APPENDIX

Statistics: The Description, Organization, and Interpretation of Data

In Chapter 1, we considered how psychologists gather data—how they design a study or an experiment, how they ensure external and internal validity, and so on. But what do they do once the data are gathered? In this appendix, we will focus on the statistical methods investigators use to organize and interpret numerical data.

Let us begin with an example. Suppose some investigators want to find out whether three-year-old boys are more physically aggressive than three-year-old girls. To find out, the investigators will first have to come up with some appropriate measure of physical aggression. They will then have to select the participants for the study. Since the investigators presumably want to say something about three-year-olds in general, not just the particular three-year-olds in their study, they must select their participants appropriately. Even more important, they must select boys and girls who are well matched to each other in all regards except gender, so that the investigators can be reasonably sure that any differences between the two groups are attributable to the difference in sex rather than to other factors (such as intellectual development, social class, and so on).

We discussed in Chapter 1 how investigators design studies and collect data. So we'll start here with what investigators do once their data have been collected. Their first task is to organize these data in a meaningful way. Suppose the study used two groups of 50 boys and 50 girls, each observed on 10 separate occasions. This means that the investigators will end up with at least 1,000 separate numerical entries, 500 for the boys and 500 for the girls. Something has to be done to reduce this mass of numbers into some manageable form. This is usually accomplished by some process of averaging scores.

The next step involves statistical interpretation. Suppose the investigators find that the average score for physical aggression is greater for the boys than for the girls. (It probably will be.) How should this fact be interpreted? Should it be taken seriously, or might it just be a fluke, some sort of accident? For it is just about certain that the data contain *variability*: the children within each group will not perform identically to each other; furthermore, the same child may very well behave differently on one occasion than on another. Thus, the number of aggressive acts for the boys might be, say, 5.8 on average, but might vary from a low of 1.3 (the score from completely calm Calvin) to a high of 11.4 (the score from awfully aggressive Albert). The average number of aggressive acts for the girls might be 3.9 (and so lower than the boys' average), but this derives from a range of scores that include 0 (from serene Sarah) and 6.2 (from aggressive Agnes).

Is it possible that this difference between boys and girls is just a matter of chance, an accidental by-product of this variability? For example, what if boys and girls are, in fact, rather similar in their levels of aggression, but—just by chance—the study happened to include four or five extremely aggressive boys and a comparable number of extremely unaggressive girls? After all, we know that our results would have been different if Albert had been absent on the day of our testing; the boys' average, without his contribution, would have been lower. Likewise, Agnes's twin sister was not included in our test group because of the random process through which we selected our research participants. If she had been included, and if she was as aggressive as her twin, then the girls' average would have been higher. Is it possible that accidents like these are the real source of the apparent difference between the groups? If so, then another study, without these same accidents, might yield a different result. One of the main reasons for using statistical methods is to deal with questions of this sort, to help us draw useful conclusions about behavior despite the unavoidable variability, and, specifically, allowing us to ask in a systematic way whether our data pattern is reliable (and so would emerge in subsequent studies) or just the product of accidents.

DESCRIBING THE DATA

In the example above, we assumed that the investigators would be collecting numerical data. We made this assumption because much of the power of statistics results from the fact that numbers can be manipulated using the rules of arithmetic, unlike open-ended responses in an interview, videotapes of social interactions, or lists of words recalled in a memory experiment. (How could you average together one participant's response of "Yes, I like them" with another's response of "Only on weekends"?) As a result, scientists prefer to use numerical response measures whenever possible. Consider our hypothetical study of physical aggression. The investigators who watched the research participants might rate their physical aggression in various situations from 1 to 5, with 1 being "extremely docile" and 5 being "extremely aggressive," or they might count the number of aggressive acts (say, hitting or kicking another child). This operation of assigning numbers to observed events is called *scaling*.

There are several types of scales that will concern us. They differ by the arithmetical operations that can be performed on them.

Categorical and Ordinal Scales

Sometimes the scores assigned to individuals are merely *categorical* (also called *nominal*). For example, when respondents to a poll are asked to name the television channel they watch most frequently, they might respond "4," "2," or "13." These numbers serve only to group the responses into categories. They can obviously not be subjected to any arithmetic operations. (If a respondent watches channels 2 and 4 equally often, we can't summarize this by claiming that, on average, she watches channel 3!)

Ordinal scales convey more information, in that the relative magnitude of each number is meaningful—not arbitrary, as in the case of categorical scales. If individuals are asked to list the ten people they most admire, the number 1 can be assigned to the most admired person, 2 to the runner-up, and so on. The smaller the number assigned, the more the person is admired. Notice that no such statement can be made of television channels: channel 4 is not more anything than channel 2, just different from it.

Scores that are ordinally scaled cannot, however, be added or subtracted. The first two persons on the most-admired list differ in admirability by 1; so do the last two. Yet the individual who has done the ranking may admire the first person far more than the other nine, all of whom might be very similar in admirability. Imagine, for example, a child who, given this task, lists his mother first, followed by the starting lineup of the Chicago Cubs. In this case, the difference between rank 1 and rank 2 is enormous; the difference between rank 2 and rank 3 (or any other pair of adjacent ranks) is appreciably smaller. Or, to put it another way, the difference of eight between person 2 and person 10 probably represents a smaller difference in judged admirability than the difference of one obtained between persons 1 and 2 (at least so the mother hopes).

Interval Scales

Scales in which equal differences between scores, or intervals, can be treated as equal units are called *interval scales*. Response time is a common psychological variable that is usually treated as an interval scale. In some memory experiments, for example, the participant must respond as quickly as possible to each of several words, some of which she has seen earlier in the experiment; the task is to indicate, by pressing the appropriate button, whether each word had appeared earlier or not.

Suppose that someone requires an average of 2 seconds to respond to nouns, 3 seconds to verbs, and 4 seconds to adjectives. The difference in decision time between verbs and nouns ($3 - 2 = 1$ second) is the same as the difference in decision time between adjectives and verbs ($4 - 3 = 1$ second). We can make this statement—which in turn suggests various hypotheses about the factors that underlie such differences—precisely because response time can be regarded as an interval scale.

Ratio Scales

Scores based on an interval scale allow subtraction and addition. But they do not allow multiplication and division. Consider the Celsius scale of temperature. The difference between 10 and 20 degrees Celsius is equal to that between 30 and 40 degrees Celsius. But can one say that 20 degrees Celsius is twice as high a temperature as 10 degrees Celsius? The answer is no, for the Celsius scale of temperature is only an interval scale. It is not a *ratio scale*, which allows statements such as 10 feet is one-fifth as long as 50 feet, or 15 pounds is three times as heavy as 5 pounds. To make such statements, one needs a true zero point. Such a ratio scale with a zero point does exist for temperature—the Kelvin absolute temperature scale, whose zero point (*absolute zero* to chemists and physicists) is about −273 degrees Celsius.

Some psychological variables can be described by a ratio scale. For example, it does make sense to say that the rock music emanating from your neighbor's dorm room is four times as loud as your roommate singing in the shower. But there are many psychological variables that cannot be described in ratio terms. For example, let us say that we assemble a list of behaviors commonly associated with clinical depression, and we find that, say, Person 1 displays 8 of these behaviors, while Person 2 displays 16 of them. We could legitimately say that there is a difference of 8 behaviors here—this is an interval scale. But we should not say that Person 2's score is twice as worrisome as that of Person 1, because we really don't know the zero point for this scale. More specifically, what we need to know is how many of these behaviors can be observed in people who do not suffer from depression. If we knew that people without depression showed none of these behaviors, then zero would be the true starting point for our scale (and so, in this scenario, it would appear that Person 2 does have twice as many of the relevant behaviors as Person 1). But if we found that people without depression showed 7 of these behaviors, then that would be the starting point for our scale (and so Person 1, with only 1 behavior more than this starting point, would appear to be vastly better off than Person 2, with 9 behaviors beyond the starting point).

ORGANIZING THE DATA

We have considered the ways in which psychologists describe the data provided by their studies by assigning numbers to them (scaling). Our next task is to see how these data are organized.

The Frequency Distribution

Suppose that an investigator wanted to determine whether visual imagery aids memory. (See Chapter 7 for some actual research on this topic.) To find out he designed an experiment that required participants to memorize a list of words and later to recall as many of these words as possible. Members of the experimental group were instructed to form visual images connecting each word to the preceding word. Members of the control group were not given any imagery instructions. Let us say that there are ten people in each group, and so the scores from the control group might have been

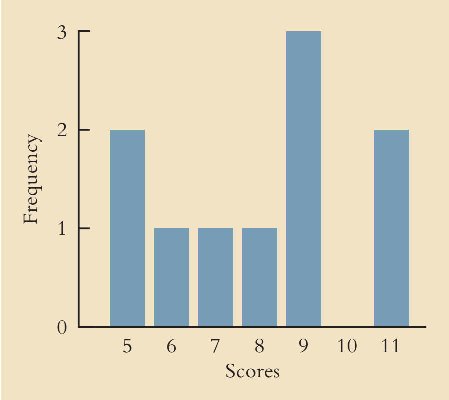
8, 11, 6, 7, 5, 9, 5, 9, 9, 11.

TABLE A.1 Frequency Distribution

Score	Frequency
11	2
10	0
9	3
8	1
7	1
6	1
5	2

A first step in organizing these data is to list all the possible scores and the frequencies with which they occurred, as shown in Table A.1. Such an arrangement is called a *frequency distribution* because it shows the frequency with which each number of words was recalled (e.g., how many of the participants recalled 11 words, how many recalled 10 words, and so on).

The frequency distribution can also be expressed graphically. A common means for doing this is a *histogram*, which uses a series of rectangles to depict the frequency distribution (Figure A.1). The values of the dependent variable (the number of words recalled) are shown by the location of each rectangle on the *x*-axis. The frequency of each score is shown by the height of each rectangle, as measured on the *y*-axis. This is simple enough for our example, but in practice graphic presentation often requires a further step. The number of possible values the dependent variable can assume is often very large. As a result, it is possible that every specific score in the data list appears just once! For example, in a response-time study, there might be only one response in the entire data set that took exactly 224.01 milliseconds, just one that took exactly 224.02 milliseconds, and so on. If the investigator created a frequency distribution showing how often each score occurred, the resulting histogram would be very wide (with many rectangles), very flat (since all rectangles would have the same height), and not at all informative. To get around this, it is common for investigators to group together similar observations, and this is usually done by dividing the data into "bins." Thus, the histogram might plot the frequency of observing a response time between, say, 200 and 225 milliseconds (that would be one bin), the frequency of observing a time between 225.01 and 250 milliseconds, and so on.



A.1 Histogram In a histogram, a frequency distribution is graphically represented by a series of rectangles. The location of each rectangle on the x-axis indicates a score, while its height shows how often that score occurred.

Measures of Central Tendency

For many purposes we want a description of an experiment's result that is more concise than a frequency distribution. We might, for example, wish to describe how a typical or average participant behaved. This sort of data summary is provided by a *measure of central tendency*, which locates the center of the distribution. Three measures of central tendency are commonly used: the *mode*, the *median*, and the *mean*.

The mode is simply the score that occurs most frequently. In our example, the mode for the control group is 9. More people (to be exact, 3) recalled 9 words than recalled any other number of words.

The median is the point that divides the distribution into two equal halves, when the scores are arranged in increasing order. To find the median in our example, we first list the scores:

$$5,\ 5,\ 6,\ 7,\ 8,\underset{\uparrow}{\ }\ 9,\ 9,\ 9,\ 11,\ 11$$

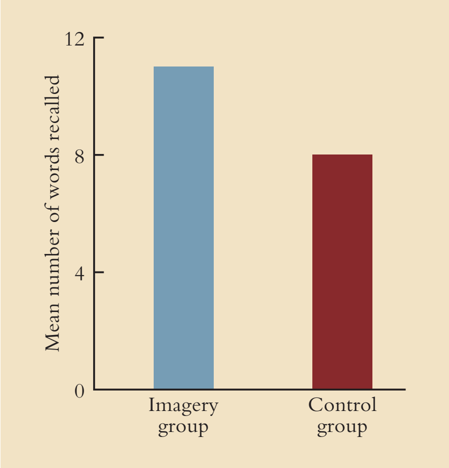
Since there are ten scores, the median lies between the fifth and sixth scores, that is, between 8 and 9, as indicated by the arrow. Any score between 8 and 9 would divide the distribution into two equal halves, but it is conventional to choose the number in the center of the interval between them, that is, 8.5. When there is an odd number of scores this problem does not arise, and the middle number is used.

The third measure of central tendency, the mean (M), is the familiar arithmetic average. If *N* stands for the number of scores, then

$$M = \frac{\text{sum of scores}}{N}$$

$$= \frac{5+5+6+7+8+9+9+9+11+11}{10} = \frac{80}{10} = 8.0$$

The mean is the measure of central tendency most commonly used by psychologists, in part because a number of further calculations can be based on this measure. It is common, therefore, for the results of experiments like our imagery example to be displayed as shown in Figure A.2. The values of the independent variable (in this case, getting imagery instructions) are indicated on the *x*-axis, and the values of the dependent variable (mean number of words recalled) on the *y*-axis.



A.2 The results of an experiment on memorizing Participants in the imagery group, who were asked to form visual images of the words they were to memorize, recalled an average of 11 words. Participants in the control group, who received no special instructions, recalled an average of 8 words.

Despite the common use of the mean, each of these measures of central tendency has its own advantages. The mode is used relatively rarely, because the modes of two samples can differ greatly even if the samples have very similar distributions. If one of the 3 participants who recalled 9 words recalled only 5 instead, the mode would have been 5 rather than 9. But the mode does have its uses. For example, a home builder might decide to include a two-car garage on a new house because 2 is the mode for the number of cars owned by American families; more people will be content with a two-car garage than with any other size.

The median and the mean differ most in the degree to which they are affected by extreme scores. If the highest score in our sample were changed from 11 to 111, the median would be unaffected, whereas the mean would jump from 8.0 to 18.0. Most people would find the median (which remains 8.5) a more compelling "average" than the mean in such a situation, since most of the scores in the distribution are close to the median but are not close to the mean (18.0). This is why medians are often preferred when the data become highly variable, even though the mean has computational advantages.

The advantages of the median become particularly clear with distributions of scores that contain a few extreme values. Such distributions are said to be *skewed*, and a classic example is income distribution, since there are only a few very high incomes but many low ones. Suppose we sample ten individuals from a neighborhood and find their yearly incomes (in thousands of dollars) to be

10, 12, 20, 20, 40, 40, 40, 80, 80, 4,000

The median income for this sample is 40 ($40,000), since both the fifth and sixth scores are 40. This value reflects the income of the typical individual. The mean income for this sample, however, is (10 + 12 + 20 + 20 + 40 + 40 + 40 + 80 + 80 + 4,000)/10 = 418, or $418,000. A politician who wants to demonstrate that her neighborhood has prospered might—quite accurately—use these data to claim that the average (mean) income is $418,000. If, on the other hand, she wished to plead for financial relief, she might say—with equal accuracy—that the average (median) income is only $40,000. There is no single "correct" way to find an "average" in this situation, but it is obviously important to know which average (that is, which measure of central tendency) is being used.

When deviations in either direction from the mean are equally frequent, the distribution is said to be *symmetrical*. In such distributions, the mean and the median are likely to be close to each other in actual value, and so either can be used in describing the data. Many psychological variables have symmetrical distributions, but for variables with skewed distributions, like income, measures of central tendency must be chosen with care.

Measures of Variability

In reducing an entire frequency distribution to an average score, we have discarded a lot of very useful information. Suppose the National Weather Service measures the temperature every day for a year in various cities and calculates a mean for each city. This tells us something about the city's climate, but certainly does not tell us everything. This is shown by the fact that the mean temperature in both San Francisco and Albuquerque is 56 degrees Fahrenheit. But the climates of the two cities differ considerably, as indicated in Table A.2.

The weather displays much more variability in the course of a year in Albuquerque than in San Francisco, but, of course, this variability is not reflected in the mean. One

TABLE A.2 Temperature Data for Two Cities (Degrees Fahrenheit)

City	Lowest Month	Mean	Highest Month	Range
Albuquerque, New Mexico	35	56	77	42
San Francisco, California	48	56	63	15

way to measure this variability is the *range*, the highest score minus the lowest. The range of temperatures in San Francisco is 15, while in Albuquerque it is 42.

A shortcoming of the range as a measure of variability is that it reflects the values of only two scores in the entire sample. As an example, consider the following distributions of ages in two college classes:

Distribution *A*: 19, 19, 19, 19, 19, 20, 25

Distribution *B*: 17, 17, 17, 20, 23, 23, 23

Intuitively, distribution *A* has less variability, since all scores but one are very close to the mean. Yet the range of scores is the same (6) in both distributions. The problem arises because the range is determined by only two of the seven scores in each distribution.

A better measure of variability would incorporate every score in the distribution rather than just two. One might think that the variability could be measured by asking how far each individual score is away from the mean, and then taking the average of these distances. This would give us a measure that we could interpret (roughly) as ìon average, all the data points are only two units from the mean (or "... three units ..." or whatever it turned out to be). The most straightforward way to measure this would be to find the arithmetic difference (by subtraction) between each score and the mean (that is, computing [score − M] for each score), and then to take the average of these differences (that is, add up all of these differences, and divide by the number of observations):

$$\frac{\text{sum of (score} - \text{M)}}{N}$$

This hypothetical measure is unworkable, however, because some of the scores are greater than the mean and some are smaller, so that the numerator is a sum of both positive and negative terms. (In fact, it turns out that the sum of the positive terms equals the sum of the negative terms, so that the expression shown above always equals zero.) The solution to this problem is simply to square all the terms in the numerator, thus making them all positive.* The resulting measure of variability is called the *variance* (*V*):

$$\text{V} = \frac{\text{sum of (score} - \text{M)}^2}{N} \qquad (1)$$

* An alternative solution would be to sum the *absolute value* of these differences, that is, to consider only the magnitude of this difference for each score, not the sign. The resulting statistic, called the average deviation, is little used, however, primarily because absolute values are not easily dealt with in certain mathematical terms that underlie statistical theory. As a result, statisticians prefer to transform negative into positive numbers by squaring them.

TABLE A.3 Calculating Variance

Score	Score – Mean	(Score – Mean)^2
8	$8 - 8 = 0$	$0^2 = 0$
11	$11 - 8 = 3$	$3^2 = 9$
6	$6 - 8 = -2$	$(-2)^2 = 4$
7	$7 - 8 = -1$	$(-1)^2 = 1$
5	$5 - 8 = -3$	$(-3)^2 = 9$
9	$9 - 8 = 1$	$1^2 = 1$
5	$5 - 8 = -3$	$(-3)^2 = 9$
9	$9 - 8 = 1$	$1^2 = 1$
9	$9 - 8 = 1$	$1^2 = 1$
11	$11 - 8 = 3$	$3^2 = 9$
		sum $= 44$

$$\mathrm{V} = \frac{\text{sum of (score} - \text{mean)}^2}{N} = \frac{44}{10} = 4.4$$

The calculation of the variance for the control group in the word-imagery experiment is shown in Table A.3. As the table shows, the variance is obtained by subtracting the mean (M, which equals 8) from each score, squaring each result, adding all the squared terms, and dividing the resulting sum by the total number of scores (N, which equals 10), yielding a value of 4.4.

Because deviations from the mean are squared, the variance is expressed in units different from the scores themselves. If our dependent variable were a distance, measured in centimeters, the variance would be expressed in square centimeters. As we will see in the next section, it is convenient to have a measure of variability that can be added to or subtracted from the mean; such a measure ought to be expressed in the same units as the original scores. To accomplish this end, we employ another measure of variability, the *standard deviation*, or *SD*. The standard deviation is derived from the variance by taking the square root of the variance. Thus

$$\mathrm{SD} = \sqrt{\mathrm{V}}$$

In our example, the standard deviation is about 2.1, the square root of the variance which is 4.4.

Converting Scores to Compare Them

Suppose a person takes two tests. One measures his memory span—how many digits he can remember after one presentation. The other test measures his reading speed—how quickly he can read a 200-word essay. It turns out that he can remember 8 digits and needs 140 seconds for the essay. Is there any way to compare these two numbers, to decide whether he can remember digits as well (or worse or equally well) as he can read? On the face of it, the question seems absurd; it seems like comparing apples and oranges. But for some purposes, we would want to compare these numbers. For example, a first step toward identifying people with dyslexia is documenting that their reading ability is markedly lower than we would expect, based on their intellectual performance in other areas. For this purpose, a comparison much like the one just sketched might be useful. But how do we compare digits-remembered to number-of-seconds-needed-for-reading?

In fact, there is a way to make this comparison, starting with an assessment of how each of these two scores compares to the scores of other persons who have been given the same two tests.

PERCENTILE RANKS

One way of doing this is by transforming each of the two scores into a *percentile rank*. The percentile rank of a score indicates the percentage of all scores that lie below that given score. Let us assume that 8 digits is the 78th percentile. This means that 78 percent of the relevant comparison group remembers fewer digits. Let us further assume that a score of 140 seconds in the reading task is the 53rd percentile of the same comparison group. We can now answer the question with which we started. This person can remember digits more effectively than he can read. By converting into percentile ranks we have rendered incompatible scores compatible, allowing us to compare the two.

STANDARD SCORES

For many statistical purposes there is an even better method of comparing scores or of interpreting the meaning of individual scores. This is to express them by reference to the mean and standard deviation of the frequency distribution of which they are a part. This is done by converting the individual scores into *standard scores* (often called *z-scores*). The formula for calculating a *z-score* is:

$$z = \frac{(\text{score} - \text{M})}{\text{SD}} \tag{2}$$

Suppose you take a test that measures aptitude for accounting and are told your score is 36. In itself, this number cannot help you decide whether to pursue or avoid a career in accounting. To interpret your score you need to know both the average score and how variable the scores are. If the mean is 30, you know you are above average, but how far above average is 6 points? This might be an extreme score or one attained by many, depending on the variability of the distribution.

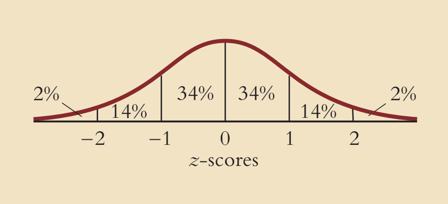
Let us suppose that the standard deviation of the distribution is 3. Your *z*-score on the accounting test is $(36 - 30)/3 = +2$. That is, your score is 2 SDs above the mean.

But how to use this information? Let us say that you are still unsure whether to become an accountant, and so you take a screen test to help you decide whether to become an actor instead. Here, your score is 100. This is a larger number than the 36 you scored on the earlier test, but it may not reveal much acting aptitude. Suppose the mean score on the screen test is 80, and the standard deviation is 20; then your *z*-score is $(100 - 80)/20 = +1$. In acting aptitude, you are 1 SD above the mean (that is, $z = +1$)—above average but not by much. In accounting aptitude, you are 2 SDs above the mean (that is, $z = +2$), and so the use of *z*-scores makes your relative abilities clear.

Percentile rank and a *z*-score give similar information, but, to convert one into the other, we need a bit more information.

The Normal Distribution

Frequency histograms can have a wide variety of shapes, but many variables that interest psychologists have a *normal distribution* (often called a *normal curve*), which is a



A.3 Normal distribution Values taken from any normally distributed variable (such as those presented in Table A.4) can be converted to z-scores by the formula z = (score – M)/(SD). The figure shows graphically the proportions that fall between various values of z.

TABLE A.4 Normally Distributed Variables

Variable	Mean	Standard Deviation	Values Correspoinding to Specific z-scores				
			−2	−1	0	1	2
IQ	100	15	70	85	100	115	130
SAT	500	100	300	400	500	600	700
Height (women)	160 cm	5 cm	150	155	160	165	170

symmetrical distribution of the shape shown in Figure A.3. (For more on normal curves, see Chapter 14.) The graph is smooth, unlike the histogram in Figure A.1, because it describes the distribution of scores from a very large sample. The normal curve is bell shaped, with most of its scores near the mean; the farther a score is from the mean, the less likely it is to occur. Among the many variables whose distributions are approximately normal are IQ, scholastic aptitude test (SAT) scores, and women's heights (see Table A.4).*

These three variables—IQ, SAT score, and height—obviously cannot literally have the same distribution, since their means and standard deviations are different (Table A.4 gives plausible values for them). In what sense, then, can they all be said to be normally distributed? The answer is that the shape of the distributions for all these variables is the same. For example, an IQ of 115 is 15 points, or 1 SD, above the IQ mean of 100; a height of 165 centimeters is 5 centimeters, or 1 SD, above the height mean of 160 centimeters. Both scores, therefore, have *z*-scores of 1. And crucially, the percentage of heights between 160 and 165 centimeters is the same as the percentage of IQ scores between 100 and 115, that is, 34 percent. This is true not just for these two variables, but in general: it is the percentage of scores that lie between the mean and 1 SD above the mean for any normally distributed variable.

THE PERCENTILE RANK OF A *Z*-SCORE

In fact, this point can be put more generally: each normal curve has its own mean and its own standard deviation. But all normal curves have the same shape, and, as a result, the percentage of scores that fall between the mean and +1 standard deviation (and so have *z*-scores between 0 and 1.0) is always the same: 34 percent. Likewise, for all normal curves, the percentage of the scores that fall between +1 standard deviation and +2 standard deviations (and so have *z*-scores between 1.0 and 2.0) is always the same: 14 percent. And, since normal curves are symmetrical, the same proportions hold for below the mean (and so 34 percent of the scores have *z*-scores between 0 and −1, and so on). These relationships are illustrated in Figure A.3.

These facts allow us to convert any *z*-score directly into a percentile rank. A *z*-score of 1 has a percentile rank of 84. That is, 84 percent of all the scores are below this particular score. (This is true because 34 percent of the scores lie between the mean and

* Men's heights are also normally distributed, but the distribution of the heights of *all* adults is not. Such a distribution would have two peaks, one for the modal height of each sex, and would thus be shaped quite differently from the normal curve. Distributions with two modes are called *bimodal.*

$z = 1$, and 50 percent of the scores lie blow the mean). Likewise, a z-score of -1 (1 SD below the mean) corresponds, in a normal distribution, to a percentile rank of 16: only 16 percent of the scores are lower. And so on.

HOW THE NORMAL CURVE ARISES

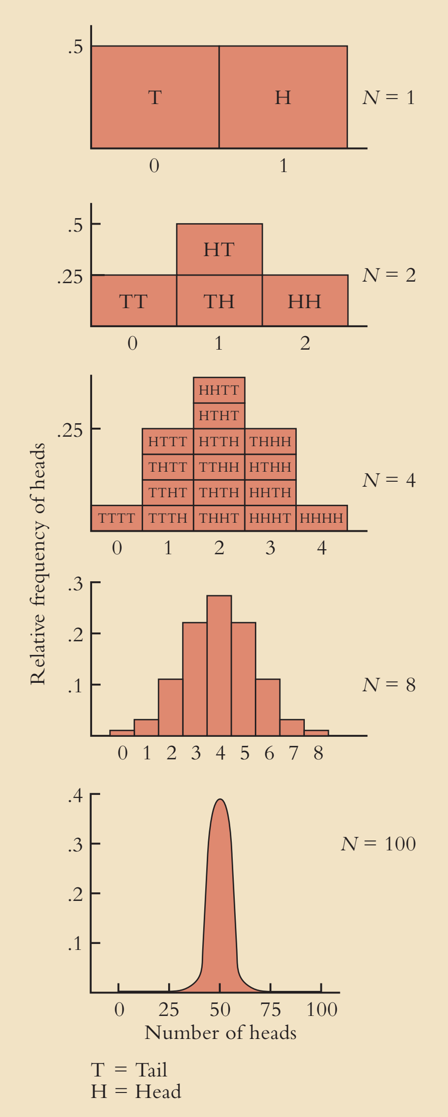
Why should variables such as height or IQ (and many others) form distributions that have this particular shape? Mathematicians have shown that whenever a given variable is the sum of many smaller variables, its distribution will be close to that of the normal curve. One example is lifetime earnings—obviously the sum of what one has earned on many prior occasions. A different example is height. Height can be thought of as the sum of the contributions of the many genes and the many environmental factors that influence this trait; it, therefore, satisfies the general condition. The basic idea is that the many different factors that influence a given measure (such as the genes for height) operate independently of the others, and, for each of these factors, it is a matter of chance whether the factor applies to a particular individual or not. Thus, if someone's father had a certain height-promoting gene on one chromosome but not on the other chromosome in the pair, then it would literally be a matter of chance whether the person inherited this gene or not (and likewise for each of the other genes—and surely there are many—that determine height). The person's height would also depend on accidents in his experience—for example, whether, just by bad luck, he happened to catch the flu at an age that interrupted what would have otherwise been a strong growth spurt.

In essence, then, we can think of each person's height as dependent on a succession of coin tosses, with each toss describing whether that person received the height-promoting factor or not—inherited the gene or not, got the flu at just the wrong time or not, and so on. Of course, each factor contributes its own increment to the person's height, and so his ultimate height depends on how many of these factors fell the right way. Thus, if we want to predict the person's height, we need to explore the (relatively simple) mathematics that describe how these chance events unfold.

Let us imagine that someone literally does toss a coin over and over, with each head corresponding to a factor that tends to increase height and each tail to a factor that tends to diminish it. Predicting the person's height, therefore, would be equivalent to predicting how many heads, in total, the person will obtain after a certain number of tosses. If the coin is tossed only once, then there will be either 0 heads or 1 head, and these are equally likely. The resulting distribution is shown in the top panel of Figure A.4.

If the number of tosses (which we will call N) is 2, then 0, 1, or 2 heads can arise. However, not all these outcomes are equally likely: 0 heads come up only if the sequence tail-tail (TT) occurs; 2 heads only if head-head (HH) occurs; but 1 head results from either HT or TH. The distribution of heads for $N = 2$ is shown in the second panel of Figure A.4. The area above 1 head has been subdivided into two equal parts, one for each possible sequence containing a single head.*

As N increases, the distribution of the number of heads looks more and more like the normal distribution, as the subsequent panels of Figure A.4 show. When N becomes as large as the number of factors that determine height, the distribution of the number of heads is virtually identical to the normal distribution, and this gives us just the claim we were after: as we have described, this logic of coin tossing corresponds reasonably well to the logic of the factors governing height, and so, just as the distribution of coin tosses will (with enough tosses) be normally distributed, so will height. The same logic applies to many other



A.4 Histograms showing expected number of heads in tossing a fair coin N times In successive panels, N = 1, 2, 4, and 8. The bottom panel illustrates the case when N = 100 and shows a smoothed curve.

* The distribution of the number of heads is called the *binomial distribution*, because of its relation to the binomial theorem: the number of head-tail sequences that can lead to k heads is the $(k + 1)$st coefficient of $(a + b)^N$.

measures of interest to psychologists—the distribution of people's intelligence or personality traits, the distribution of response times in an experimental procedure, the distribution of students' scores on a mid-term exam. These, too, are influenced by a succession of chance factors, and so, just like the coin tosses, they will be normally distributed.

DESCRIBING THE RELATION BETWEEN TWO VARIABLES: CORRELATION

So far, we have focused on how psychologists measure a single variable—what scales they use, how they measure the variable's average or its variability. In general, though, investigators want to do more than this—they want to ask how two (or more) variables are related to each other. Is there a relationship between the sex of a child (the independent variable) and how physically aggressive (the dependent variable) that child is? Is there a relationship between using visual imagery (the independent variable) and memory (the dependent variable)? One way to measure this relationship is by examining the *correlation* between the two variables.*

Positive and Negative Correlation

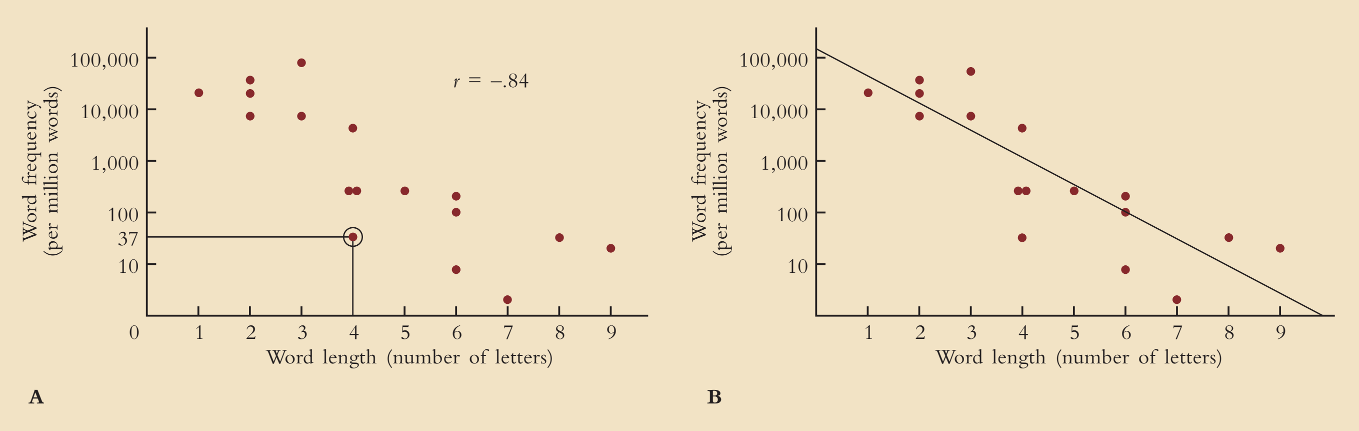
Imagine that a manager of a taxicab company wants to identify drivers who will earn relatively large amounts of money (for themselves and, of course, for the company). The manager makes the plausible guess that one relevant factor is the driver's knowledge of the local geography, so she devises an appropriate test of street names, routes from place to place, and so on, and administers the test to each driver. The question is whether this test score is related to the driver's job performance as measured by his weekly earnings. To decide, the manager has to find out whether there is a correlation between the test score and the earnings—that is, whether they tend to vary together.

In the taxicab example, the two variables will probably be positively correlatedóas the independent variable (test score) increases, the dependent variable (earnings) will generally increase too. But other variables may be negatively correlatedówhen one increases, the other will tend to decrease. An example is a phenomenon called Zipfís law, which states that words that occur frequently in a language tend to be relatively short. The two variablesóword length and word frequencyóare negatively correlated, since one variable tends to increase as the other decreases.

Correlational data are often displayed in a *scatter plot* (or *scatter diagram*) in which values of one variable are shown on the x-axis and variables of the other on the y-axis. Figure A.5A is a scatter plot of word frequency versus word length for the words in this sentence.** Each word is represented by a single point. An example is provided by the word *plot,* which is four letters long and occurs with a frequency of 37 times per million

* In Chapter 1, we contrasted experimental and correlational designs; correlational designs are those which exploit differences that exist independently of the investigator's manipulations. Thus a comparison between boys and girls is a correlational design (because the sex difference certainly exists independently of the investigator's procedures), and so is a comparison between, say, young children and old children. All of this is different from the *statistical technique* that computes correlations. The statistic is just a specific means of exploring the relationship between two variables. Correlational designs often use correlational statistics, but often do not.

** There is no point for the "word" A.5A in this sentence. The frequencies of the other words are taken from H. Kucera and W. N. Francis, *Computational Analysis of Present-Day American English* (Providence, R. I.: Brown University Press, 1967).



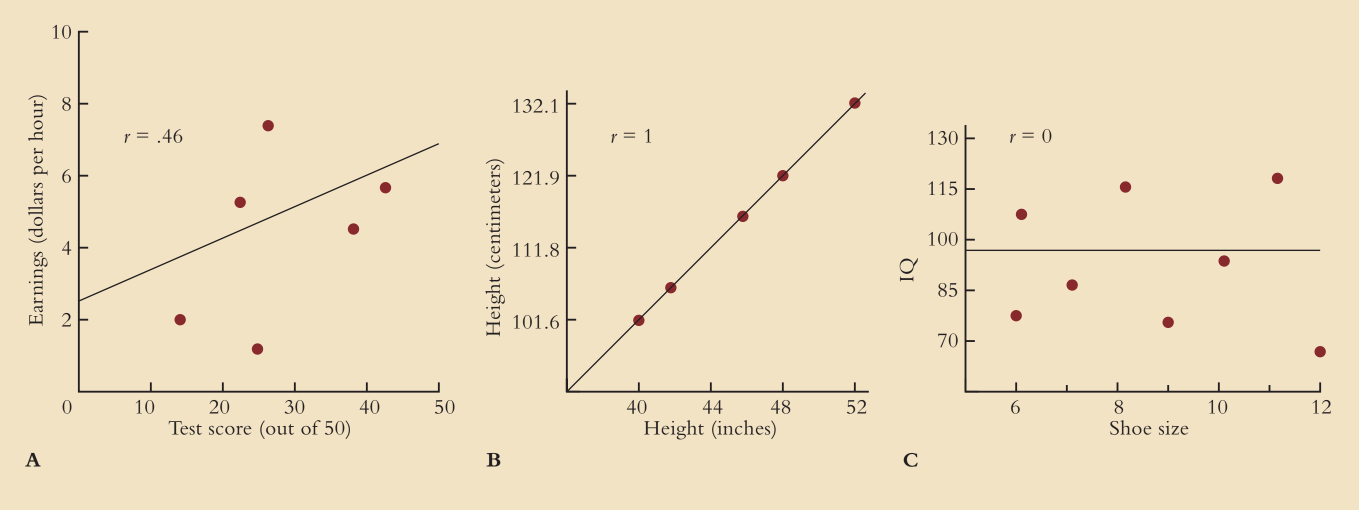
A.5 Scatter plot of a negative correlation between word length and word frequency.

words of English text (and is represented by the circled dot). The points on the graph display a tendency to decrease on one variable as they increase on the other, although the relation is by no means perfect.

It is helpful to draw a line through the various points in a scatter plot that comes as close as possible to all of them (Figure A.5B). The line is called a *line of best fit*, and it indicates the general trend of the data. Here, the line slopes downward because the correlation between the variables is negative.

The three panels of Figure A.6 are scatter plots showing the relations between other pairs of variables. In Figure A.6A hypothetical data from the taxicab example show that there is a positive correlation between test score and earnings (since the line of best fit slopes upward). Test score is not a perfect predictor of on-the-job performance, however, since the points are fairly widely scattered around the line. Points above the line represent individuals who earn more than their test score would lead one to predict; points below the line represent individuals who earn less.

The examples in Figures A.5 and A.6A illustrate moderate correlations; in contrast, panels B and C of Figure A.6 illustrate extreme cases. Figure A.6B shows data



A.6 Scatter plots of various correlations (A) The scatter plot and line of best fit show a positive correlation between a taxi-driving test and earnings. (B) A perfect positive correlation. The line of best fit passes through all the points. (C) A correlation of zero. The line of best fit is horizontal.

from a hypothetical experiment conducted in a fourth-grade class to illustrate the relation between metric and English units of length. The heights of five children are measured twice, once in inches and once in centimeters; each point on the scatter plot gives the two height measurements for one child. All the points in the figure fall on the line of best fit, because height in centimeters always equals 2.54 times height in inches. The two variables, height in centimeters and height in inches, are perfectly correlatedóone can be perfectly predicted from the other. Thus, once you know your height in inches, there is no information to be gained by measuring yourself in centimeters.

Figure A.6C presents a relation between IQ and shoe size. These variables are unrelated to each other; people with big feet have neither a higher nor a lower IQ than people with small feet. The line of best fit is therefore horizontal: The best guess of an individualís IQ is the same no matter what his or her shoe sizeóit is the mean IQ of the population.

The Correlation Coefficient

Correlations are usually described by a *correlation coefficient*, denoted *r*, a number that expresses the strength and the direction of the correlation. For positive correlations, *r* is positive; for negative correlations, it is negative; for variables that are completely uncorrelated, *r* equals 0. The largest positive value *r* can have is +1.00, which represents a perfect correlation (as in Figure A.6B); the largest possible negative value is −1.00, which is also a perfect correlation. The closer the points in a scatter plot come to falling on the line of best fit, the nearer *r* will be to +1.00 or −1.00 and the more confident we can be in predicting scores on one variable from scores on the other. The values of *r* for the scatter plots in Figures A.5 and A.6A are given on the figures.

The method for calculating *r* between two variables, *X* and *Y*, is shown in Table A.5 (on the next page). The formula is:

$$r = \frac{\text{sum } (z_x z_y)}{N} \qquad (3)$$

The variable z_x is the *z*-score corresponding to *X*; z_y is the *z*-score corresponding to *Y*. To find *r*, each *X* and *Y* score must first be converted to a *z*-score by subtracting the mean for that variable and then dividing by the standard deviation for the variable. Then the product of z_x and z_y is found for each pair of scores. The average of these products (the sum of the products divided by *N*, the number of pairs of scores) is *r*.

Interpreting and Misinterpreting Correlations

It is tempting to assume that if two variables are correlated, then one is the cause of the other. This certainly seems plausible in our taxicab example, in which greater knowledge of local geography would improve the driver’s performance, which in turn would lead to greater earnings. Cause-and-effect relationships are also reflected in other real-life correlations. There is, for example, a correlation between how much loud music you listen to as an adolescent and the sensitivity of your hearing in later life. (The correlation is negative—more loud music is associated with less sensitive hearing.) And, in fact, there is a causal connection here, because listening to loud music can damage your hearing. Similarly, there is a correlation between the vividness of your visual imagery while awake and how often you remember your dreams on

TABLE A.5

Calculation of the Correlation Coefficient

1. Data (from Figure A.6A).

Test score (X)	*Earnings* (Y)
45	6
25	2
15	3
40	5
25	6
30	8

2. Find the mean and standard deviation for X and Y.

For X, mean = 30, standard deviation = 10
For Y, mean = 5, standard deviation = 2

3. Convert each X and each Y to a z-score, using $z = \dfrac{(\text{score} - \text{M})}{SD}$

X	Y	z-*score for* X (z_x)	z-*score for* Y (z_y)	$z_x z_y$
45	6	1.5	0.5	0.75
25	2	−0.5	−1.5	0.75
15	3	−1.5	−1.0	1.50
40	5	1.0	0.0	0.00
25	6	−0.5	0.5	−0.25
30	8	0.0	1.5	0.00
				2.75

4. Find the product $z_x z_y$ for each pair of scores.

5. $r = \dfrac{\text{sum }(z_x z_y)}{N} = \dfrac{2.75}{6} = .46$

awakening (Cory, Ormiston, Simmel, & Dainoff, 1975). This correlation is positive—greater vividness is associated with more frequent dream recall. And here, too, there may be a causal connection: vivid waking imagery creates a mental perspective similar to the nighttime experience of dreaming, and this similarity of perspective facilitates recall.

However, as we emphasized in Chapter 1 and again in many other contexts in this book, often a correlation does *not* indicate a cause-and-effect relationship, or, if it does, the direction of causation is ambiguous. For example, consider the negative correlation between obesity and life expectancy: people who are overweight tend to die younger than people who are not overweight. For many years, this was interpreted as a cause-and-effect relationship: being overweight caused early death. Newer evidence, however, suggests that this is incorrect. Instead, it turns out that obesity is often associated with inactivity, and inactivity is what causes the problems. Overweight people who are active actually have lower mortality rates than normal-weight people who are sedentary (Kampert, Blair, Barlow, & Kohl, 1996).

Thus, a correlation, by itself, cannot indicate a cause-and-effect relationship. Some correlations do indicate causation, but many do not. As a result, correlational results are important and instructive but must be interpreted with care.

INTERPRETING THE DATA

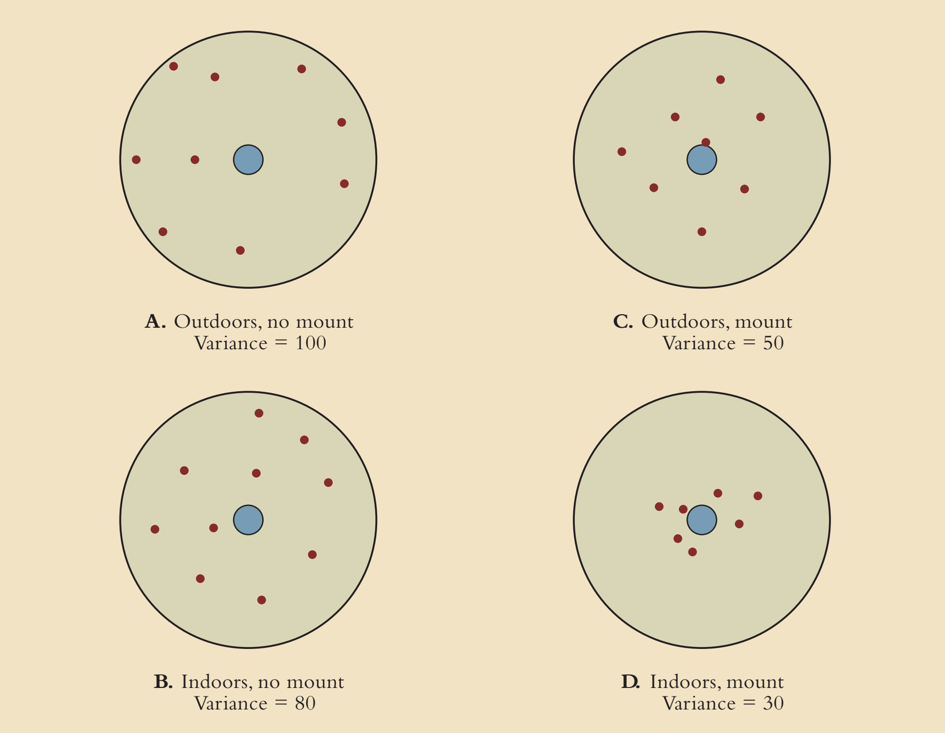
Any data collected in the real world contain variability, and data in psychology are no exception. In memory experiments, for example, different research participants recall different numbers of items, and the same participant is likely to perform differently if tested again later. But investigators nonetheless hope to draw general conclusions from data despite this variability. Nor is variability necessarily the enemy, because as we shall see, understanding the sources of variability in one's data can provide insights into the factors that influence the data.

Let us first consider how the pattern of variability can be used as a source of information concerning why the data are as they are. From this base, we will turn to the specific procedures that researchers use in implementing this logic, as they seek to ask whether their data are reliable or not and whether their data will support their conclusions or not. (Some readers may prefer to focus just on the procedures necessary for statistical analysis, rather than the underlying conceptualization; those readers can skip ahead to the heading, "Hypothesis testing.")

Accounting for Variability

As an example of how variability may be explained, consider a person shooting a pistol at a target. Although she always aims at the bull's-eye, the shots scatter around it (Figure A.7A). Assuming that the mean is the bull's-eye, the variance of these shots is the average squared deviation of the shots from the center. Suppose we find this variance to be 100; we next must explain it.

If the shooting was done outdoors, the wind may have increased the spread; moving the shooter to an indoor shooting range produces the tighter grouping shown in Figure A.7B. The new variance is 80, a reduction of 20 percent. This means that the wind accounts for 20 percent of the original variance.



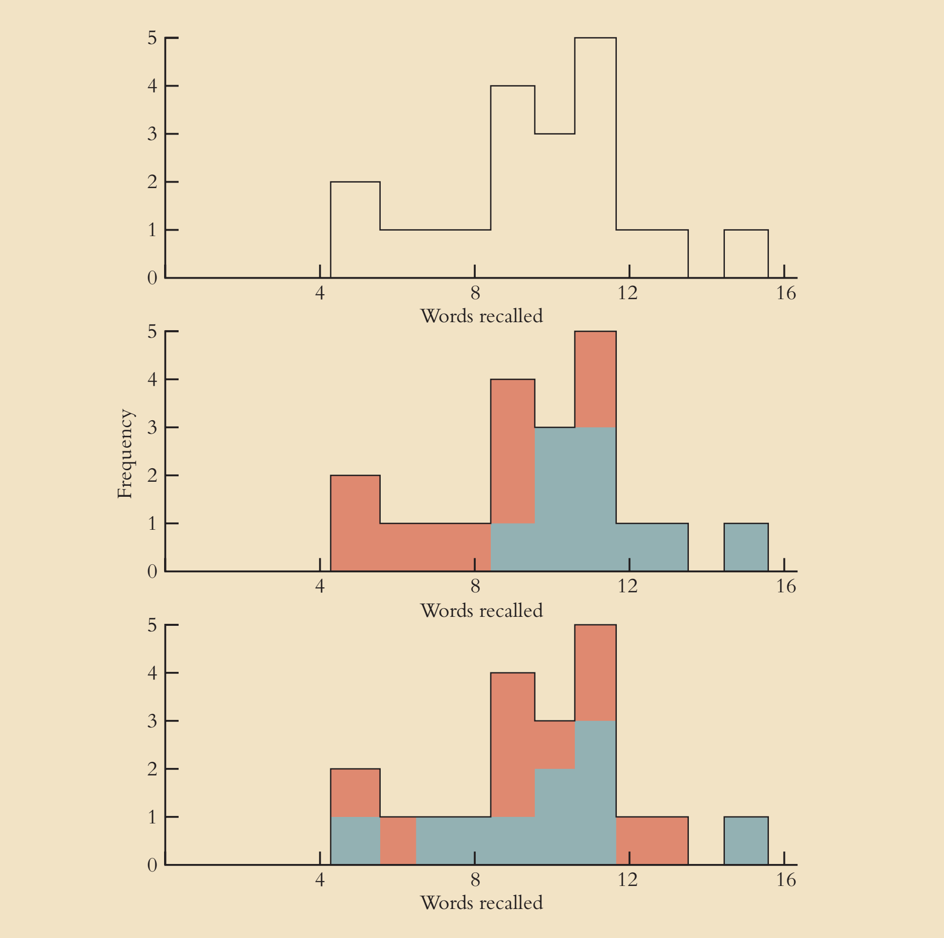
A.7 Results of target shooting under several conditions In each case, the bull's-eye is the mean, and the variance is the average squared deviation of the shots from the bull's-eye.

In addition, some of the initial variance may have resulted from the unsteady hand of the shooter, so we now mount the gun (although still leaving it outdoors). This yields a variance of 50 (Figure A.7C), a reduction of 50 percent. So 50 percent of the variance can be attributed to the shaky hand of the shooter. To find out how much of the variance can be explained by both the wind and the shaking, we mount the gun and move it indoors; now we may find a variance of only 30 (Figure A.7D). This means we have explained 70 percent of the variance, leaving 30 percent unaccounted for.*

But not all changes in the situation will reduce the variance. For example, if we find that providing the shooter with earmuffs leaves the variance unchanged, we know that none of the original variance was due to the noise of the pistol.

VARIANCE AND EXPERIMENTS

Figure A.8 shows how this approach can be applied to the experiment on visual imagery described earlier (see pp. A4–A5). Figure A.8A shows the distribution of scores for all twenty people in the experiment lumped together; the total variance of this overall distribution is 6.25. But as we saw, the ten members of the experimental group had been instructed to use visual imagery in memorizing, whereas the ten members of the control group were given no special instructions. How much of the overall variance can be explained by the difference in these instructions? In Figure A.8B, the distributions



A.8 **Accounting for variance in an experiment on memorizing** (A) The distribution of number of words recalled is shown for all twenty participants lumped together; the variance of this distribution is 6.25. (B) The distributions of the experimental and control groups are displayed separately. The number of words recalled by the group that received imagery instructions is shown in blue; the number recalled by the control group that received no special instructions is shown in red. Within each of these groups, the variance is about 4.00. (C) The distribution of words recalled is plotted separately for men and women regardless of how they were instructed. Blue indicates the number of words recalled by women, red the number recalled by men. The variance is 6.25.

* We are grateful to Paul Rozin for suggesting this example.

are no longer lumped together. They are instead presented as two separate histograms; the people who received imagery instructions are shown in blue, while those who did not are indicated in red. As the figure shows, there is less variability within either the imagery group or the control group than within the overall distribution that lumped both kinds of participants together. While the variance in the overall distribution is 6.25, the variance within the two subgroups averages to only 4.0. We conclude that the difference between the two sets of instructions accounted for 36 percent of the variance and that 64 percent (4 ÷ 6.25) still remains unexplained.

Figure A.8C shows a situation in which an independent variable (in this case, sex) accounts for little or none of the variance. In this figure, the participantsí scores are again presented as two histogramsóseparately depicting the scores of the men and the women (regardless of whether they were instructed to use imagery or not). The menís scores are shown in red, the womenís in blue. Now the variance of the two subgroups (that is, men versus women) averages to 6.25, a value identical to that found for the overall distribution. We conclude that the participantís sex accounts for none of the overall variance in recall.

VARIANCE AND CORRELATION

The technique of explaining the variance in one variable by attributing it to the effect of another variable can also be applied to correlational studies. Here, the values of one variable are explained (that is, accounted for) when the values of the other variable are known. Recall the taxicab example, in which a correlation of +.46 was found between taxi drivers' earnings and their scores on an aptitude test. Since the correlation is neither +1.00 nor 0, some but not all of the variance in job performance can be explained by the test scores. The greater the magnitude of *r*, the more variance is accounted for. The rule is that the proportion of variance that is explained equals r^2. If $r = +.46$, one variable accounts for $(.46)^2 = .21$ (21 percent) of the variance of the other. (Just why this proportion is r^2 is beyond the scope of this discussion.)

To put this another way, suppose all the cab drivers were identical in their performance on the aptitude test, which measured knowledge of local geography. This means that the variance on that variable would be zero. As a result, the variability on the second variable, earnings, would be reduced, and the formula tells us by how much. The original variance on earnings can be determined from the data in Figure A.6A. It is 4. Its correlation with the aptitude test is +.46. If we remove the variance caused by differences in how much the cab drivers know about local geography, the variability on earnings will be $4 - (.46)^2 \times 4 = 3.16$. The drop in the variance from 4 to 3.16 is a reduction of 21 percent. So the aptitude test does help us to predict taxicab earnings, for it accounts for 21 percent of the variance. But a good deal of the variance, 79 percent, is still unexplained.

Hypothesis Testing

The logic we have described—cast in terms of explaining the variance—lies at the heart of many techniques used for statistical analysis. For example, we saw that the participant's sex accounts for none of the overall variance in recall in our (hypothetical) imagery experiment; this is what tells us that we can reject the hypothesis that sex is relevant to performance in this task. Conversely, we saw that the variance is reduced if we divide the data according to experimental group (imagery group versus control group); this tells us that imagery is relevant here.

But how exactly is this logic put into practice? In this section, we tackle this question by means of some simple examples.*

Much behavioral research attempts to answer questions such as: Does the amount of food a person eats depend on the effort required to eat it? Can people learn while they are sleeping? Is drug *X* more effective than drug *Y*? Each of these questions suggests an experiment.

TESTING HYPOTHESES ABOUT SINGLE SCORES

We will begin by testing a hypothesis about single scores. Consider the problem in identifying people with dyslexia. As one step in identifying such people, we might give each person a test of reading comprehension. If the person's score was unusually low, this might be an indication of dyslexia (although several other tests would be needed to confirm this possibility). The question, though, is how low a score must be before it is "unusually low."

We know from the start that reading scores among nondyslexic readers vary—some read rather well, others read at some middle level, and some read rather poorly. As a result, it is possible that a poor reader is not dyslexic at all; he is simply at the low end of the normal range for reading skills. How can we evaluate this possibility?

Suppose we tested a large number of nondyslexic readers and found that the average reading score is 50, that the standard deviation of these scores is 10, and that the scores are normally distributed. We now look at the reading score from an individual we are concerned about. Let us say that her score is 40. How likely is it that she has dyslexia? This is equivalent to asking: How *un*likely is a score of 40 within the distribution of scores obtained by the general population (that is, a population of people who we believe are *not* dyslexic)? To answer these questions, we can convert her score to a *z*-score by computing its distance from the mean and dividing this difference by the standard deviation. The resulting *z*-score is (40 − 50)/10 or −1 SD. Since the distribution is normal, Figure A.3 tells us that 16 percent of the general population would score as low or even lower than this. Under the circumstances, it's plausible that a score of 40 does not indicate dyslexia; this score is common enough even among people without dyslexia. Our conclusion might be different, though, if the score were 30 or below. For then the *z*-score would be (30 − 50)/10 or −2, 2 SDs below the mean for the general population. Only 2 percent of the population obtain scores this low, and so we might now feel more comfortable concluding that a person with this particular score is likely not to be drawn from the general population. Instead, we might conclude that this score is likely to have been drawn from a *different* population—the population of people who do in fact suffer from dyslexia.

In this example we have to decide between two hypotheses about this individual's score. One hypothesis is that the score was drawn from the population of nondyslexic readers. True, a score of 40 or even 30 might seem atypical, but, on this view, this is

* The logic of explaining variance is crucial for most statistical procedures, but this logic turns out to be most visible with more complicated cases—for example, cases involving the comparison of two different groups (as in the example illustrated in Figure A.8), or the analysis of experiments in which two variables are manipulated. (For example, an experimenter might ask whether imagery instructions are as helpful for children as they are for adults; in this case, the experiment's design would have four groups: children and adults, and, then, within each of these groups, some participants given imagery instructions and some not). In the following pages, however, we have chosen to use simpler examples. This makes the underlying logic, in terms of explaining the variance, a bit less obvious, but it also makes the statistical procedures themselves much easier to grasp!

merely a reflection of the ordinary variability around the mean of the broader population. This is the *null hypothesis*, the hypothesis that there really is no systematic difference between the particular observation we are interested in and other observations we have made on other occasions and with other individuals. The alternative hypothesis is that the null hypothesis is *false* and that the score now before us is far enough away from the other scores for us to conclude that it did not arise by chance and is instead in a different category (in our example, the category of scores obtained by people with dyslexia).

As we have already suggested, the choice between these two hypotheses turns out to be a matter of probabilities. In essence, we start by adopting the working assumption that the null hypothesis is correct, and ask, within this assumption, what the probability would be of obtaining the score we have before us. If this probability—computed from the *z*-score—is relatively high (that is, if this score would be observed relatively often if the null hypothesis were correct), we conclude that the score poses no challenge to the null hypothesis, and so we accept the null hypothesis as probably correct. If, on the other hand, we start by assuming the null hypothesis, but then calculate that the score would be extremely rare under the null hypothesis, then we have two choices: either we have just observed an extremely rare event or the null hypothesis is false. Since the first of these choices is, by definition, very unlikely, we opt for the second choice.

With this logic, all depends on the *z*-score associated with our observation, and so, in the context of hypothesis testing, the *z*-score is referred to as the *critical ratio*. Behavioral scientists generally stipulate a critical ratio of 2 as the cutoff point. If it is 2 or more, they generally reject the null hypothesis and conclude that the test observation is systematically different from the control observations. Critical ratios of 2 or more are considered *statistically reliable*, which is just another way of saying that the null hypothesis can be rejected. Critical ratios of less than 2 are considered too small to allow the rejection of the null hypothesis.*

This general procedure is not foolproof. It is certainly possible for an individual to have a reading score of 30 (a critical ratio of 2) or even lower without being dyslexic. According to Figure A.3, this will happen about 2 percent of the time. Raising the cutoff value to a critical ratio of 3 or 4 would make such errors less common but would not eliminate them entirely; furthermore, raising the critical value might mean failure to detect some individuals with dyslexia. One of the important consequences of the variability in psychological data can be seen here: the investigator who has to decide between two interpretations of the data (the null hypothesis and the alternative hypothesis) cannot be correct all the time. Using statistics, in other words, is a matter of playing the odds.

TESTING HYPOTHESES ABOUT MEANS

In the preceding discussion, our concern was with hypotheses about single scores. We now turn to the more commonly encountered problems in which the hypotheses involve means.

*Many authors use the term *statistically significant* instead of *statistically reliable*, and the decision process we are describing is sometimes referred to as *significance testing*. However, the term we are using, *reliability*, seems preferable for two reasons. First, what the statistics are measuring really is a matter of reliability—that is, whether the observation before us is likely to be an accident (and so probably would not reappear if we ran the test again), or whether it is reliable (and so would reappear if we retested). Second, the term *significance* implies that a result is important, consequential, worth publicizing. The statistical tests tell us none of those things, and so a "statistically significant" result might, in truth, be entirely insignificant in the eyes of the world! Hence the label of *statistical significance* seems a misnomer.

In many experiments, the investigator compares two or more groups—participants tested with or without a drug, with or without imagery instructions, and so on. Suppose we get a difference between the groups. How do we decide whether the difference is genuine rather than due merely to chance?

Let us return to the experiment in which memory for words was tested with and without instructions to imagine the items. To simplify, we will here consider a modified version of the experiment in which the same participants serve in both the imagery and the nonimagery conditions. Each participant memorizes a list of 20 words without instructions, then memorizes a second list of 20 words under instructions to visualize. What we want to know is whether the participants show any improvement with the imagery instructions. There is no separate control group in this experiment. Because each personís score while using imagery can be compared with his score without using imagery, each provides his own control.*

Table A.6 gives data for the ten participants in the experiment. For each one, the table lists the number of words recalled without imagery instructions, the number recalled with such instructions, and the improvement (the difference between the two scores). The mean improvement overall is 3 words, from a mean of 8 words recalled

TABLE A.6 Number of Items Recalled With and Without Imagery Instruction, for Ten Participants

Subject	Score with Imagery	Score without Imagery	Improvement
Alphonse	11	5	6
Betsy	15	9	6
Cheryl	11	5	6
Davis	9	9	0
Earl	13	6	7
Fred	10	11	−1
Germaine	11	8	3
Hortense	10	11	−1
Imogene	8	7	1
Jerry	12	9	3
Mean	11	8	3

$$\text{Variance of improvement scores} = \frac{\text{sum of (score} - 3)^2}{10} = 8.8$$

$$\text{Standard deviation of improvement scores} = \sqrt{8.8} = 2.97$$

* This sort of design, in which participants serve in more than one condition, is called a *within-subjects design*, in contrast to a *between-subjects design* in which different people serve in the different conditions. Within-subjects designs have certain advantages; among them, we can obviously be certain that the participants in one group are identical to the participants in the other group. But within-subjects designs also introduce their own complications. For example, if the participants serve in one condition first, then in the other condition, then this creates a confound: any differences observed might be due to the effects of practice, which obviously benefits the second condition. For present purposes, we ignore these complications (and also the steps needed to control for this confound). For more on this issue, though, see Chapter 1. In any case, the logic of the statistics here is very similar to the logic relevant to between-subjects designs, and so we will use this (simpler) case as a way to convey this logic.

without imagery to a mean of 11 words with imagery. But note that this does not hold for all participants. For example, for Fred and Hortense, the "improvement" is negative: they both do better without imagery instructions. But is there an imagery facilitation effect overall? Put in other words, is the difference between the two conditions statistically reliable?

As one way to approach this question, note that, ultimately, we are not trying to draw conclusions about the specific ten people we ran in the experiment. Instead, we want to draw broader conclusions, about the population at large. One way to make sure our data justify such broad conclusions would be to test the entire population in our study—every adult in North America, justifying claims about North America, or every adult in Europe, justifying claims about Europe, and so on.

Of course, we could not run these huge studies—they would require absurd amounts of time and effort. What we do instead is test a sample of individuals, and so we observe a mean *for this sample*. But can we extrapolate from this sample? It is useful to keep in mind here that we might easily have run a different sample, and it would have produced its own mean, or some other sample, with its own mean, and on and on and on for the vast number of samples we could have tested. Each sample would produce a mean (called, for obvious reasons, a *sample mean*), and, if we did in fact run sample after sample, we would end up with a set of sample means. From that set—from the distribution of sample means—we could compute a mean of all the means, and this would tell us something about the broader population (and so this mean of means, averaging together all the samples we might gather, is called the *population mean*). We could also ask how variable this set is—by computing a standard deviation for the distribution of sample means.

What we really want to ask, therefore, is whether the sample mean we actually obtained (based on the data in Table A.6) is representative of the population mean—the average that we would observe if we ran sample after sample after sample and averaged them all together. The possibility that we hope for is that our sample mean *is* representative of this larger group, which is equivalent to claiming that we would get roughly the same result if we were to do the experiment a second, third, or fourth time. A different possibility, though, is that our sample mean is just a lucky accident—showing an apparent difference between the conditions that would not show up reliably if we performed the experiment again and again. This latter possibility is, in this context, the null hypothesis. As we mentioned, the null hypothesis is, in general, a claim that there is no systematic difference between the observations we are comparing. In the dyslexia case, this was a claim that the person we had tested was not systematically different from the broader population. In the present case, it is a claim that there is no systematic difference between memory with imagery and memory without. It is, therefore, the claim that, if we conducted the memory experiment again and again, we would not observe a difference, and therefore the difference that has emerged in our data is just a fluke.

We test the null hypothesis in the memory experiment in the same way that we did in our dyslexia example. In that example, we computed a critical ratio (that is, a z-score) based on the difference between the score we had actually observed and the mean that was predicted by the null hypothesis. The null hypothesis claimed that the individual we tested was not dyslexic, and so the relevant mean was the mean for the broad population of nondyslexics. In the present example, we follow the same logic. The null hypothesis claims that, if we run the test over and over, we will not observe a difference, and so the mean we should expect, on this hypothesis, is zero. (In other words, the null hypothesis claims that the population mean, in this case the mean difference between the imagery and control conditions, is zero.)

The formula we will use for the present case is the standard one:

$$Z = \frac{(\text{score} - \text{M})}{\text{SD}}$$

The score we will use in this calculation will be the sample mean we actually obtained—a value of 3 (see Table A.6). The mean (M) will be the mean assumed by the null hypothesis—in this case, zero. But what is the denominator? It is the standard deviation from the set of all the sample means, a measure of how variable the data would be as we move from one sample to another. This value—the standard deviation of a distribution of sample means—is called the *standard error* (*SE*) of the mean. Its value is determined by two factors: the standard deviation of the sample and the size of that sample. Specifically,

$$\text{SE} = \frac{\text{SD}}{\sqrt{N - 1}} \tag{4}$$

Why this formula takes this particular form is beyond the scope of our discussion. Two elements of the formula should, however, seem sensible. First, in calculating the standard error, all we have to go on in most cases is information about the actual sample we have observed. We have no information about all those other samples that we might have tested (but actually did not!). Therefore, it is plausible that our estimate of how variable the data would be in general (and this is, of course, exactly what the standard error measures) depends heavily on how variable our original sample is (that is, what the standard deviation of the sample measures). It should, in other words, be no surprise that the standard error is proportional to the standard deviation.

Second, it should also seem right that the standard error goes down as the size of our particular sample goes up. If our sample included only two or three observations, then it is entirely likely that our sample mean has been drawn off by one or more atypical scores. If our sample was larger, then the impact of these atypical scores would be diluted within the larger data set. In that case, our sample would more likely be reflective of the population at large, and our estimate of the standard error is correspondingly lowered.

In any case, with the standard error now defined, we can conclude our analysis of the results of our memorization experiment. The critical ratio to be evaluated is

$$\text{Critical ratio} = \frac{\text{obtained sample mean} - \text{population mean}}{\text{SE}}$$

Since the population mean is assumed to be zero (by the null hypothesis), this expression becomes

$$\text{Critical ratio} = \frac{\text{obtained sample mean}}{\text{SE}} \tag{5}$$

To compute the standard error, we first find the standard deviation of the imagery scores; this turns out to be 2.97, as shown in Table A.6. Then equation (4) tells us

$$\text{SE} = \frac{\text{SD}}{\sqrt{N - 1}} = \frac{2.97}{\sqrt{10 - 1}} = .99$$

The critical ratio is now the obtained mean difference divided by the standard error, or 3/.99 = 3.03. This is clearly larger than 2.0, so we conclude that the observed difference in memory between the imagery and control conditions probably should not be attributed to chance. Said differently, the sample we have run (which does show a difference between conditions) is probably representative of the data we would get if we ran the experiment again, with a new group of participants and a new set of stimuli. Thus the pattern is deemed reliable, and so we can conclude that giving visual imagery instructions does improve recall.*

CONFIDENCE INTERVALS

In using statistics to test hypotheses, we ask whether a certain sample mean could have been drawn by chance from a set of sample means distributed around some assumed population mean. (When testing the null hypothesis, this assumed population mean is zero.) But there is another way of phrasing the entire issue: given a sample of data with its own sample mean, can we extrapolate from this in a fashion that allows us to specify, with reasonable confidence, what the possible range of values might be for the population mean? If we know the standard error of the mean, the answer is yes. We have already seen that about 2 percent of the scores in a normal distribution are more than 2 SDs above the distribution's mean (see Figure A.3). Similarly, about 2 percent of the scores have values lower than 2 SDs below the mean. Since this is so, we can conclude that the chances are roughly 96 in 100 that the population mean is within an interval whose largest value is 2 SEs above the sample mean and whose lowest value is 2 SEs below. Because we can be fairly (96 percent) confident that the actual population mean will fall within this specified range, it is often called the *confidence interval*.

As an example, consider the prediction of elections. During election campaigns, polling organizations report the current standing of various candidates by statements such as the following: "In a poll of 1,000 registered voters, 57 percent favored candidate Smith; the margin of error was 3 percent." This margin of error is the confidence interval around the proportion (that is, ± 3 percent).

To determine this confidence interval, the pollsters compute the standard error of the proportion they found (in this case, .57). This standard error is analogous to the standard error of a mean we discussed in the previous section. Given an N of 1,000, this standard error happens to be .015.** Since $2 \times .015$ is .03 or 3 percent, the appropriate confidence interval for our example is the interval from 54 to 60 percent. Under the circumstances, candidate Smith can be fairly confident that she has the support of at least 50 percent of the electorate, since 50 percent is well below the poll's confidence interval (see Figure A.9).

Some Implications of Statistical Inference

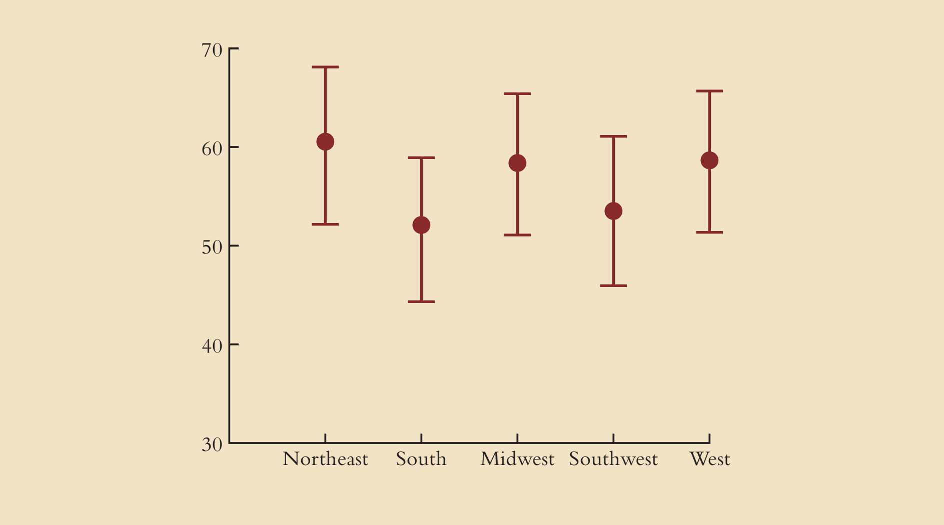
The methods of testing hypotheses and estimating confidence intervals that we just described are routinely employed in evaluating the results of psychological research. But they have several characteristics that necessarily affect the interpretation of all such results.

* There are several simplifications in this account. One is that the critical ratio described here does not have an exactly normal distribution. When the sample size is large, this effect is unimportant, but for small samples (like the one in the example) they can be material. To deal with these and related problems, statisticians often use measures that refer to distributions other than the normal one. An example is the *t*-test, a kind of critical ratio based on what is called the *t*-distribution.

** The standard error of a proportion (e.g., the proportion of polled voters who express pro-*X* sentiments) is analogous to the standard error of the mean and measures the precision with which our sample proportion estimates the population proportion. The formula for the standard error of a proportion p is:

$$\mathrm{SE}_p = \sqrt{\frac{p \times (1 - p)}{N}}$$

In our example, $p = .57$ and $N = 1{,}000$, so $\mathrm{SE}_p = .015$.



A.9 A candidate's poll results and her confidence intervals The results of a mythical poll conducted for a no-less-mythical candidate Smith by randomly sampling 200 people in each of five regions of the United States. The figure shows the pro-Smith proportions in each region, together with the confidence intervals around them, and indicates that she is ahead in all five samples. But there are two regions where she cannot be confident that she is ahead in the population—the South and the Southwest, where the confidence intervals of the pro-Smith proportion dip below 50 percent.

THE PROBABILISTIC NATURE OF HYPOTHESIS TESTING AND CONFIDENCE INTERVALS

As we have already noted, the nature of statistical testing always leaves the possibility of error. In our dyslexia case, we discussed the fact that it is unlikely that someone with a score 2 SDs below the population mean is, in truth, drawn from that population, but it is still possible. (In fact we know exactly how often this sort of unusual occurrence would occur: 2 percent of the time.) Likewise, if we use a confidence interval of ± 2 SEs, the chance that the population mean (or proportion, or whatever) falls outside of that interval is less than 4 or 5 in 100. This is a small chance for error, but it is still a chance.

Do we want to be more confident than this? If so, we might use a confidence interval of ± 3 SEs, where the equivalent chance is only 1 in 1,000. The same holds for critical ratios. If we want to be cautious, we might insist that the critical ratio be larger than the usually assumed value of 2—perhaps 3 (a chance factor of 1 in 2,000) or 4 (1 in 20,000), and so on. But the likelihood of these chance occurrences is never zero, and so, as long as there is some unexplained variance, there is some possibility of error.

The probabilistic nature of statistical reasoning has another consequence. Even if we can come to a correct conclusion about the mean of a population (or a proportion, as in polls), we cannot generalize to individuals. The variability within the population (or within our sample) simply prohibits us from applying claims, true for the population, to each individual within the population. Thus, a study which shows that men have higher scores than women on spatial relations tests does not preclude the existence of brilliant female artists or architects.

THE ROLE OF SAMPLE SIZE

A last point concerns the role of sample size in affecting how the results are interpreted. The larger the sample, the smaller the standard error and the smaller the confidence interval around the mean or the proportion. This can have major effects on hypothesis testing.

Suppose that, in the population, a certain independent variable produces a very small difference. As an example, suppose that the population difference between men and women on a certain test of spatial relations is 1 percent. We would probably be unable to reject the null hypothesis (that is, the hypothesis that there is no sex difference on the test) with samples of moderate size. But if the sample size were sufficiently large, we could reject the null hypothesis. For an *N* of such magnitude would lead to a decrease in the standard errors of the sample means, which in turn would lead to an increase in the critical ratio. Someone who read a report of this experiment would now learn that, by using thousands of participants, we discovered a reliable difference of 1 percent. A fair reaction to this bit of intelligence would be that the null hypothesis can indeed be rejected, but that the psychological significance of this finding is rather slight. The moral is simple: Statistical reliability does indicate a difference and, moreover, indicates that the difference is unlikely to be a fluke or chance occurrence. But statistical reliability, by itself, does not indicate whether the effect discovered is of psychological significance or of any practical importance.

SUMMARY APPENDIX

- Statistical methods concern the ways in which investigators describe, organize, and interpret collections of numerical data. A crucial concern of statistical endeavors is to interpret the *variability* that is encountered in all research.
- An early step in processing numerical data is *scaling*, a procedure for assigning numbers to psychological responses. Scales can be *categorical*, *ordinal*, *interval*, or *ratio scales*. These differ in the degree to which they can be subjected to arithmetical operations.
- An important step in organizing the data is to arrange them in a *frequency distribution*, often displayed in graphic form, as in a *histogram*. Frequency distributions are summarized by a *measure of central tendency*. The common measure of central tendency is the *mean* (*M*), though sometimes another measure, the *median*, may be preferable, as in cases when the distribution is *skewed*. Important measures of variability are the *variance* (*V*) and the *standard deviation* (*SD*).
- One way of comparing two scores drawn from different distributions is to convert both into *percentile ranks*. Another is to transform them into *z*-scores, which express the distance of a score from its mean in standard deviations. The percentile rank of a *z*-score can be computed if the shape of that score's distribution is known. An important example is the *normal distribution*, graphically displayed by the *normal curve*, which describes the distribution of many psychological variables and is basic to much of statistical reasoning.
- In some studies, the relation between variables is expressed in the form of a *correlation*, which may be positive or negative. It is measured by *r*, the correlation coefficient, a number that can vary from +1.00 to −1.00. Correlations reflect the extent to which two variables vary together, but they do not necessarily indicate that one of them causes the other.
- One of the main functions of statistical methods is to help test hypotheses about a population given information about the sample. An important example is the difference between mean scores obtained under two different conditions. Here, the investigator has to decide between the *null hypothesis*, which asserts that the difference was obtained by chance, and the *alternative hypothesis*, which asserts that the difference is genuine and exists in the population. The decision is made by dividing the obtained mean difference by the *standard error* (*SE*), a measure of the variability of that mean difference. If the resulting ratio, called the *critical ratio*, is large enough, the null hypothesis is rejected, the alternative hypothesis is accepted, and the difference is said to be *statistically reliable*. A related way of making statistical decisions is by using a *confidence interval*, or margin of error. This is based on the variability of the scores from a sample and determines the interval within which the population mean or proportion probably falls.