

Using Statistics to Answer Questions

CHAPTER

9

Descriptive Statistics

• Scales of Measurement • Measures of Central Tendency • Graphing Your Results • Calculating and Computing Statistics • Measures of Variability

Correlation

• The Pearson Product–Moment Correlation Coefficient

Inferential Statistics

• What Is Significant? • The t Test
• One-Tail Versus Two-Tail Tests of Significance
• The Logic of Significance Testing
• When Statistics Go Astray: Type I and Type II Errors

Effect Size

Just as detectives seek out clues and leads and gather data to help solve a case, psychologists, too, gather data to help answer research questions. After they have gathered the evidence, detectives must determine whether it is real (meaningful). Likewise, in later chapters, we will examine several statistical methods used to determine whether the results of an experiment are meaningful (significant). As we have seen, the term *significant* is used to describe those instances in which the statistical results are likely to have been caused by our manipulation of the independent variable (IV).

To understand better the nature of statistical significance, a closer look at statistics is in order. **Statistics** is a branch of mathematics that involves the collection, analysis, and interpretation of data. Researchers use various statistical techniques to aid them in several ways during the decision-making processes that arise when conducting research.

The two main branches of statistics assist your decisions in different ways. **Descriptive statistics** summarize any set of numbers so you can understand and talk about them more intelligibly. Researchers use **inferential statistics** to analyze data after they have conducted an experiment to determine whether the IV had a significant effect. Although we assume that you already have some familiarity with statistics, we have included several relevant formulas in Appendix B. We encourage you to review them at this time and as needed.

Statistics The branch of mathematics that involves the collection, analysis, and interpretation of data.

Descriptive statistics Procedures used to summarize a set of data.

Inferential statistics Procedures used to analyze data after an experiment is completed in order to determine whether the independent variable has a significant effect.

Descriptive Statistics

We use descriptive statistics when we want to summarize a set or distribution of numbers in order to communicate their essential characteristics. One of these essential characteristics is a measure of the typical or representative score, called a *measure of central tendency*. A

second essential characteristic that we need to know about a distribution is how much *variability* or spread exists in the scores. Before we discuss these measures of central tendency and variability, however, we must examine the measurements on which they are based.

Scales of Measurement

Measurement The assignment of symbols to events according to a set of rules.

Scale of measurement A set of measurement rules.

Nominal scale A scale of measurement in which events are assigned to categories.

Ordinal scale A scale of measurement that permits events to be rank ordered.

Interval scale A scale of measurement that permits rank ordering of events with the assumption of equal intervals between adjacent events.

We can define **measurement** as the assignment of symbols to events according to a set of rules. Your grade on a test is a symbol that stands for your performance; it was assigned according to a particular set of rules (the instructor's grading standards). The *particular* set of rules used in assigning a symbol to the event in question is known as a **scale of measurement**. The four scales of measurement that are of interest to psychologists are nominal, ordinal, interval, and ratio scales. How you choose to measure (i.e., which scale of measurement you use) the dependent variable (DV) directly determines the type of statistical test you can use to evaluate your data after you have completed your research project.

Nominal Scale The **nominal scale** is a simple classification system. For example, if you are categorizing the furniture in a classroom as tables or chairs, you are using a nominal scale of measurement. Likewise, recording responses to an item on a questionnaire as "agree," "undecided," or "disagree" reflects the use of a nominal scale of measurement. You assign the items being evaluated to mutually exclusive categories.

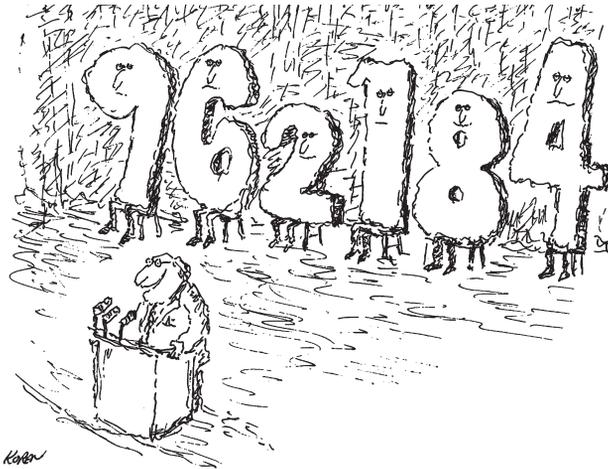
Ordinal Scale When you can rank order the events in question, you are using an **ordinal scale** of measurement. Notice that we indicated *only* that the events under consideration could be rank ordered; we did not indicate that the intervals separating the units were comparable. Although we can rank the winners in a track meet (i.e., first, second, third, fourth), this rank ordering does not tell us anything about how far apart the winners were. Perhaps it was almost a dead heat for first and second; maybe the winner was far ahead of the second-place finisher.

Interval Scale When you can rank order the events in question and equal intervals separate adjacent events, you are using an **interval scale**. For example, the temperatures on a Fahrenheit thermometer form an interval scale; rank ordering has been achieved *and* the difference between any two adjacent temperatures is the same, 1 degree. Notice that the interval scale does not have a true zero point, however. When you reach the "zero" point on a Fahrenheit thermometer, does temperature cease to exist? No, it's just very cold. Likewise, scores on tests such as the SAT and ACT are interval-scale measures.



Assume you are on a college admissions committee and you are reviewing applications. Each applicant's ACT score forms an integral part of your review. You have just come across an applicant who scored 0 on the verbal subtest. What does this score tell you?

The score of 0 should not be interpreted as meaning that this individual has absolutely no verbal ability. Because ACT scores are interval-scale measurements, there is no true zero. A score of 0 should be interpreted as meaning that the individual is very low in that ability. The same could be said for 0 scores on the wide variety of tests, questionnaires, and personality inventories routinely used by psychologists in personality research. The presence of a true zero is characteristic only of the ratio scale of measurement.



"Tonight, we're going to let the statistics speak for themselves."

© The New Yorker Collection 1974 Edward Koren from cartoonbank.com. All Rights Reserved.

Ratio Scale The **ratio scale** of measurement takes the interval scale one step farther. Like the interval scale, the ratio scale permits the rank ordering of scores with the assumption of equal intervals between them, *but* it also assumes the presence of a true zero point. Physical measurements, such as the amplitude or intensity of sound or light, are ratio measurements. These measurements can be rank ordered, and there are equal intervals between adjacent scores. However, when a sensitive measuring device reads 0, there is nothing there. Because of the true zero point, the ratio scale allows you to make ratio comparisons, such as "twice as much" or "half as much."

Ratio scale A scale of measurement that permits rank ordering of events with the assumptions of equal intervals between adjacent events and a true zero point.

Our discussion of scales of measurement has progressed from the nominal scale, which provides the least amount of information, to the ratio scale, which provides the greatest amount of information. When psychologists evaluate changes in the DV, they try to use a scale of measurement that will provide the most information; frequently, they select interval scales because they do not use measurements that have a true zero.

We now turn to the topic of central tendency. Keep in mind that the scales of measurement directly determine which measure of central tendency you will use.

Measures of Central Tendency

Measures of central tendency, such as the mode, median, and mean, tell us about the typical score in a distribution.

Mode The score in a distribution that occurs most often.

Mode The **mode** is the number or event that occurs most frequently in a distribution. If students reported the following work hours per week

12, 15, 20, 20, 20

the mode would be 20.

$$\text{Mode} = 20$$

Although the mode can be calculated for any scale of measurement, it is the only measure of central tendency that can be used for nominal data.

Median The number that divides a distribution in half.

Median The **median** is the number or score that divides the distribution into equal halves. To be able to calculate the median, you must first rank order the scores. Thus, if you started with the following scores

56, 15, 12, 20, 17

you would have to rank order them as follows:

12, 15, 17, 20, 56

Now it's an easy task to determine that 17 is the median (Mdn):

$$\text{Mdn} = 17$$

What if you have an even number of scores, as in the following distribution?

1, 2, 3, 4, 5, 6

In this case the median lies halfway between the two middle scores (3 and 4). Thus, the median would be 3.5, halfway between 3 and 4. The median can be calculated for ordinal, interval, and ratio data.

Mean The arithmetic average of a set of numbers; found by adding all the scores in a set and then dividing by the number of scores.

Mean The **mean** is defined as the arithmetic average. To find the mean we add all the scores in the distribution and then divide by the number of scores we added. For example, assume we start with

12, 15, 18, 19, 16

We use the Greek letter sigma, Σ , to indicate the sum. If X stands for the numbers in our distribution, then ΣX means to add the numbers in our distribution. Thus, $\Sigma X = 80$. If N stands for the number of scores in the distribution, then the mean would equal $\Sigma X/N$. For the previous example, $80/5 = 16$. The sum of these numbers is 80, and the mean is 16 ($80/5$). The mean is symbolized by M .

You may recall from your statistics class that \bar{X} stood for the mean. We haven't arbitrarily changed symbols on you. Because M stands for the mean in APA-format papers (see Chapter 14), we chose to use it instead of \bar{X} . Thus, $M = 16$. You can calculate the mean for interval and ratio data, but not for nominal and ordinal data.

Choosing a Measure of Central Tendency Which measure of central tendency should you choose? The answer to that question depends on the type of information you are seeking and the scale of measurement you are using. If you want to know which score occurred most often, then the mode is the choice. However, the mode may not be very representative of the other scores in your distribution. Consider the following distribution:

1, 2, 3, 4, 5, 11, 11

In this case the mode is 11. Because all the other scores are considerably smaller, the mode does not accurately describe the typical score.

The median may be a better choice to serve as the representative score because it takes into account all the data in the distribution; however, there are drawbacks with this choice. The median treats all scores alike; differences in magnitude are not taken into account. Thus, the median for *both* of the following distributions is 14:

Distribution 1: 11, 12, 13, **14**, 15, 16, 17 Mdn = 14

Distribution 2: 7, 8, 9, **14**, 23, 24, 25 Mdn = 14

When we calculate the mean, however, the *value* of each number is taken into account. Although the medians for the two distributions above are the same, the means are not:

Distribution 1: 11, 12, 13, 14, 15, 16, 17

$\Sigma X = 98$ $M = 98/7$ $M = 14$

Distribution 2: 7, 8, 9, 14, 23, 24, 25

$\Sigma X = 110$ $M = 110/7$ $M = 15.71$

The fact that the mean of Distribution 2 is larger than that of Distribution 1 indicates that the value of each individual score has been taken into account.

Because the mean takes the value of each score into account, it usually provides a more accurate picture of the typical score, and it is the measure of central tendency favored by psychologists. On the other hand, there are instances in which the mean may be misleading. Consider the following distribution of charitable donations:

Charitable donations: \$1, \$1, \$1, \$5, \$10, \$10, \$100

Mode = \$1

Mdn = \$5

Mean = $\$128/7$ $M = \$18.29$

If you wanted to report the “typical” gift, would it be the mode? Probably not. Even though \$1 is the most frequent donation, this amount is *substantially* smaller than the other donations, and more people made contributions over \$1 than made the \$1 contribution. What about the median? You see that \$5 appears to be more representative of the typical donation; there are an equal number of higher and lower donations. Would the mean be better? In this example the mean is substantially inflated by one large donation (\$100); the mean is \$18.29 even though six of the seven donations are \$10 or under. Although reporting the mean in this case may look good on a report of giving, it does not reflect the typical donation.

The lesson to be learned from this example is that when you have only a limited number of scores in your distribution, the mean may be inflated (or deflated) by extremely large (or extremely small) scores. The median may be a better choice as your measure of central tendency in such instances. As the number of scores in your distribution increases, the influence of extremely large (or extremely small) scores on the mean decreases. Look what happens if we collect two additional \$5 donations:

Charitable donations: \$1, \$1, \$1, \$5, \$5, \$5, \$10, \$10, \$100

Mode = \$1 and \$5

Mdn = \$5

Mean = $\$138/9$ $M = \$15.33$

Note we now have two values (\$1 and \$5) for the mode (i.e., a *bimodal distribution*). The median stays the same (\$5); however, the mean has decreased from \$18.29 to \$15.33; the addition of only two more low values moved the mean closer to the median.

Graphing Your Results

After you have calculated a measure of central tendency, you can convey this information to others. If you have only one set of scores, the task is simple: You write down the value as part of your paper or report.

What if you are dealing with several groups or sets of numbers? Now the task is complicated, and the inclusion of several numbers in a paragraph of text might be confusing. In such cases a graph or figure can be used to your advantage; a picture may well be worth a thousand words. It is not uncommon to see a detective use a chart or graph to help make a point in the solution of a case. In preparing a research report, psychologists also use graphs effectively. There are several types of graphs for the researcher to choose from. Your choice of graphs will be determined by which one depicts your results most effectively *and* by the scale of measurement you used. For example, if you used a nominal scale of measurement, then you would probably use a pie chart, a histogram, a bar graph, or a frequency polygon.

Pie chart Graphical representation of the percentage allocated to each alternative as a slice of a circular pie.

Pie Chart If you are dealing with percentages that total 100%, the familiar pie chart may be a good choice. The **pie chart** depicts the percentage represented by each alternative as a slice of a circular pie. The larger the slice, the greater the percentage. For example, if you surveyed college men to determine their TV viewing preferences, you might display your results as a pie chart (see Figure 9-1). From the hypothetical data presented in Figure 9-1 we can see that the mode is sports programs.



Take another look at Figure 9-1. Why would it not be appropriate to describe a mean preference in this instance?

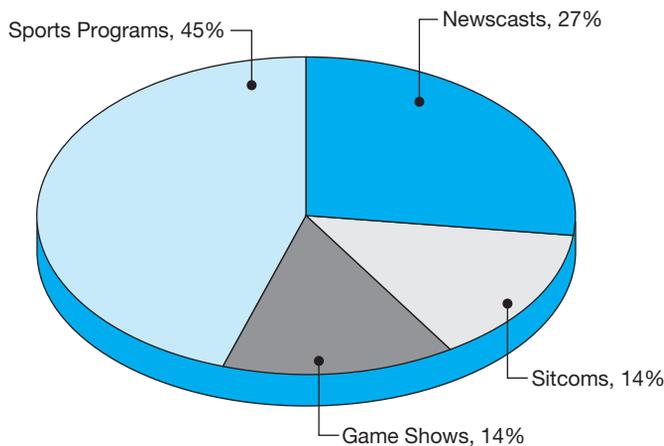


FIGURE 9-1 TV Viewing Preferences of College Men.

To answer this question, ask yourself what scores we would use to obtain the mean. We know there are four categories of TV preference, and we know the percentage preferring each category. We could add the percentages for each category and then divide by the number of categories. The resulting number would tell us that there is an average of 25% per category; unfortunately, that number does not tell us much about the “mean preference.” We would have to have individual scores, in the form of interval or ratio data, before we could calculate a mean preference. These data are simply not available.

Histogram We can use a **histogram** to present our data in terms of frequencies per category. When we study a *quantitative variable*, we construct a histogram. Quantitative categories are ones that can be numerically ordered. The levels or categories of a quantitative variable must be arranged in a numerical order. For example, we may choose to arrange our categories from smallest to largest, or vice versa. Figure 9-2 shows a histogram for the age categories of participants in a developmental psychology research project.

Histogram A graph in which the frequency for each category of a quantitative variable is represented as a vertical column that touches the adjacent column.

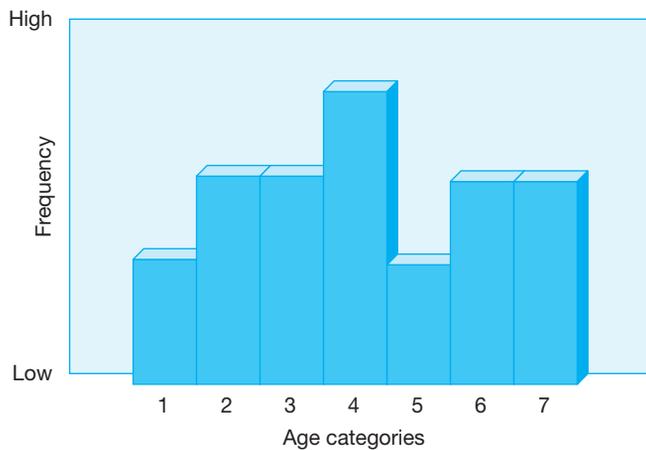


FIGURE 9-2 Histogram Depicting the Frequency of Participants in Various Age Categories in a Developmental Psychology Research Project. Note that the sides of adjacent columns touch.

Bar Graph The **bar graph** also presents data in terms of frequencies per category; however, we are using *qualitative categories* when we construct a bar graph. Qualitative categories are ones that cannot be numerically ordered. For example, single, married, divorced, and remarried are qualitative categories; there is no way to order them numerically.

Figure 9-3 shows a bar graph for the sports and fitness preferences of women and girls who are frequent participants in such activities. Placing a space between the bars lets the reader know that qualitative categories are being reported. You can see at a glance that the number per category and the type of activities differ dramatically between the two groups. Think of how many words it would take to write about these results rather than present them in graph form.

Bar graph A graph in which the frequency for each category of a qualitative variable is represented as a vertical column. The columns of a bar graph do not touch.

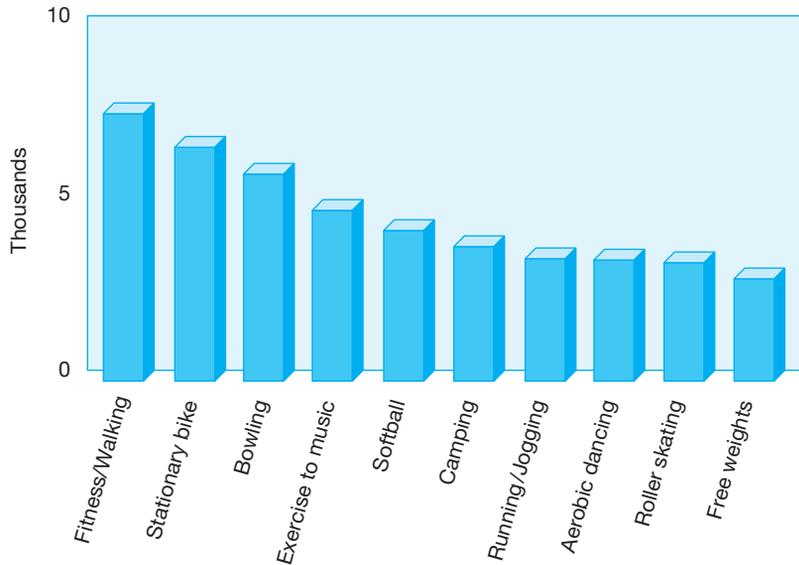


FIGURE 9-3 Bar Graphs Depicting the Sports and Fitness Preferences of Women and Girls Who Are Frequent Participants in Such Activities. Because a bar graph depicts a qualitative variable, the bars do not touch.

Source: *The American Enterprise*, a national magazine of politics, business, and culture, September/October 1993, p. 101. TAEmag.com.

Frequency polygon A graph that is constructed by placing a dot in the center of each bar of a histogram and then connecting the dots.

Line graph A graph that is frequently used to depict the results of an experiment.

Ordinate The vertical or y axis of a graph.

Abscissa The horizontal or x axis of a graph.

Frequency Polygon If we mark the middle of the crosspiece of each bar in a histogram (see Figure 9-4A) with a dot, connect the dots, and remove the bars, we have constructed a **frequency polygon** (see Figure 9-4B).

The frequency polygon, like the histogram, displays the frequency of each number or score. The only difference between the two is that we used bars in the histogram and connected dots in the frequency polygon.

Line Graph Researchers frequently present the results of psychological experiments as a **line graph**. In constructing a line graph, we start with two axes or dimensions. The vertical or y axis is known as the **ordinate**; the horizontal or x axis is known as the **abscissa** (see Figure 9-5). We plot our scores or data (the DV) on the ordinate. The values of the variable we manipulated (the IV) are plotted on the abscissa.

How tall should the y axis be? How long should the x axis be? A good rule of thumb is for the y axis to be approximately two-thirds as tall as the x axis is long (see Figures 9-5 and 9-6). Other configurations will give a distorted picture of the data. For example, if the ordinate is considerably shorter, differences between groups or treatments will be obscured (see Figure 9-7A, page 180), whereas lengthening the ordinate tends to exaggerate differences (see Figure 9-7B).

In Figure 9-6 we have plotted the results of a hypothetical experiment that evaluated the effects of different levels of stress on making correct landing decisions by air traffic controllers. As you can see, as stress increased, the number of correct responses increased. What if we had tested two different groups of participants,

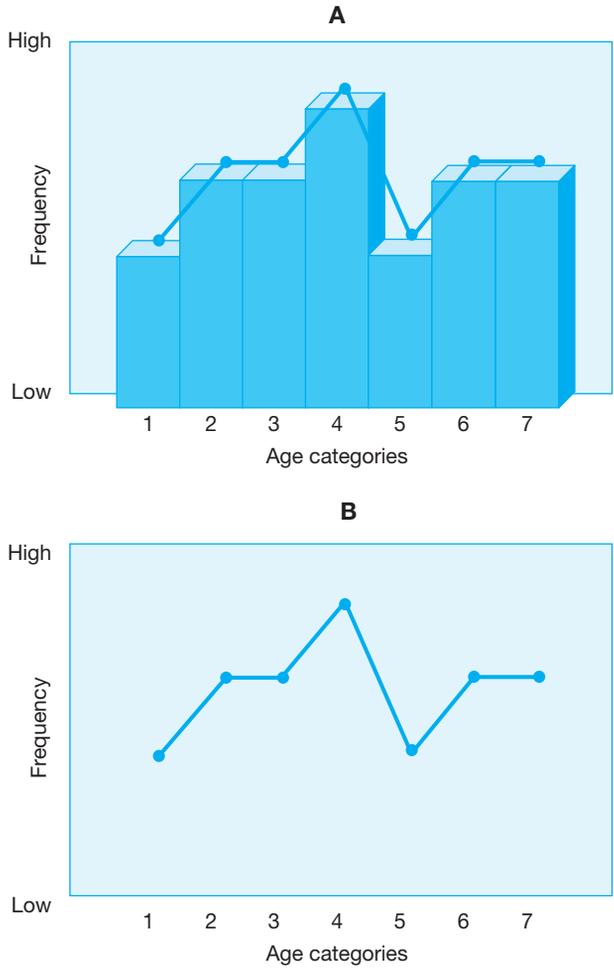


FIGURE 9-4 The Frequency Polygon Is Constructed by Placing a Dot in the Center of Each Bar of a Histogram and Connecting the Dots (A) and Removing the Bars (B). The frequency polygon, like the histogram, displays the frequency of each score or number.

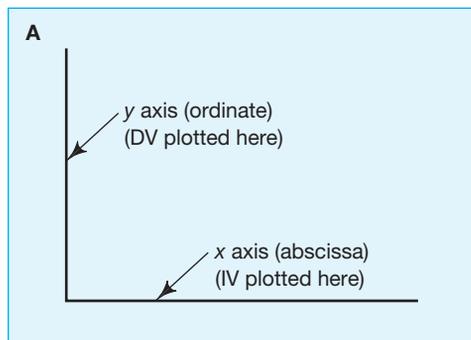


FIGURE 9-5 The Ordinate or y Axis and Abscissa or x Axis. The ordinate should be about $\frac{2}{3}$ the size of the abscissa to portray the data as clearly as possible.

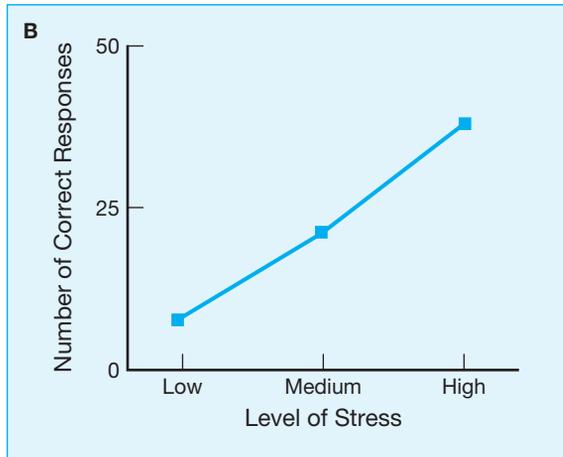


FIGURE 9-6 Results of a Hypothetical Experiment Investigating the Effects of Stress on Correct Response by Air Traffic Controllers.

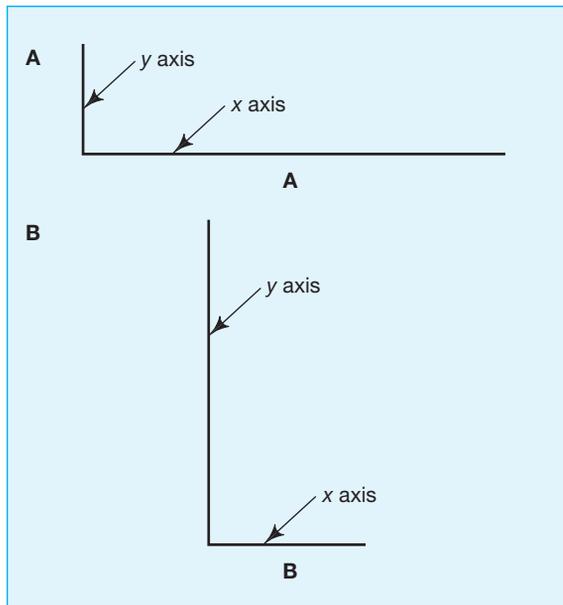


FIGURE 9-7 Altering the x (Abscissa) or y (Ordinate) Axis Can Distort the Results of an Experiment. **A.** If the ordinate is considerably shorter than the abscissa, significant effects can be obscured. **B.** If the ordinate is considerably longer than the abscissa, very small effects can be exaggerated.

college students and air traffic controllers? How would we display the results of both groups on the same graph? No problem. All we must do is add the data points for the second group and a legend or box that identifies the groups (see Figure 9-8). Now, we can see at a glance that the air traffic controllers, whose occupation is very stressful, made more correct responses as stress levels increased; the converse was true for the college students.

When you graph the results of an experiment in which more than one variable is used, how do you know which IV to plot on the abscissa? Although there is no fixed rule, a good guideline is to plot the variable having the greater number of levels on the abscissa. Thus, in Figure 9-8 the three levels of stress were plotted on the abscissa, rather than the two levels of participants.

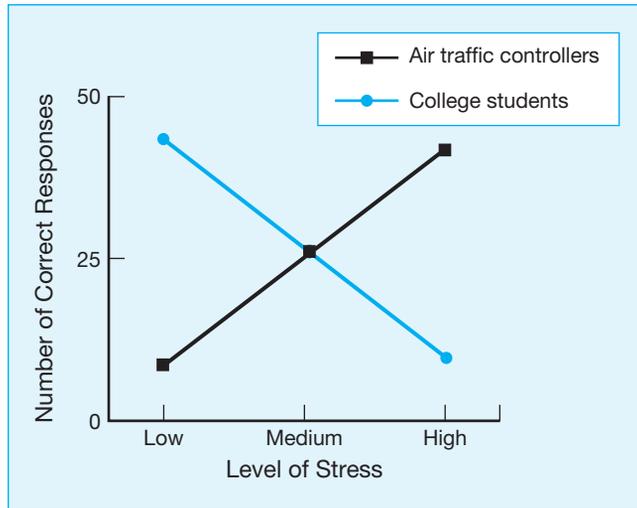


FIGURE 9-8 Results of a Hypothetical Experiment Investigating the Effects of Stress on Correct Response by Air Traffic Controllers and College Students.



Why would you choose to plot the IV with the greatest number of levels on the abscissa?

By plotting the variable with the greatest number of levels on the abscissa, you reduce the number of lines that will appear on your graph. The fewer lines, the less difficulty the reader will have in interpreting your graph. For example, had we plotted the type of participants in our stress experiment on the abscissa, then Figure 9-8 would have had three lines, one for each level of stress. We will discuss the accepted APA format for preparing graphs and tables in Chapter 14.

Calculating and Computing Statistics

Remember, this is *not* a statistics text—we assume you have already taken a statistics course; therefore, we will not review formulas for all the various statistical calculations and tests you will encounter in this book. You can find those formulas in Appendix B. Use them as needed. Your calculating skills may be a little rusty, but all statistical formulas merely require addition, subtraction, multiplication, division, and finding square roots—not all that challenging for a college student . . . especially one with a calculator.

By the same token, most psychologists (and probably most psychology students) rarely use hand computation techniques for statistics after their initial statistics course; the vast majority use a computer package of some sort to analyze data they collect. Of course, these computer packages vary widely. You may have access to a large and powerful statistics package owned by your school or department (some standard packages are SPSS, SAS, and BMD; you probably have Microsoft Excel® on your computer). Alternatively, you may have access to

a smaller statistics program; some schools even require students to buy a statistics software program when they take the statistics course. In any case, you are likely to have access to a computerized statistical analysis program. We cannot begin to give instructions about how to operate the particular program you might have access to—there are simply too many programs. Throughout the chapters that deal with statistics, we will attempt to give you some *general* hints about how to interpret the output you receive from such programs.

Measures of Variability

Although measures of central tendency and graphs convey considerable information, there is still more we can learn about the numbers we have gathered. We also have to know about the variability in our data.

Imagine that our instructor just returned your last psychology test; your score is 64. What does that number tell you? By itself it may not mean very much. You ask your professor for additional information and find that the class mean was 56. You feel better because you were above the mean; however, after a few moments of reflection you realize you need still more information. How were the other scores grouped? Were they all clustered close to the mean or did they spread out considerably? The amount of **variability** or spread in the other scores will have a bearing on the standing of your score. If most of the other scores are very close to the mean, then your score will be among the highest in the class. If the other scores are spread out widely around the mean, then your score will not be one of the strongest. Obviously, you need a measure of variability to provide a complete picture of these data. Range and standard deviation are two measures of variability frequently reported by psychologists.

Variability The extent to which scores spread out around the mean.

Range A measure of variability that is computed by subtracting the smallest score from the largest score.

Range The **range** is the easiest measure of variability to calculate; you rank order the scores in your distribution and then subtract the smallest score from the largest to find the range. Consider the following distribution:

1, 1, 1, 1, 5, 6, 6, 8, 25

When we subtract 1 (the smallest score) from 25 (the largest score), we find that the range is 24:

$$\text{Range: } 25 - 1 = 24$$

However, other than telling us the difference between the largest and smallest scores, the range does not provide much information. Knowing the range is 24 does not tell us about the distribution of the scores we just considered. Consider Figure 9-9.

The range is the same in Parts A and B of Figure 9-9; however, the spread of the scores differs drastically between these two distributions. Most of the scores cluster in the center of the first distribution (Figure 9-9A), whereas the scores are spread out more evenly in the second distribution (Figure 9-9B). We must turn to another measure, the standard deviation, to provide additional information about how the scores are distributed.

Variance A single number that represents the total amount of variation in a distribution; also the square of the standard deviation, σ^2 .

Variance and Standard Deviation To obtain the standard deviation, we must first calculate the variance. You can think of the **variance** as a single number that represents the total amount of variability in the distribution. The larger the number, the greater the total spread of the scores. The variance and standard deviation are based on how much each score in the distribution deviates from the mean.

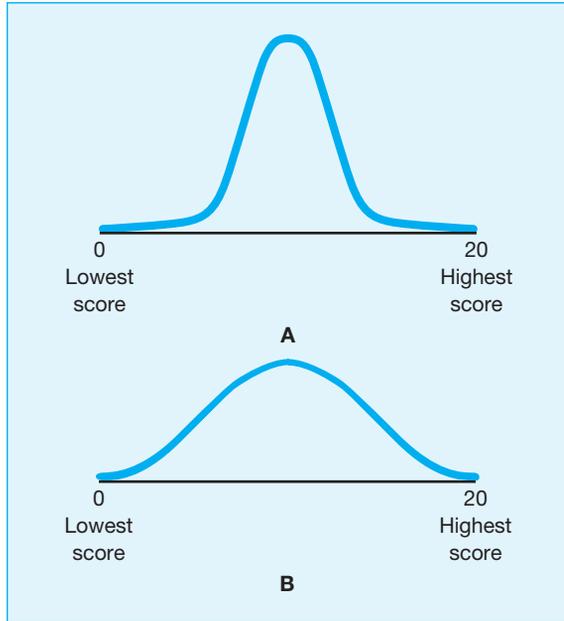


FIGURE 9-9 The Range Does Not Provide Much Information About the Distribution Under Consideration. Even though the range is the same, these two distributions differ drastically.

When researchers conduct experiments, they use a sample of participants to provide information (an estimate) about an entire population. As an example we calculated the variance of the set of nine numbers for which we computed the range and found it to be 58.25. (See Appendix B for the formula to calculate the variance.) Once we have the variance, we can use it to find the standard deviation.

Interpreting the Standard Deviation To find the **standard deviation (SD)**, all we have to do is take the square root of the variance. Using our variance of 58.25,

$$\begin{aligned}
 SD &= \sqrt{\text{variance}} \\
 &= \sqrt{58.25} \\
 &= 7.63
 \end{aligned}$$

Standard deviation
 Square root of the variance; has important relations to the normal curve.

As with the variance, the larger the standard deviation is, the greater the variability or spread of scores will be.

A sample computer printout listing the mean, variance, and standard deviation for the nine scores we used when we calculated the range appears in Table 9-1. Rather than providing printouts from a specific statistics software package, we are using generic printouts in the text. Statistics packages provide slightly different information; we show you the information you might reasonably expect to find on your printout. As you can see, the computer informs us that we entered nine numbers. The mean of these nine numbers is 6.00, the variance is 58.25, and the standard deviation is 7.63.

| TABLE 9-1 | | Computer Printout Showing Mean, Standard Deviation, and Variance | | |
|------------------|------|-------------------------------------------------------------------------|-------|----------|
| Mean | SD | Variance | Range | <i>N</i> |
| 6.00 | 7.63 | 58.25 | 24.00 | 9 |

Normal distribution A symmetrical, bell-shaped distribution having half the scores above the mean and half the scores below the mean.

Now that we have found the standard deviation, what does it tell us? To answer that question we must consider the normal distribution (also called the *normal curve*). The concept of the **normal distribution** is based on the finding that as we increase the number of scores in our sample, many distributions of interest to psychologists become symmetrical or bell shaped. (Sometimes the normal distribution is called the *bell curve*.) The majority of the scores cluster around the measure of central tendency, with fewer and fewer scores occurring as we move away from it. As you can see from Figure 9-10, the mean, median, and mode of a normal distribution all have the same value.

Normal distributions and standard deviations are related in interesting ways. For example, distances from the mean of a normal distribution can be measured in standard deviation units (*SD*). Consider a distribution with an *M* of 56 and an *SD* of 4; a score of 60 falls 1 *SD* above the mean (+1 *SD*), whereas a score of 48 is 2 *SD* below the mean (−2 *SD*), and so on. As you can see from Figure 9-11, 34.13% of all the scores in *all* normal distributions occur between the mean and 1 *SD* above the mean.

Likewise, 34.13% of all the scores in a distribution occur between the mean and 1 *SD* below the mean. Another 13.59% of the scores occur between 1 and 2 *SD* above the mean; another 13.59% of the scores occur between 1 and 2 *SD* below the mean. Thus, slightly over 95% of all the scores in a normal distribution occur between 2 *SD* below the mean and 2 *SD*

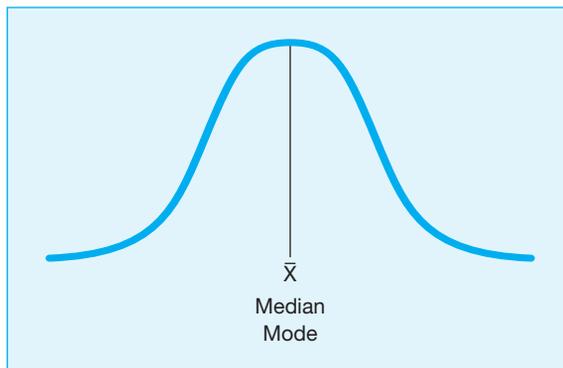


FIGURE 9-10 A Symmetrical or Bell-Shaped Normal Distribution. Note that the mean, median, and mode coincide in a normal distribution.

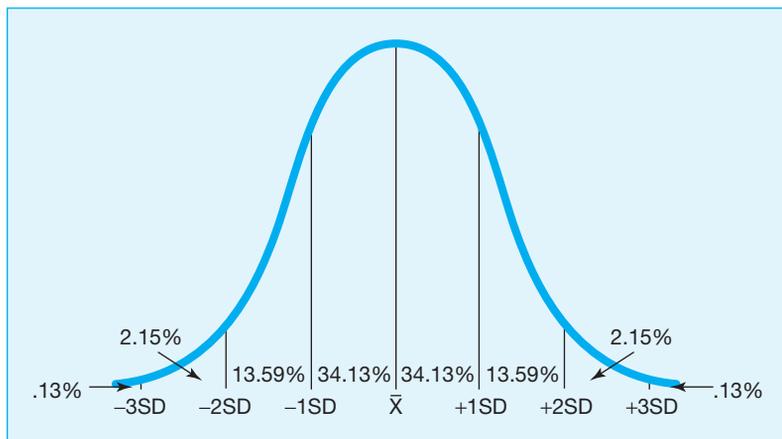


FIGURE 9-11 The Relation of Standard Deviations and the Normal Distribution.

above the mean. Exactly 2.28% of the scores occur *beyond* 2 SD above the mean; another 2.28% of the scores occur beyond 2 SD below the mean. It is important to remember that these percentages hold true for *all* normal distributions.



Review Figure 9-11 for a moment. Why isn't the percentage of scores from 0 to 1 SD above (or below) the mean the same as the percentage of scores from 1 to 2 SD above (or below) the mean?

As we move away from the mean (either above or below), the scores become progressively different from the mean. Because larger scores occur less frequently and the scores between 1 and 2 SD are larger than those from 0 to 1 SD, the percentage of scores from 1 to 2 SD will be lower than the percentage of scores from 0 to 1 SD.

Now, let's return to your test score of 64. You know the mean of the class is 56. If the instructor also tells you that the $SD = 4$, what would your reaction be? Your score of 64 would be 2 SD above the mean; you should feel pretty good. Your score of 64 puts you in the top 2.28% of the class (100% minus 50% of the scores below the mean and minus 34.13% from the mean to 1 SD above the mean and minus 13.59% that occurs between 1 and 2 SD above the mean). (See Figure 9-12A.)

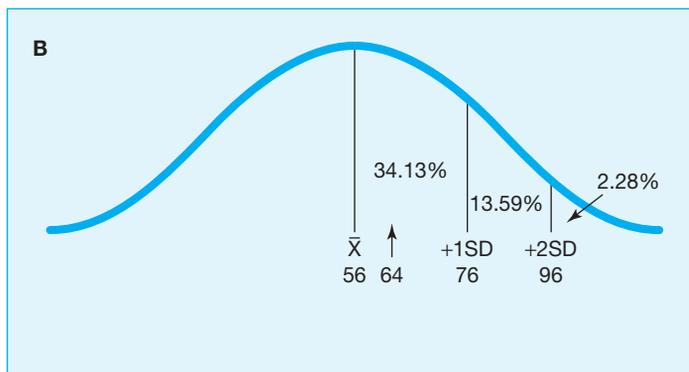
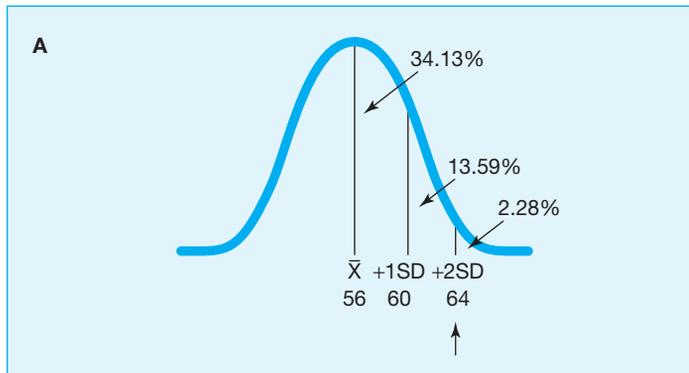


FIGURE 9-12 **A.** A score of 64 is exceptionally good when the mean is 56 and the SD is 4. **B.** The same score is not as highly regarded when the SD is 20.

What if your instructor told you the SD was 20? Now your score of 64 does not stand up as well as it did when the SD was 4. You are above the mean but a long way from being even 1 SD above the mean (see Figure 9-12B).

Because the percentage of the scores that occurs from the mean to the various SD units is the same for *all* normal distributions, we can compare scores from different distributions by discussing them in terms of standard deviations above or below the mean. Consider the following scores:

| Test # | Your Score | M | SD | Relation of Your Score to the M |
|--------|------------|-----|------|-----------------------------------|
| 1 | 46 | 41 | 5 | 1 SD above |
| 2 | 72 | 63 | 4 | Over 2 SD above |
| 3 | 93 | 71 | 15 | Over 1 SD above |

Even though your scores, the means, and the standard deviation values differ considerably, we can determine how many SD units away from the mean each of your scores is. In turn, we can compare these differences. When these comparisons are made, we find that your scores are consistently 1 SD or more above the mean. Thus, you are consistently in at least the top 15.87% of the class (100% minus 50% of the scores below the mean and minus 34.13% of the scores from the mean to 1 SD above the mean). By comparing scores from various distributions in this manner, we are able to see patterns and suggest what might occur in the future. Another type of descriptive statistic, the correlation coefficient, is also used for predictive purposes. We turn to this topic next.



Courtesy of Sidney Harris.

Unlike the warning on this truck, psychologists view data and statistical procedures as tools to help answer research questions.

■ REVIEW SUMMARY

1. **Statistics** involves the collection, analysis, and interpretation of data.
2. **Measurement** is the assignment of symbols to events according to a set of rules. A **scale of measurement** is a particular set of measurement rules.
3. A **nominal scale** is a simple classification system, whereas events can be rank ordered when an **ordinal scale** is used. Equal intervals separate rank-ordered events in an **interval scale**. The addition of a true zero to an interval scale results in a **ratio scale**.
4. **Descriptive statistics**, which summarize sets of numbers, include measures of central tendency and variability.
5. The **mode** is the most frequent score, whereas the **median** divides a distribution into two equal halves. The **mean** is the arithmetic average. Depending on the nature of the distribution, these measures of central tendency may not reflect the typical score equally well. They are, however, identical in a normal distribution.
6. Graphs, such as the **pie chart**, **bar graph**, **histogram**, and **frequency polygon**, are often used to depict frequencies or percentages.
7. The **line graph** is used to depict experimental results. The DV is plotted on the vertical (y) axis (**ordinate**), and the IV is plotted on the horizontal (x) axis (**abscissa**). A 2:3 relation of y to x axes produces a representative figure.
8. Measures of **variability** include the **range** (difference between high and low scores) and **standard deviation** (*SD*; square root of the variance). The **variance** is a single number that represents the total amount of variability that is present.
9. The standard deviation conveys considerable information about the **normal distribution** that is under consideration.

■ Check Your Progress

1. Matching

| | |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ol style="list-style-type: none"> 1. inferential statistics 2. descriptive statistics 3. measurement 4. nominal scale 5. ordinal scale 6. interval scale 7. ratio scale | <ol style="list-style-type: none"> A. assignment of symbols to events B. rank order C. putting events into categories D. equal intervals plus a true zero E. used to summarize a set of numbers F. equal intervals G. used to analyze data after an experiment |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
2. The number that occurs most frequently is the

| | | | |
|---------|-----------|---------|-------------|
| a. mean | b. median | c. mode | d. harmonic |
|---------|-----------|---------|-------------|
3. When you are dealing with a normal distribution of scores, which measure of central tendency is preferred? Why?
4. A _____ presents data in terms of frequencies per category.

| | | | |
|--------------|---------------|-------------------------|--------------|
| a. pie chart | b. line graph | c. bimodal distribution | d. histogram |
|--------------|---------------|-------------------------|--------------|

5. You are constructing a line graph to depict the results of an experiment you just completed. What is the ordinate? What is the abscissa? What will be plotted on each of them?
6. Why does the range not convey much information about variability?
7. The _____ is a single number that represents the total amount of variability in a distribution.
 - a. variance
 - b. standard deviation
 - c. range
 - d. mean
8. How does the standard deviation relate to the normal curve?

Correlation

Just as it does in the successful completion of a detective case, prediction plays an important role in psychology. Nowhere is this aspect of psychology more apparent than when moving from high school to college. You probably took a college entrance examination while you were in high school. Based on the results of this exam, a prediction about your grades in college was made. Similarly, should you plan to go on for graduate training after you complete your undergraduate degree, you probably will take another entrance examination. Depending upon your area of interest, you might take the Graduate Record Examination (GRE), the Law School Admission Test (LSAT), the Medical College Admission Test (MCAT), or some other similar test.

Such predictions are based on the correlation coefficient. Sir Francis Galton (1822–1911) developed the basic ideas of correlation. Galton, who was independently wealthy, devoted his time to studying and investigating those things that interested him. According to E. G. Boring (1950), the eminent historian of psychology, Galton “was a free-lance and a gentleman scientist. He was forever seeing new relationships and working them out, either on paper or in practice. No field was beyond his possible interest, no domain was marked off in advance as being out of his province” (p. 461). For example, Galton studied such varied topics as the weather and fingerprints. He also proposed that a person’s intelligence was directly related to the quality of the nervous system: The better the nervous system, the higher the intelligence. To be able to measure the predictive relation between these two variables, Galton’s assistant, Karl Pearson (1857–1936), developed the correlation coefficient.

A **correlation coefficient** is a single number that represents the degree of relation (i.e., “co-relation”) between two variables. The value of a correlation coefficient can range from -1 to $+1$.

A correlation coefficient of -1 indicates that there is a *perfect negative relation* (see Figure 9-13) between the two variables of interest. That is, whenever we see an increase of 1 unit in one variable, there is always a proportional decrease in the

other variable.

Consider the following scores on Tests X and Y:

| | Test X | Test Y |
|-----------|--------|--------|
| Student 1 | 49 | 63 |
| Student 2 | 50 | 61 |
| Student 3 | 51 | 59 |
| Student 4 | 52 | 57 |
| Student 5 | 53 | 55 |

Correlation coefficient

A single number representing the degree of relation between two variables.

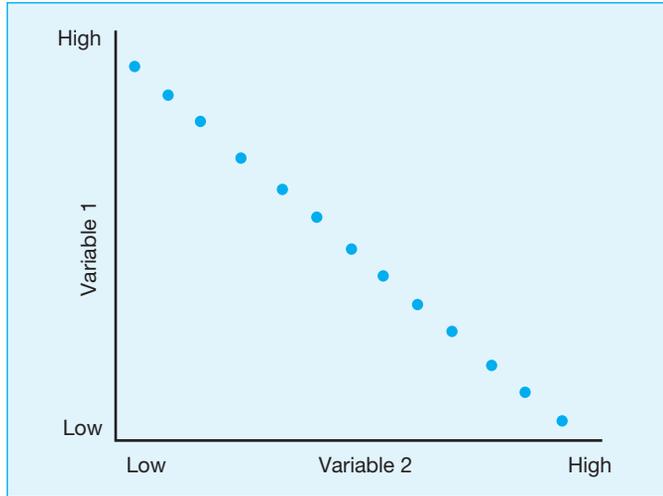


FIGURE 9-13 A Perfect Negative Correlation.

For each unit of increase in a score on Test X, there is a corresponding *decrease* of 2 units in the score on Test Y. Given this information, you are able to predict that if Student 6 scores 54 on Test X, that student’s score on Test Y will be 53.

As you saw in Chapter 4, a *zero correlation* means that there is *little or no relation* between the two variables (see Figure 9-14). As scores on one variable increase, scores on the other variable may increase, decrease, or be the same. Hence, we are not able to predict how you will do on Test Y by knowing your score on Test X. A correlation coefficient does not have to

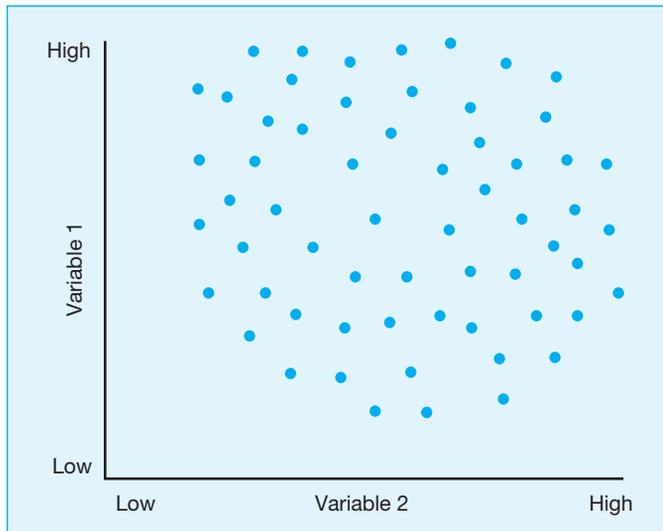


FIGURE 9-14 A Zero Correlation.

be exactly 0 to be considered a zero correlation. The inability to make good predictions is the key consideration. Two sets of scores having a near-zero correlation might look like this:

| | Test X | Test Y |
|-----------|--------|--------|
| Student 1 | 58 | 28 |
| Student 2 | 59 | 97 |
| Student 3 | 60 | 63 |
| Student 4 | 61 | 60 |
| Student 5 | 62 | 50 |

In this case the correlation between Test X and Test Y is 0.04. A correlation that small indicates that you will not be able to predict Test Y scores by knowing Test X scores; you are dealing with a zero correlation or no relation.

A correlation coefficient of +1 indicates that there is a *perfect positive relation* between the two sets of scores (see Figure 9-15). That is, when we see an increase of 1 unit in one variable, we always see a proportional increase in the other variable. Consider the following scores on Tests X and Y:

| | Test X | Test Y |
|-----------|--------|--------|
| Student 1 | 25 | 40 |
| Student 2 | 26 | 43 |
| Student 3 | 27 | 46 |
| Student 4 | 28 | 49 |
| Student 5 | 29 | 52 |

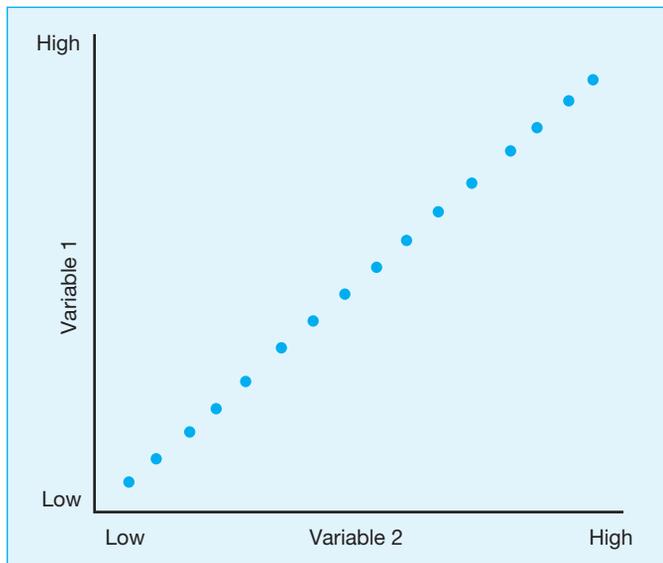


FIGURE 9-15 A Perfect Positive Correlation.

In this example, there is an *increase* of 3 units in the score on Test Y for every unit increase on Test X. The perfect positive correlation leads you to predict that if Student 6 scores 30 on Test X, then his or her score on Test Y will be 55.



Now that we have reviewed the basics of correlation, we would like you to think about the following question. Do perfect (either positive or negative) correlations occur frequently in the real world? Why or why not?

The existence of a perfect correlation indicates that there are no other factors present that influence the relation we are measuring. This situation rarely occurs in real life. Think about correlating test scores. To obtain a perfect correlation, all of the participants would have to study and learn the same amount for each test. This situation is not likely to occur. Such factors as fatigue, illness, boredom, and distractions will likely have an effect and make the correlation less than perfect.

The Pearson Product–Moment Correlation Coefficient

The most common measure of correlation is the Pearson product–moment correlation coefficient (r), which was developed by Galton’s assistant Karl Pearson. This type of correlation coefficient is calculated when both the X variable and the Y variable are interval or ratio scale measurements and the data appear to be linear. Other correlation coefficients can be calculated when one or both of the variables are not interval or ratio scale measurements or when the data do not fall on a straight line.

Examples of computer printouts for perfect positive and perfect negative correlations appear in Table 9-2. As you can see, the correlation of Tests X and Y with themselves is always

TABLE 9-2 Computer Printout for (A) Perfect Negative Correlation and (B) Perfect Positive Correlation

A. Perfect Negative Correlation

PEARSON CORRELATION MATRIX

| | TEST X | TEST Y |
|--------|--------|--------|
| TEST X | 1.00 | |
| TEST Y | –1.00 | 1.00 |

NUMBER OF OBSERVATIONS: 5

B. Perfect Positive Correlation

PEARSON CORRELATION MATRIX

| | TEST X | TEST Y |
|--------|--------|--------|
| TEST X | 1.00 | |
| TEST Y | 1.00 | 1.00 |

NUMBER OF OBSERVATIONS: 5

1.00; however, the correlation of Test X with Test Y is -1.00 (Table 9-2A) when the relation is perfect negative and $+1.00$ (Table 9-2B) when the relation is perfect positive. Because some computers do not provide a probability for correlations, you may have to consult a table of correlation probabilities to determine whether a particular correlation is significant (see Table A-3 in Appendix A).

Review Figures 9-13, 9-14, and 9-15; they will help you visualize the various correlations we have discussed. Perfect positive and perfect negative correlations always fall on a straight line, whereas nonperfect correlations do not. You will find, however, that the more the scores cluster close together and form a straight line, the stronger (i.e., larger) the correlation coefficient will be. For positive correlations, the trend of the points is from lower left to upper right, whereas for negative correlations, the trend is from upper left to lower right. There is no consistent pattern for a zero correlation.

Although descriptive statistics can tell us a great deal about the data we have collected, they cannot tell us everything. For example, when we conduct an experiment, descriptive statistics cannot tell us whether the IV we manipulated had a significant effect on the behavior of the participants we tested or whether the results we obtained would have occurred by chance. To make such determinations we must conduct an inferential statistical test.

Inferential Statistics

After you have conducted an experiment, you perform a statistical test on the data that you have gathered. The results of this test will help you decide whether the IV was effective. In other words, we shall decide whether our statistical result is *significant*.

What Is Significant?

An inferential statistical test can tell us whether the results of an experiment would occur frequently or rarely by chance. Inferential statistics with small values occur *frequently by chance*, whereas inferential statistics with large values occur *rarely by chance*. If the result occurs often by chance, we say that it is not significant and conclude that our IV did not affect the DV. In this case we would accept the **null hypothesis**, which says that the differences between groups are due to chance (i.e., not the operation of the IV). If, however, the result of our inferential statistical test occurs rarely by chance (i.e., it is significant), we can conclude that some factor other than chance is operative. If we have conducted our experiment properly and exercised good control (see Chapters 6 and 7), then our significant statistical result gives us reason to believe the IV we manipulated was effective (i.e., did affect the DV scores).

Null hypothesis A hypothesis that says that all differences between groups are due to chance (i.e., not the operation of the IV).

When do we consider that an event occurs rarely by chance? Traditionally, psychologists say that any event that occurs by chance alone 5 times or fewer in 100 occasions is a rare event. Thus, you will see frequent mention of the “.05 level of significance” in journal articles. This statement means that a result is considered significant if it would occur 5 or fewer times by chance in 100 replications of the experiment when the null hypothesis is true. As the experimenter, you decide on the level of significance before the experiment is conducted.

You will encounter several significance tests in later chapters in this book. For the present we will use the *t* test to illustrate their use.



Reprinted by permission of Warren Street.

The t Test

For years you have heard the old saying that “clothes make the person.” You decide to test this adage experimentally by determining whether type of clothing influences the time it takes a salesperson to wait on customers. To set up the study, imagine that you randomly select 16 salesclerks from a store at the mall and then randomly assign these clerks to one of two groups of eight clerks. Group A will wait on customers in dressy clothes; Group B will wait on customers in sloppy clothes. Because you formed the groups randomly at the start of the experiment, you assume they are comparable before they are exposed to the IV.

The students in Group A wear dressy clothes to the shopping mall, whereas the students in Group B wear sloppy clothes to the shopping mall. Because the students in Group A have no relation to, or effect on, the students in Group B, these groups are *independent* of each other. Each student enters a store in the mall and uses a silent, inconspicuous stopwatch to measure the time (in seconds) it takes a salesperson to offer service. (Keep in mind that these data were *recorded* by the student shoppers; the salesclerks actually produced the

data.) The “latency-to-service” scores (this is a latency DV—see Chapter 6) for the two groups appear below.

| Group A (Dressy Clothes) | Group B (Sloppy Clothes) |
|-----------------------------|-----------------------------|
| 37 | 50 |
| 38 | 46 |
| 44 | 62 |
| 47 | 52 |
| 49 | 74 |
| 49 | 69 |
| 54 | 77 |
| 69 | 76 |
| $\Sigma X = 387$ | $\Sigma Y = 506$ |
| $M = 48.38$ | $M = 63.25$ |

Do you think the clerks waited on the students in dressy clothes more quickly? Just looking at the differences between the groups suggests that this might be the case; the mean score of Group B is higher than that of Group A. (Higher scores reflect longer times before a salesperson offered service.) On the other hand, there is considerable overlap between the two groups; several of the latency-to-service scores were similar for students dressed in sloppy clothes and the students in the dressy clothes. Is the difference you obtained large enough to be genuine, or is it just a chance happening? Merely looking at the results will not answer that question.

***t* test** An inferential statistical test used to evaluate the difference between two means.

The ***t* test** is an inferential statistical test used to evaluate the difference between the means of *two groups* (see Chapter 10 for research designs using two groups). Because the two groups in our latency-to-service experiment were independent, we will use an independent-groups *t* test. (We discuss the correlated-groups *t* test in Chapter 10.) The computer printout for our *t* test appears in Table 9-3.

You can see that our *t* value is 2.61 and that the probability of this *t* value is .021. Because the probability of this result occurring by chance is less than .05, we can conclude that the two groups differ significantly.

If your computer program does not provide the probability of your result as part of the printout, you will have to make this determination by yourself. Recall that our *t* value is 2.61.

TABLE 9-3 Computer Printout for Independent-Groups *t* Test

| GROUP | <i>N</i> | <i>M</i> | <i>SD</i> |
|------------|----------|-----------|------------|
| Dressy | 8 | 48.38 | 9.46 |
| Sloppy | 8 | 63.25 | 11.73 |
| $t = 2.61$ | | $df = 14$ | $p = .021$ |

After we have obtained our t value, we must follow several steps in order to interpret its meaning:

1. Determine the degrees of freedom (df) involved. (Because some statistical packages may not automatically print the degrees of freedom for you, it is important to keep this formula handy.) For our clothing research:

$$\begin{aligned} df &= (N_A - 1) + (N_B - 1) \\ &= (8 - 1) + (8 - 1) \\ &= 14 \end{aligned}$$

2. We use the degrees of freedom (we will discuss the meaning of degrees of freedom after we have completed the problem) to enter a t table (see Table A-1 in Appendix A). This table contains the t values that occur by chance. We will compare our t value to these chance values. To be significant, the calculated t must be equal to or larger than the one given in Table A-1.
3. We enter the t table on the row for 14 degrees of freedom. By reading across this row we find that a value of 2.145 occurs by chance 5% of the time (.05 level of significance). Because our value of 2.61 is larger than 2.145 (the .05 value in the table for 14 df), we can conclude that our result is significant (has a probability of occurring by chance less than .05). Thus, the type of clothing had a significant effect on latency to service. This result is one that occurs fewer than 5 times in 100 by chance. Had we chosen a different level of significance, such as once in 100 occurrences (.01), the table value would have been 2.977, and we would have concluded that our result is not significant. In many instances, your computer program will print the probability of your t statistic automatically, and you will not have to consult the t table.

Although it is easy to follow a formula to calculate the degrees of freedom, the meaning of this term may not be clear, even if you have already had an introductory statistics course. We will try to help you understand its meaning. By **degrees of freedom** we mean the ability of a number in a given set to assume any value. This ability is influenced by the restrictions imposed on the set of numbers. For every restriction, one number is determined and will assume a fixed or specified value. For example, assume we have a set of 10 numbers and we know the sum of these numbers to be 100. Knowing that the sum is 100 is a restriction; hence, one of the numbers will be determined or fixed. In Example 1 below, the last number must be 15 because the total of the first 9 numbers (which can assume any value) is 85. In Example 2, the first 9 numbers have assumed different values. What is the value of the last number?

Degrees of freedom The ability of a number in a specified set to assume any value.

| Numbers | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Sum |
|-----------|----|----|----|---|---|----|----|---|----|----|-----|
| Example 1 | 6 | 12 | 11 | 4 | 9 | 9 | 14 | 3 | 17 | 15 | 100 |
| Example 2 | 21 | 2 | 9 | 7 | 3 | 18 | 6 | 4 | 5 | ? | 100 |

As in the first example, the first nine numbers can assume any value. In this example, the sum of the first nine numbers is 75. That means that the value of the last number is fixed at 25.

One-Tail Versus Two-Tail Tests of Significance

Recall from Chapter 5 that you state your experimental hypothesis in either a directional or a nondirectional manner. If you use the directional form, you are specifying exactly how (i.e., the direction) the results will turn out. For the example we have been considering, an experimental hypothesis, stated in general implication form (see Chapter 5), might be as follows:

If students wear dressy clothes to the shopping mall, then the time it takes a salesperson to offer to serve them will be shorter than the latency to service for students dressed in sloppy clothes.

Because we predict that the latency to service for the students wearing dressy clothes will be shorter than that for the students wearing sloppy clothes, we have a directional hypothesis. If we simply indicate that we expect a difference between the two groups and do not specify the exact nature of that difference, then we are using a nondirectional hypothesis.

Now, how do directional and nondirectional hypotheses relate to the t test? If you remember discussing one-tail and two-tail tests of significance in your statistics class, you're on the right track. A one-tail t test evaluates the probability of only one type of outcome, whereas the two-tail t test evaluates the probability of both possible outcomes. If you've associated directional hypotheses with one-tail tests and nondirectional hypotheses with two-tail tests, you're right again.

Figure 9-16 depicts the relation between the type of experimental hypothesis (directional versus nondirectional) and the type of t test used (one-tail versus two-tail). As you can see, the region of rejection is larger and occurs only in one tail of the distribution when a one-tail test is conducted (Figure 9-16A). The probability of the result's occurring by chance alone is split in half and distributed equally to the two tails of the distribution when a two-tail test is conducted (Figure 9-16B).

Although the calculations for a one-tail test of significance and a two-tail test of significance are the same, you would consult different columns in the t table. For the shopping center example, we conducted a two-tail test of significance; 2.145 was our critical value at the .05 level of significance. Hence, a t value equal to or greater than 2.145 is significant (see Table A-1 in Appendix A). Had we done a one-tail test, our critical value at the .05 level of significance would have been 1.761 (see Table A-1).

Because a lower value is required for significance with a one-tail test of significance, it is somewhat easier to find a significant result. If this is the case, why don't experimenters always state directional experimental hypotheses? The main reason is that researchers don't always know exactly how an experiment will turn out. If we knew the outcome of each experiment before it was conducted, there would be no need to do the experiment. If you state a directional hypothesis and then obtain the opposite result, you have to reject your hypothesis. Had you stated a nondirectional hypothesis, your experiment would have confirmed your prediction that there would be a difference between the groups. When conducting a t test, researchers are usually interested in either outcome; for example, what if casually dressed students were actually waited on more quickly?

The Logic of Significance Testing

Remember, we consider the result of an experiment to be statistically significant when it occurs rarely by chance. In such instances we assume that our IV produced the results.

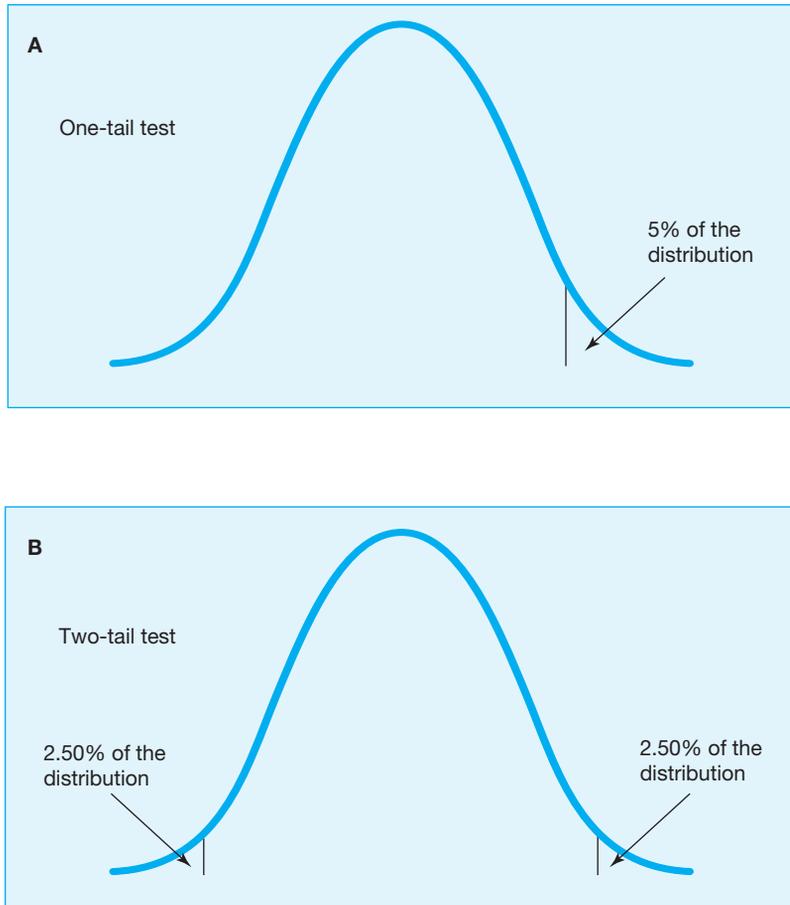


FIGURE 9-16 Regions of Rejection for (A) a One-Tail Test and (B) a Two-Tail Test. **A.** With a one-tail test the region of rejection of the null hypothesis is located in one tail of the distribution. Directional hypotheses, such as $A > B$, are associated with one-tail tests. **B.** With a two-tail test the region of rejection of the null hypothesis is distributed evenly to both tails of the distribution. Nondirectional hypotheses, such as $A \neq B$ (A does not equal B), are associated with two-tail tests.

Although Sherlock Holmes wasn't speaking of a psychological experiment, he captured the intent of significance testing when he asked, "How often have I said to you that when you have eliminated the impossible, whatever remains, *however improbable*, must be the truth?" (Doyle, 1927, p. 111).

Typically our ultimate interest is not in the samples we have tested in an experiment but in what these samples tell us about the population from which they were drawn. In short, we want to generalize, or *infer*, from our samples to the larger population.

We have diagrammed this logic in Figure 9-17. First, samples are randomly drawn from a specified population (Figure 9-17A). We assume that random selection has produced two equivalent groups: Any differences are due solely to chance factors. In Figure 9-17B we see

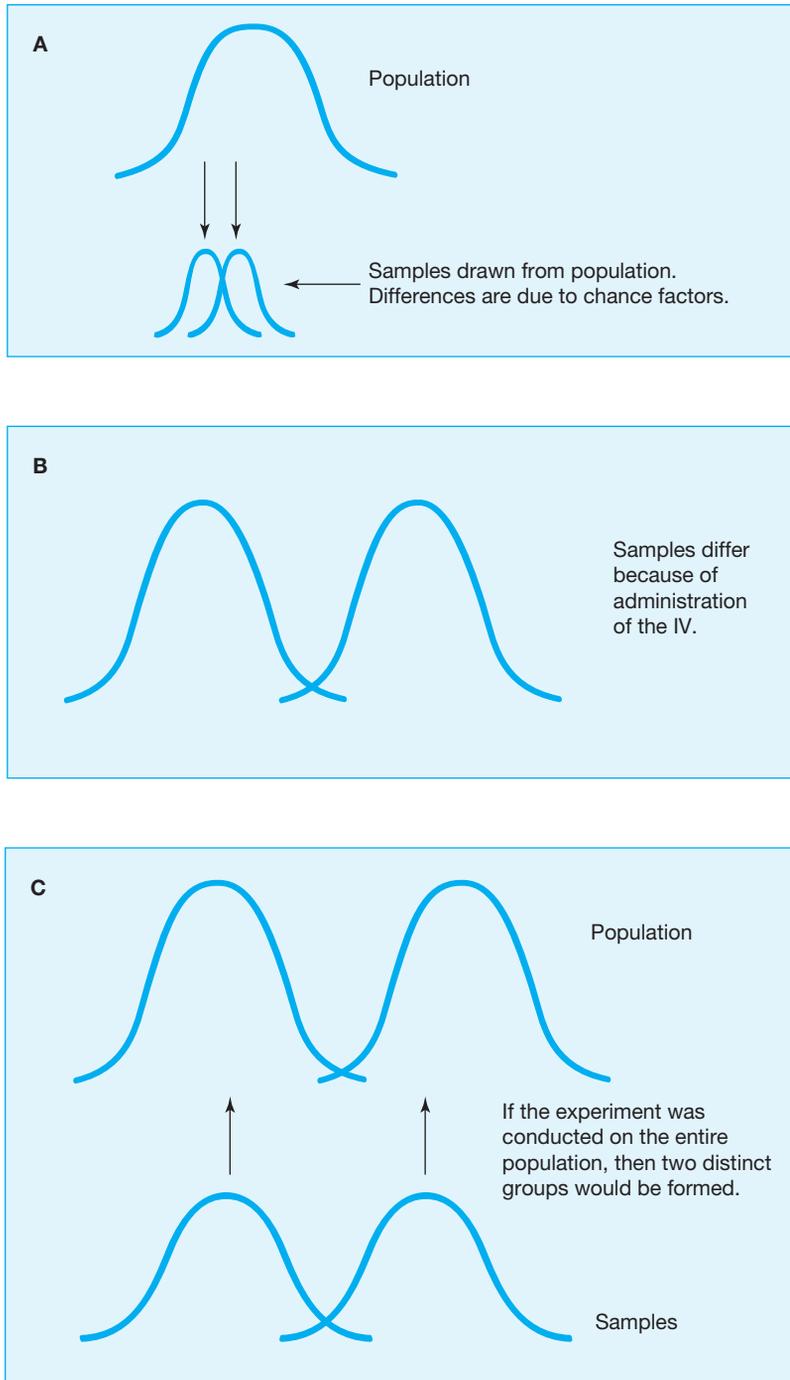


FIGURE 9-17 **A.** Random samples are drawn from a population. **B.** The administration of the IV causes the samples to differ significantly. **C.** The experimenter generalizes the results of the experiment to the general population.

the results of our experiment; the manipulation of the IV caused the groups to be significantly different. At this point generalization begins. Based on the significant difference that exists between the groups, we *infer* what would happen if our treatments were administered to all individuals in the population. In Figure 9-17C we have generalized from the results of our research using two samples to the entire population (see Chapter 8). We are inferring that two separate groups would be created in the population as a result of the administration of our IV.

When Statistics Go Astray: Type I and Type II Errors

Unfortunately, not all our inferences will be correct. Recall that we have determined that an experimental result is significant when it occurs rarely by chance (i.e., 5 times or less in 100). There always is the chance that your experiment represents 1 of those 5 times in 100 when the results did occur *by chance*. Hence, the null hypothesis is true, and you will make an error in accepting your experimental hypothesis. We call this faulty decision a **Type I error** (alpha, α). The experimenter directly controls the probability of making a Type I error by setting the significance level. For example, you are less likely to make a Type I error with a significance level of .01 than with a significance level of .05.

Type I error Accepting the experimental hypothesis when the null hypothesis is true.

On the other hand, the more extreme or critical you make the significance level (e.g., going from .05 to .01) to avoid a Type I error, the more likely you are to make a Type II or beta (β) error. A **Type II error** involves rejecting a *true* experimental hypothesis. Unlike Type I errors, Type II errors are not under the direct control of the experimenter. We can indirectly cut down on Type II errors by implementing techniques that will cause our groups to differ as much as possible. For example, using a strong IV and testing larger groups of participants are two techniques that will help avoid Type II errors.

Type II error Accepting the null hypothesis when the experimental hypothesis is true.

We will have more to say about Type I and Type II errors in subsequent chapters. They are summarized here as follows:

| | | TRUE STATE OF AFFAIRS | |
|---------------|---------------------------------|---------------------------------|---------------------------|
| | | Experimental Hypothesis Is True | Null Hypothesis Is True |
| YOUR DECISION | Experimental Hypothesis Is True | Correct Decision | Type I (α) Error |
| | Null Hypothesis Is True | Type II (β) Error | Correct Decision |

You should keep in mind that the typical practice is to set the alpha level at .05 because it places the probability of Type I and II errors at acceptable levels.

Effect Size

Effect size The magnitude or size of the experimental treatment.

Before concluding this chapter, we want to introduce you to a statistical concept and procedure that currently is gaining in popularity and frequency of usage. **Effect size** is a statistical measure that conveys information concerning the *magnitude* of the effect produced by the IV.



Doesn't obtaining significance with an inferential test give us the same information? After all, significance indicates the IV had an effect, and that's our concern. Why do we need anything else?

Unfortunately, a significant statistical test tells us only that the IV had an effect; it does not tell us about the size of the significant effect. Moreover, whether an effect is significant or not may depend on factors other than the IV. For example, you just saw that you are more likely to obtain significance (i.e., avoid a Type II error) when you use larger samples, even though the influence of the IV remains the same. The American Psychological Association's *Publication Manual* (2001) indicates that "neither of the two types of probability values [your selected alpha level and the probability level associated with the inferential statistic you calculate] reflects the importance or magnitude of an effect because both depend on sample size" (p. 18).

Such considerations have encouraged researchers to report effect size in addition to the inferential statistics that are used. In fact, some statisticians (e.g., Kirk, 1996) envision a time when the reporting of effect size will be more common than significance testing. Indeed, the APA's *Publication Manual* (2001) states, "You are encouraged to provide effect-size information" (p. 18) when you prepare your research report.

There are several different ways to calculate effect size. Here are two that should give you no problems. Cohen's *d* (Cohen, 1977) is easy to compute when you use two groups and calculate a *t* test. In these cases:

$$d = \frac{t(N_1 + N_2)}{\sqrt{df} \sqrt{N_1 N_2}}$$

or, when the two samples are of equal size:

$$d = \frac{2t}{\sqrt{df}}$$

Cohen (1977) indicated that $d = .20$ to $.50$ is a small effect size, $d = .50$ to $.80$ is a medium effect size, and d values greater than $.80$ reflect large effect sizes.

A second technique for determining effect size is appropriate when you calculate a Pearson product-moment correlation (r): r^2 gives you an estimate of the proportion of the variance accounted for by the correlation in question (Rosenthal & Rosnow, 1984). For example, even though $r = .30$ is significant ($p < .01$) with 90 pairs of scores, this correlation accounts for only 9% ($.30^2 = .09 = 9\%$) of the variance. This figure means that 91% of the variability in your research results is accounted for by *other* variables, a rather small effect size indeed.

■ REVIEW SUMMARY

1. A **correlation coefficient** is a single number that represents the degree of relationship between two variables. Many predictions are based on correlations.
2. A perfect negative correlation (-1) exists when an increase of 1 unit in one variable is always accompanied by a proportional decrease in the other variable. A perfect positive correlation ($+1$) exists when an increase of 1 unit is always accompanied by a proportional increase in the other variable. A correlation of 0 indicates that there is no relation between the variables under consideration.
3. The Pearson product-moment correlation coefficient is calculated when both variables are interval-scale measurements.
4. Inferential statistics help the experimenter decide whether the IV was effective. A significant inferential statistic is one that occurs rarely by chance.
5. The **t test**, which is an inferential statistic, is used to test the differences between two groups.
6. When results are significant, the experimenter hopes to be able to extend the results of the experiment to the more general population.
7. A one-tail t test is conducted when a directional hypothesis is stated, whereas a two-tail t test is conducted when a nondirectional hypothesis is stated.
8. Even though lower critical values are associated with one-tail tests, making it easier to attain significance, most experimental hypotheses are nondirectional because the researchers do not know exactly what the research will show.
9. Sometimes the results of an inferential statistical test produce an incorrect decision. An experimental hypothesis may be incorrectly accepted (**Type I error**) or incorrectly rejected (**Type II error**).

■ Check Your Progress

1. Matching

| | |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ol style="list-style-type: none"> 1. correlation coefficient 2. perfect negative correlation 3. perfect positive correlation 4. significant 5. inferential statistics 6. Type I error 7. Type II error 8. one-tail test 9. two-tail test | <ol style="list-style-type: none"> A. nondirectional hypothesis B. result occurs infrequently by chance C. rejecting a true null hypothesis D. tests conducted to determine whether the IV had an effect E. directional hypothesis F. represents the degree of relationship between two variables G. rejecting a true experimental hypothesis H. -1 I. $+1$ |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
2. Explain the difference between a positive correlation and a perfect positive correlation.

3. What does a zero correlation signify?
4. Explain the logic involved when an independent-groups t test is conducted.
5. What is meant by “level of significance”? How is the level of significance determined?
6. If it is easier to obtain a significant result with a one-tail test, why would an experimenter ever state a nondirectional experimental hypothesis and thus use a two-tail test?
7. A one-tail test of significance is associated with a
 - a. directional hypothesis
 - b. nondirectional hypothesis
 - c. positive correlation
 - d. negative correlation
8. The Type I error
 - a. is under the direct control of the experimenter
 - b. always occurs 5% of the time
 - c. is specified by the experimental hypothesis
 - d. all of the above
 - e. none of the above
9. If you could compare all men and women in the world, you would find that men are significantly more aggressive. You conduct an experiment and find no difference in aggression between men and women. You have made a
 - a. correct decision
 - b. Type I error
 - c. Type II error
 - d. Type III error

■ Key Terms

| | | |
|-----------------------------|-------------------------|------------------------------|
| Statistics, 171 | Pie chart, 176 | Normal distribution, 184 |
| Descriptive statistics, 171 | Histogram, 177 | Correlation coefficient, 188 |
| Inferential statistics, 171 | Bar graph, 177 | Null hypothesis, 192 |
| Measurement, 172 | Frequency polygon, 178 | t test, 194 |
| Scale of measurement, 172 | Line graph, 178 | Degrees of freedom, 195 |
| Nominal scale, 172 | Ordinate, 178 | Type I error, 199 |
| Ordinal scale, 172 | Abscissa, 178 | Type II error, 199 |
| Interval scale, 172 | Variability, 182 | Effect size, 200 |
| Ratio scale, 173 | Range, 182 | |
| Mode, 174 | Variance, 182 | |
| Median, 174 | Standard deviation, 183 | |
| Mean, 174 | | |

■ Looking Ahead

So far we have considered sources of researchable problems (Chapter 1), developed an experimental hypothesis (Chapter 2), considered the ethics involved in conducting research (Chapter 2), scrutinized our experiment for possible extraneous variables and nuisance variables (Chapter 6), and implemented control procedures to deal with these extraneous variables (Chapters 6 and 7). Now we are ready to combine all of these elements in an experimental design. In Chapter 10 experimental designs involving the use of two groups of participants are considered. We will consider more complex designs in subsequent chapters.