legislation has led to different policies. Section 106 of the Civil Rights Act of 1991 made it illegal to use separate norms. The act states that it is unlawful for employers

> in connection with the selection or referral of applicants or candidates for employment or promotion to adjust the scores of, use different cut-offs for, or otherwise alter the results of employment-related tests on the basis of race, color, religion, sex, or national origin.

Employers may have a variety of different objectives when making employment decisions. One goal may be to enhance the ethnic and racial diversity of their workforce. Another goal may be to hire those with the best individual profiles. Often these goals compete. The law may now prohibit employers from attempting to balance these competing objectives (Sackett & Wilk, 1994).

Beyond this rather obvious type of age-related comparison, child experts have discovered that children at the same age level tend to go through different growth patterns. Children who are small as infants often remain small and continue to grow at a slower pace than do others. Pediatricians must therefore know more than a child's age; they must also know the child's percentile within a given age group. For a variety of physical characteristics, children tend to stay at about their same percentile level, relative to other children in their age group, as they grow older. This tendency to stay at about the same level relative to one's peers is known as **tracking**. Height and weight are good examples of physical characteristics that track. Figures 2.9 and 2.10 show the expected rates of growth in terms of length (height) for boys and girls. The charts are based on national norms from the U.S. Centers for Disease Control and Prevention (CDC). Notice that the children who were the largest as babies are expected to remain the largest as they get older.

Pediatricians use the charts to determine the expected course of growth for a child. For example, if a 3-month-old boy was 24 inches in length (about 61 cm), the doctor would locate the child on the center line on the bottom half of Figure 2.9. By age 36 months, the child would be expected to be about 37.5 inches (or about 95 cm). The tracking charts are quite useful to doctors because they help determine whether the child is going through an unusual growth pattern. A boy who had a length of 24 inches at age 3 months might come under scrutiny if at age 36 months he had a length of 35 inches. He would have gone from the 50th percentile to about the 3rd percentile in relation to other boys. This might be normal for 3-year-olds if the boy had always been in the 3rd percentile, but unusual for a boy who had been in the middle of the length distribution. The doctor might want to determine why the child did not stay in his track.

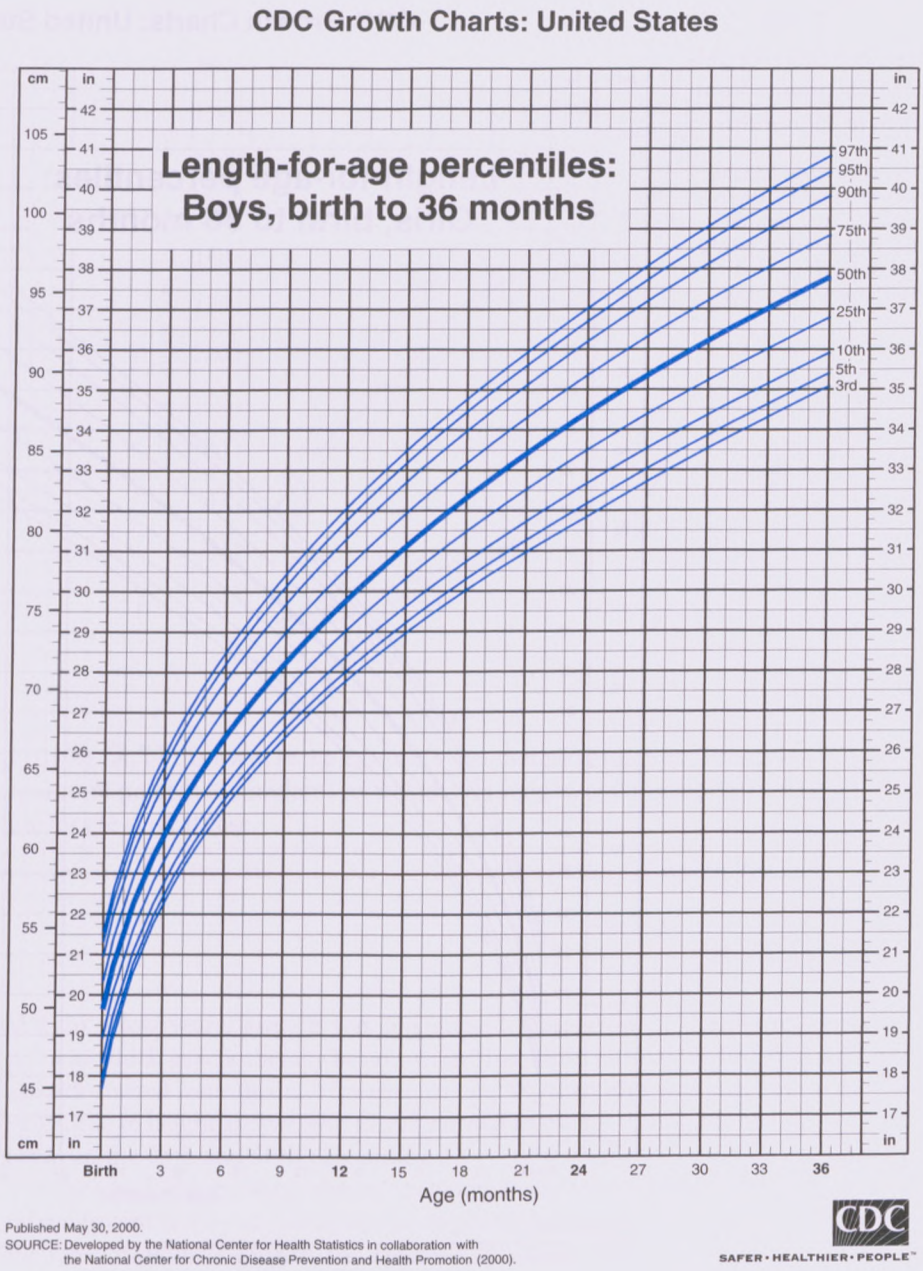


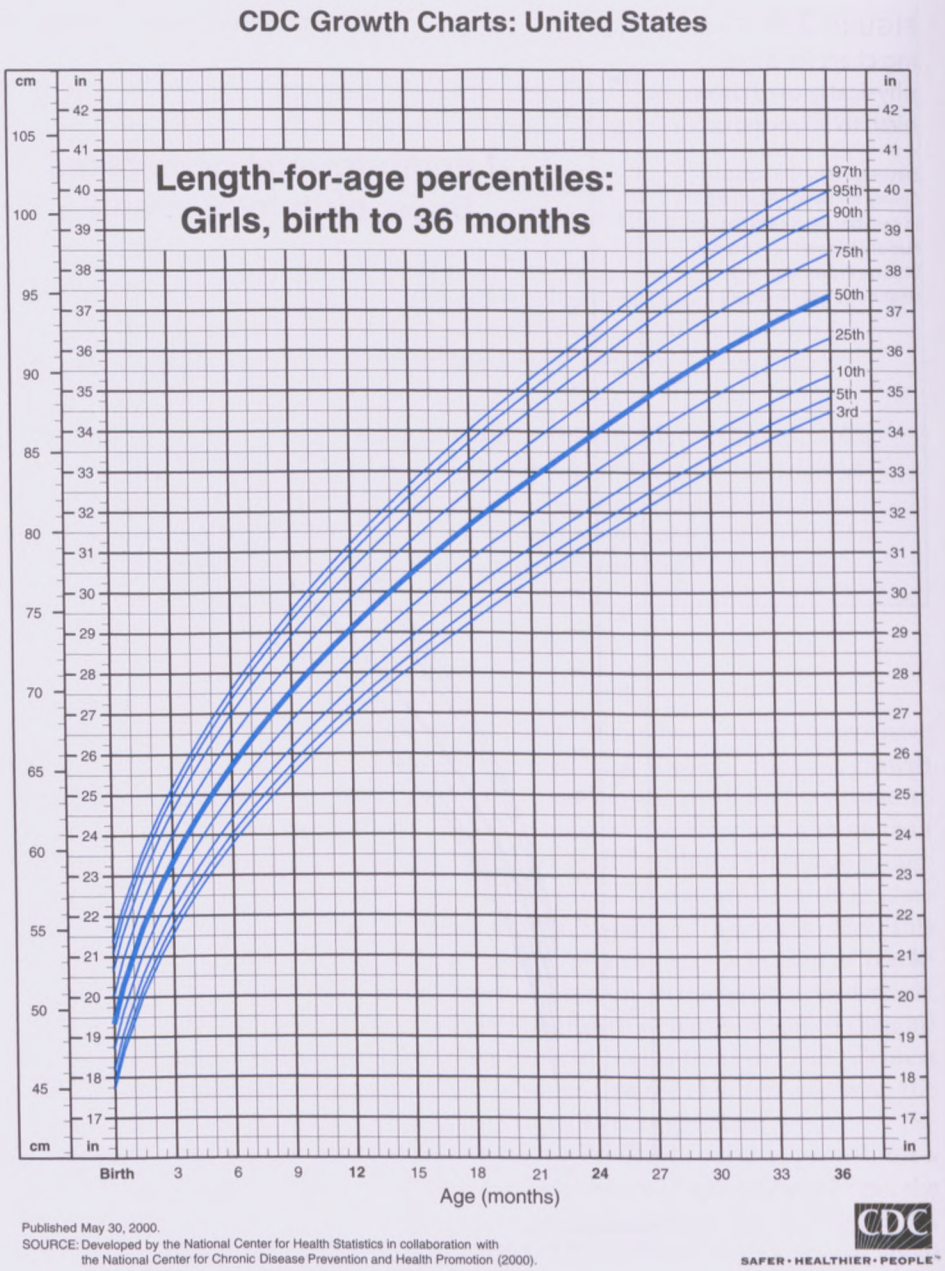
FIGURE 2.9 Tracking chart for boys' physical growth from birth to 36 months.

Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000).

Figure 2.11 shows an example of a child going off track. There is some concern that children who are fed a fat-restricted diet experience stunted growth (Kaplan & Toshima, 1992). The consequences of a slightly restricted vegetarian diet are mild if they exist at all. However, highly restricted diets may affect growth. For instance, Pugliese, Lifshitz, Grad, Fort, and Marks-Katz (1983) studied

FIGURE **2.10**
Tracking chart for girls' physical growth from birth to 36 months.

Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000).



24 adolescents who had voluntarily undergone severe caloric restrictions because they wanted to lose weight. Though they did not have anorexia nervosa, they consumed only a small percentage of the calories recommended for their age. Figure 2.11 shows the growth pattern for one of these children. As the figure suggests, the child grew normally until age 9. At that point, highly restricted dieting began.

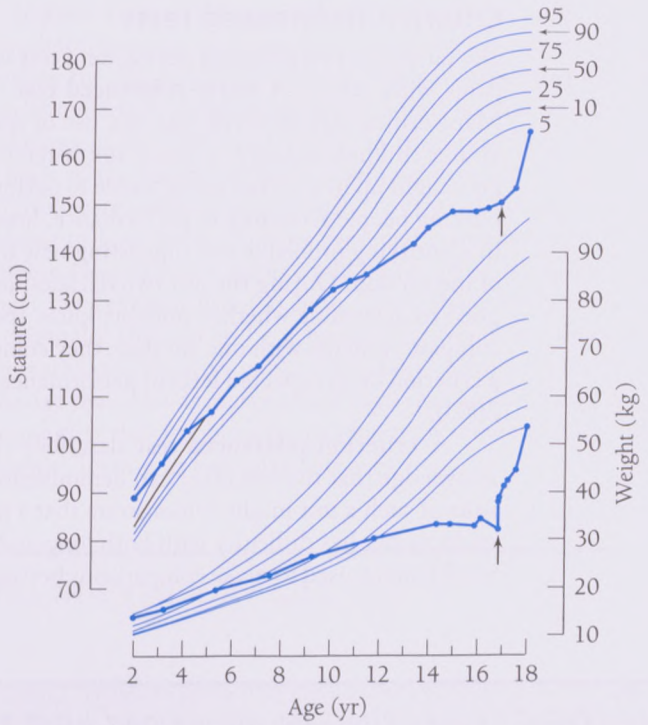


FIGURE **2.11** Growth in the case of severe dietary restriction. The scales represent percentile standards for height and weight, and the plotted values are for the clinical case.

(From Pugliese et al., 1983, p. 514; reprinted by permission of *The New England Journal of Medicine, 309,* 513–518, 1983.)

Within a few years, growth was interrupted. The arrow in the figure shows the point at which psychotherapy began. After this point, normal feeding resumed, and growth started once again. However, at age 18, the child was still below the 5th percentile in height and weight. Given normal tracking, this child should have been between the 25th and 50th percentiles.

Although the tracking system has worked well for medicine, it has stirred considerable controversy in education. Some people believe there is an analogy between the rates of physical growth and the rates of intellectual growth: Just as there are some slow growers who eventually will be shorter than average adults, there are slow learners who will eventually know less as adults. Furthermore, some suggest that children learn at different rates. Children are therefore separated early in their educational careers and placed in classrooms that correspond with these different tracks. Many educators have attacked the tracking system because it discriminates against some children. Because people use psychological tests to place children in these tracks, some tests have come under severe scrutiny and attack. We shall return to this controversy in Chapters 19 and 20.

Criterion-Referenced Tests

The purpose of establishing norms for a test is to determine how a test taker compares with others. A **norm-referenced test** compares each person with a norm. Many critics have objected that this use of tests forces competition among people. Young children exposed to many norm-referenced tests in elementary school can get caught up in a never-ending battle to perform better than average. In addition to ranking people according to performance, however, tests can play an important role in identifying problems and suggesting new directions for individualized programs of instruction. During the last two decades, interest has grown in tests that are applied to determine whether students know specific information. These tests do not compare students with one another; they compare each student's performance with a criterion or an expected level of performance (Hartman & Looney, 2003; Wiberg, 2003).

A **criterion-referenced test** describes the specific types of skills, tasks, or knowledge that the test taker can demonstrate such as mathematical skills. The results of such a test might demonstrate that a particular child can add, subtract, and multiply but has difficulty with both long and short division. The results of the test would not be used to make comparisons between the child and other members of his

PSYCHOLOGICAL TESTING IN EVERYDAY LIFE 2.5

Within High-School Norms for University Admission

Beginning in 2002, the University of California changed its admissions policy. The university had discovered that its admissions did not reflect the demographic characteristics of the state. In particular, students from underrepresented groups and those from low-income neighborhoods were not gaining admission to the university. When the university was required to give up its affirmative action program, there were serious concerns that the student classes would not reflect the diversity of the state of California.

To address this problem, the university created the Eligibility in Local Context (ELC) program. This program guarantees eligibility for university admission to the top 4% of graduates of California high schools. The plan focuses only on high-school grades and does not require the SAT test.

The purpose of this policy is to provide norming within particular high schools. In other words, students are not competing with all other students in the state but are being compared only with those who have had similar educational exposures. The policy was designed to increase the number of students from underrepresented ethnic and minority groups who were admitted to the university. Unfortunately, the program was not successful. Latino acceptance rates dropped from 68% in 1995 to 45% in 2003. African American acceptance rates were 58% in 1995 and dropped to 35% by 2003. As a result, the program was abandoned.

Details can be obtained from www.ucop.edu/sas/elc.

or her class. Instead, they would be employed to design an individualized program of instruction that focuses on division. Thus, the criterion-referenced testing movement emphasizes the diagnostic use of tests—that is, using them to identify problems that can be remedied. Criterion-referenced tests became an important trend in clinical psychology in the 1980s and 1990s. In educational testing, the same ideas formed the basis of the standards-based testing movement. Instead of comparing how well children were performing in relation to other children, schools were evaluated on the basis of how many students exceeded a criterion score. Under the No Child Left Behind legislation, schools could lose funding if the number of students failing to meet the criterion was too high. Thus, the criterion-referenced testing movement, originally thought to be a humanistic trend, became associated with a more conservative approach to public education. Advocates for standards-based testing emphasize that schools must enforce high standards. Critics of the standards movement argue that the cut points for passing high stakes tests are often arbitrary (see Psychological Testing in Everyday Life 2.6).

PSYCHOLOGICAL TESTING IN EVERYDAY LIFE 2.6

No Child Left Behind

Several contemporary issues in testing are relevant to the No Child Left Behind (NCLB) Act. NCLB was initiated by President George W. Bush with the justification that every child should receive a quality education and that "no child is left behind." In 2002, Congress passed the legislation with significant support from both Democrats and Republicans. The key provisions of the bill require greater accountability for school performance. In particular, performance information about school districts was made public. Schools were required to show how well they were doing by reporting the proportions of students who passed standardized tests. The law required that states test each child for proficiency in reading and math in Grades 3 through 8. Further, each child is tested at least once during the high school years.

On the surface, it seemed like there was little to dislike about NCLB. However, by the time the legislation was considered for reauthorization in 2007, numerous concerns had emerged. In particular, critics were very concerned about the application of testing and the effects the test results might have on school funding. NCLB uses standardized achievement tests to evaluate school performance. The tests are "standards-based." This means that children must demonstrate specific critical knowledge in order to pass the test, and passing is defined by an arbitrary cut point. Those above the cut point pass the exam, and those below the cut point fail. Because tests became such an important component of school evaluation, many critics believed that teachers and schools were simply "teaching to the test." In other words, schools and teachers focused on subjects that were covered on the test, like reading and math, while disregarding many other important topics, including the arts and some humanities.

PSYCHOLOGICAL TESTING IN EVERYDAY LIFE 2.7

Are 4th Graders Smarter than 3rd Graders?

One of the most important instruments for measuring school performance is the standardized achievement test. California uses a standardized testing and reporting (STAR) system. STAR is very important because it evaluates performance for high stakes programs such as No Child Left Behind. In an average year, nearly 5 million children are tested in the STAR program. The evaluation of information from the STAR program reveals some interesting trends. The test reveals several indicators of school performance, and these measures are taken quite seriously. Figure 2.12 summarizes the percentage of students who perform at an advanced level in different grades. The data are from three separate years: 2005, 2006, and 2007. The graph shows that each year, relatively few 3rd graders perform at the advanced level while many 4th graders perform at the advanced level. Does this mean that 4th grade students are smarter than 3rd grade students? Norm-referenced test results would never lead to this conclusion because students at each level are compared relative to other students in their same grade level. Standards-based testing does not attend to how students are performing in relation to other students. Instead, the tests consider performance relative to a defined standard. One explanation for the data shown in the graph is that there was an exceptional crop of good students in a particular 4th grade. However, this explanation seems unlikely because the effect occurs each year. A more plausible explanation is that the definition of performing at an advanced level is somewhat arbitrary. The test may be too hard for 3rd graders but perhaps too easy for 4th graders.

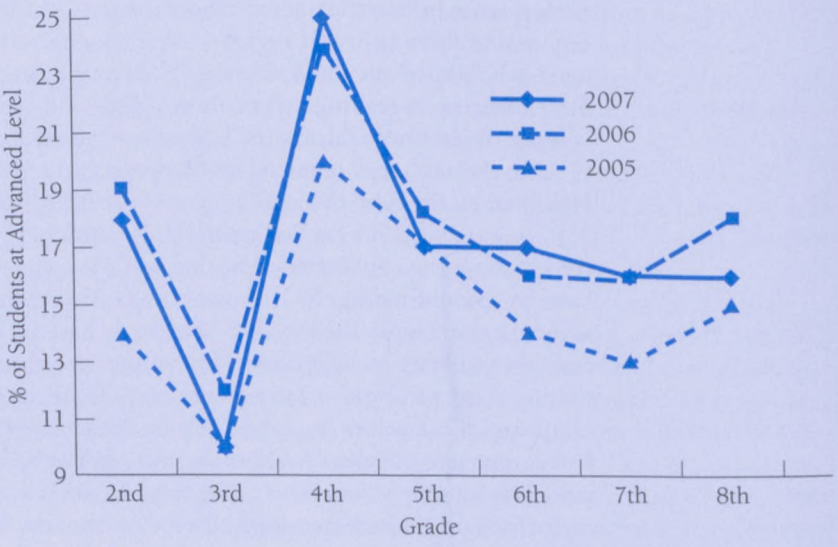


FIGURE **2.12** STAR test performance by grade in three years.

SUMMARY

In this chapter, we discussed some basic rules for translating observations of human activities into numbers. The use of number systems is important for precision in all scientific exercises. Measures of psychological processes are represented by one of four types of scales. A *nominal scale* simply assigns numbers to categories. This type of scale has none of the properties of a numbered scale. An *ordinal scale* has the property of magnitude and allows us to rank objects, but it does not have the property of equal intervals or an absolute 0. An *interval scale* can describe the distances between objects because it has the property of equal intervals in addition to the property of magnitude. A *ratio scale* has an absolute 0 in addition to equal intervals and magnitude. Any mathematical operation on a ratio scale is permissible.

To make sense out of test scores, we have to examine the score of an individual relative to the scores of others. To do this requires creating a distribution of test scores. There are several ways to display the distribution of scores, including frequency distributions and frequency polygons. We also need statistics to describe the distribution. The *mean* is the average score, the *variance* is the averaged squared deviation around the mean, and the *standard deviation* is the square root of the variance. Using these statistics, we can tell a lot about a particular score by relating it to characteristics of a well-known probability distribution known as the standard normal distribution.

Norms are used to relate a score to a particular distribution for a subgroup of a population. For example, norms are used to describe where a child is on some measure relative to other children of the same age. In contrast, *criterion-referenced tests* are used to document specific skills rather than to compare people. Criterion-referenced tests are the basis for standards-based assessment in public education. Standards-based assessment requires that students pass tests demonstrating that they have critical knowledge and skills in defined areas. Some critics believe the cut points for passing the tests are arbitrary.

In summary, this chapter reviewed basic statistical methods for describing scores on one variable. In Chapter 3, we shall discuss statistical methods for showing the relationship between two or more variables.