In Chapter 2 we examined a number of different statistics and saw how they might be used to describe a set of data or to represent the frequency of the occurrence of some event. Although the description of the data is important and fundamental to any analysis, it is not sufficient to answer many of the most interesting problems we encounter. In a typical experiment, we might treat one group of people in a special way and wish to see whether their scores differ from the scores of people in general. Or we might offer a treatment to one group but not to a control group and wish to compare the means of the two groups on some variable. Descriptive statistics will not tell us, for example, whether the difference between a sample mean and a hypothetical population mean, or the difference between two obtained sample means, is small enough to be explained by chance alone or whether it represents a true difference that might be attributable to the effect of our experimental treatment(s).

Statisticians frequently use phrases such as "variability due to chance" or "sampling error" and assume that you know what they mean. Probably you do; but if you do not, you are headed for confusion in the remainder of this book unless we spend a minute clarifying the meaning of these terms. We will begin with a simple example.

In Chapter 3 we considered the distribution of Total Behavior Problem scores from Achenbach's Youth Self-Report form. Total Behavior Problem scores are normally distributed in the population (i.e., the complete population of such scores is approximately normally distributed) with a population mean (μ) of 50 and a population standard deviation (σ) of 10. We know that different children show different levels of problem behaviors and therefore have different scores. We also know that if we took a sample of children, their sample mean would probably not equal exactly 50. One sample of children might have a mean of 49, while a second sample might have a mean of 52.3. The actual sample means would depend on the particular children who happened to be included in the sample. This expected variability that we see from sample to sample is what is meant when we speak of "variability due to chance." We are referring to the fact that statistics (in this case, means) obtained from samples naturally vary from one sample to another.

sampling error

Along the same lines, the term **sampling error** often is used in this context as a synonym for variability due to chance. It indicates that the value of a sample statistic probably will be in error (i.e., will deviate from the parameter it is estimating) as a result of the particular observations that happened to be included in the sample. In this context "error" does not imply carelessness or mistakes. In the case of behavior problems, one random sample might just happen to include an unusually obnoxious child, whereas another sample might happen to include an unusual number of relatively well-behaved children.

4.1 Two Simple Examples Involving Course Evaluations and Rude Motorists

One example that we will investigate when we discuss correlation and regression looks at the relationship between how students evaluate a course and the grade they expect to receive in that course. Many faculty feel strongly about this topic because

even the best instructors turn to the semiannual course evaluation forms with some trepidation—perhaps the same amount of trepidation with which many students open their grade report form. Some faculty think that a course is good or bad independently of how well a student feels he or she will do in terms of a grade. Others feel that a student who seldom came to class and who will do poorly as a result will also (unfairly?) rate the course as poor. Finally, there are those who argue that students who do well and experience success take something away from the course other than just a grade and that those students will generally rate the course highly. But the relationship between course ratings and student performance is an empirical question and, as such, can be answered by looking at relevant data. Suppose that in a random sample of fifty courses we find a general trend for students in a course in which they expect to do well to rate the course highly, and for them to rate courses in which they expect to do poorly as low in overall quality. How do we tell whether this trend in our small data set is representative of a trend among students in general or just an odd result that would disappear if we ran the study over? (For your own interest, make your prediction of what kind of results we will find. We will return to this issue later.)

A second example comes from a study by Doob and Gross (1968), who investigated the influence of perceived social status. They found that if an old, beat-up (low-status) car failed to start when a traffic light turned green, 84% of the time the driver of the second car in line honked the horn. However, when the stopped car was an expensive, high-status car, only 50% of the time did the following driver honk. These results could be explained in one of two ways:

1. The difference between 84% in one sample and 50% in a second sample is attributable to sampling error (random variability among samples); therefore, we cannot conclude that perceived social status influences horn-honking behavior.
2. The difference between 84% and 50% is large and reliable. The difference is not attributable to sampling error; therefore we conclude that people are less likely to honk at drivers of high-status cars.

Although the statistical calculations required to answer this question are different from those used to answer the one about course evaluations (because the first deals with relationships and the second deals with proportions), the underlying logic is fundamentally the same.

hypothesis testing

These examples of course evaluations and horn honking are two kinds of questions that fall under the heading of **hypothesis testing.** This chapter is intended to present the theory of hypothesis testing in as general a way as possible, without going into the specific techniques or properties of any particular test. I will focus largely on the situation involving differences instead of the situation involving relationships, but the logic is basically the same. You will see additional material on examining relationships in Chapter 9.

The theory of hypothesis testing is so important in all that follows that a thorough understanding of it is essential. Many students who have had one or more courses in statistics and who know how to run a number of different statistical tests still do not have a basic knowledge of what it is they are doing. As a result they have difficulty interpreting statistical tables and must learn every new procedure in a step-by-step, rote fashion. This chapter is designed to avoid that difficulty by presenting the theory

in its most general sense, without the use of any formulae. You can learn the formulae later, after you understand *why* you might want to use them. Professional statisticians might fuss over the looseness of the definitions, but any looseness will be set right in subsequent chapters. Others may object that we are considering hypothesis testing before we consider the statistical procedures that produce the test. That is precisely the intent. The material covered here cuts across all statistical tests and can be discussed independently of them. Separating the material in this way enables you to concentrate on the underlying principles without worrying about the mechanics of calculation.

We need to be explicit about what the problem is here. The reason for having hypothesis testing in the first place is that data are ambiguous. Suppose that we want to decide whether larger classes receive lower student ratings. We all know that some large classes are terrific, and others are really dreadful. Similarly, there are both good and bad small classes. So if we collect data on large classes, for example, the mean of several large classes will depend to some extent on which large courses just happen to be included in our sample. If we reran our data collection with a new random sample of large classes, that mean would almost certainly be different. A similar situation applies for small classes. When we find a difference between the means of samples of large and small classes, we know that the difference would come out slightly differently if we collected new data. So a difference between the means is ambiguous—is it greater than zero because large classes are worse than small ones, or because of the particular samples we happened to pick? Well, if the difference is quite large, it probably reflects differences between small and large classes. If it is quite small, it probably reflects just random noise. But how large is "large?" That is the problem we are beginning to explore, and that is the subject of this chapter.

As important as hypothesis testing is to psychology and other disciplines, it has long been under attack by those who think that we need something better. A hypothesis test can only tell us that a relationship is reliable or it is not, or that a difference between two groups is probably not due to chance, or that it might be. It does not really evaluate the magnitude of the difference, or its general importance. The American Psychological Association recently put together a task force to look at the general issue of hypothesis tests, and its report is now available (Wilkinson, 1999; see also http://www.apa.org/journals/amp/amp548594.html). Further discussion of this issue was included in an excellent paper by Nickerson (2000). These two documents, and others that have been published over the past 10 years, argue that *in addition* to standard hypothesis testing, researchers have an obligation to report additional information, including measures of effect size, confidence limits, and related statistics. These recommendations have influenced the coverage of material in this book, and you will see more frequent references to effect size measures than in previous editions.

If we are going to look at either of the two examples laid out above, or at a third one to follow, we need to find some way of deciding whether we are looking at a small chance fluctuation between the horn-honking rates for low- and high-status cars or a difference that is sufficiently large for us to believe that people are much less likely to honk at those they consider higher in status. If the differences are small

enough to attribute to chance variability, effect size measures don't have a strong role to play. That is why the following discussion in this chapter will emphasize hypothesis testing, while recognizing the importance of other approaches.

4.2 Sampling Distributions

In addition to course evaluations and horn honking, we will add a third example, which is one that affects a large number of children in our society. Consider the situation in which we have five students from recently divorced households. These five children have a mean of 56 on the Achenbach Youth Self-Report scale of Total Behavior Problems. This mean is over half a standard deviation above the mean (50) in the general population, and we want to know whether this finding is sufficiently deviant for us to conclude that the stress associated with divorce tends to elicit behavior problems in children at higher than normal levels. Perhaps we just came up with a peculiar sample, and another sample of children from divorced households would show normal levels of behavior. (After all, we have only five children in our sample.) Or perhaps divorce is a sufficiently stressful event in children's lives to produce serious behavior problems. To answer this kind of question, we have to use what are called **sampling distributions,** which tell us specifically what degree of sample-to-sample variability we can expect by chance as a function of sampling error.

sampling distributions

The most basic concept underlying all statistical tests is the sampling distribution of a statistic. It is fair to say that if we did not have sampling distributions, we would not have any statistical tests. Roughly speaking, sampling distributions tell us what values we might (or might not) expect to obtain for a particular statistic under a set of predefined conditions (e.g., what the obtained mean of five children might be *if* the true mean of the population from which those children come is 50). In addition, the standard deviation of that distribution (known as the "standard error" of the distribution) reflects the variability that we would expect to find in the values of that statistic over repeated trials. Sampling distributions provide the opportunity to evaluate the likelihood (given the value of a sample statistic) that such predefined conditions actually exist.

Basically, the sampling distribution of a statistic can be thought of as the distribution of values obtained for that statistic over repeated sampling (i.e., running the experiment, or drawing samples, an unlimited number of times). Although sampling distributions are almost always derived mathematically, it is easier to understand what they represent if we consider how they could, in theory, be derived empirically with a simple sampling experiment.

sampling distribution of the mean

We will take as an illustration the **sampling distribution of the mean,** because it is the most easily understood and relates directly to the example of behavior problems. The sampling distribution of the mean is nothing more than the distribution of means of an infinite number of random samples drawn under certain specified conditions (e.g., under the condition that the true mean of our population is 50 and the standard deviation is 10). Suppose we have a population with a known mean and standard deviation ($\mu = 50$, $\sigma = 10$). Further suppose that we draw a very large number (theoretically an infinite number) of random samples from this population, each sample

consisting of five scores. For each sample we will calculate its mean, and when we finish drawing all the samples, we will plot the distribution of these *means*. Such a distribution would be a sampling distribution of the mean and might look like the one presented in Figure 4.1. We can see from this figure that sample means between 48 and 52, for example, are quite likely to occur when we sample five children at random. We also can see that it is extremely unlikely that we would draw from this population a sample of five observations with a sample mean as high as 70, although there is some (quite small) probability of doing so. The fact that we know the kinds of values to expect for the mean of a sample drawn from this population is going to allow us to turn the question around and ask whether an obtained sample mean can be taken as evidence in favor of the hypothesis that we actually are sampling from this population.

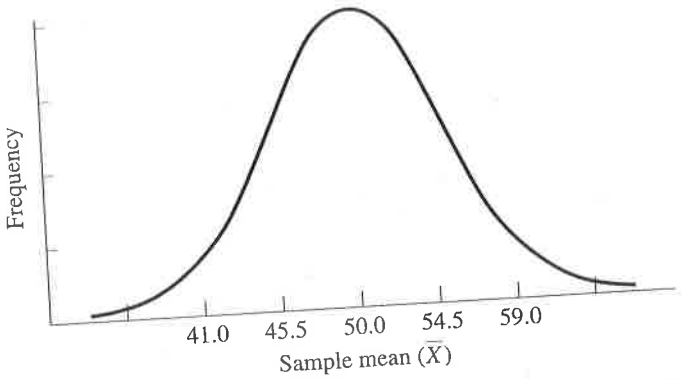

FIGURE 4.1 *Distribution of means of behavior problems, each based on 5 scores*

4.3 Hypothesis Testing

We do not go around obtaining sampling distributions, either mathematically or empirically, simply because they are interesting to look at. We have important reasons for doing so. The usual reason is that we want to test some hypothesis. Let's go back to the random sample of five highly stressed children with a mean behavior problem score of 56. We want to test the hypothesis that such a sample mean could reasonably have arisen had we drawn our sample from a population in which $\mu = 50$. This is another way of saying that we want to know whether the mean of stressed children is different from the mean of normal children. The only way we can test such a hypothesis is to have some idea of the probability of obtaining a sample mean as extreme as 56 *if* we actually sampled observations from a population in which $\mu = 50$. The answer to this question is precisely what a sampling distribution is designed to provide.

Suppose we obtained (constructed) the sampling distribution of the mean for samples of five children from a population whose mean (μ) is 50 (the distribution plotted in Figure 4.1). Suppose further we then determined from that distribution the probability of a sample mean as high as 56. For the sake of argument, suppose this probability is 0.15. Our reasoning could then go as follows: "If we did in fact sample from a population with $\mu = 50$, the probability of obtaining a sample mean as high as 56

is 0.15—a fairly likely event. Because a sample mean that high is often obtained from a population with a mean of 50, we have no reason to doubt that this sample came from such a population."

Alternatively, suppose we obtained a sample mean of 62 and calculated from the sampling distribution that the probability of a sample mean as high as 62, when the population mean is 50, was only 0.0188. Our argument could then go like this: "*If* we did sample from a population with $\mu = 50$, the probability of obtaining a sample mean as high as 62 is only 0.0188—an unlikely event. Because a sample mean that high is unlikely to be obtained from such a population, we can reasonably conclude that this sample probably came from some other population (one whose mean is not 50)."

It is important to realize the steps in this example, because the logic is typical of most tests of hypotheses. The actual test consisted of several stages:

research hypothesis

1. We wanted to test the hypothesis, often called the **research hypothesis,** that children under stress are more likely than normal children to exhibit behavior problems.
2. We obtained a random sample of children under stress.

null hypothesis

3. We set up the hypothesis (called the **null hypothesis,** H_0) that the sample was in fact drawn from a population whose mean, denoted μ_0, equals 50. This is the hypothesis that stressed children do not differ from normal children in terms of behavior problems.
4. We then obtained the sampling distribution of the mean under the assumption that H_0 (the null hypothesis) is true (i.e., we obtained the sampling distribution of the mean from a population with $\mu_0 = 50$).
5. Given the sampling distribution, we calculated the probability of a mean *at least as large* as our actual sample mean.
6. On the basis of that probability, we made a decision: either to reject or fail to reject H_0. Because H_0 states that $\mu = 50$, rejection of H_0 represents a belief that $\mu > 50$, although the actual value of μ remains unspecified.

The preceding discussion is oversimplified in the sense that we generally would prefer to test the research hypothesis that children under stress are *different from* (rather than just *higher than*) other children, but we will return to that point shortly. It is also oversimplified in the sense that in practice we also would need to take into account (either directly or by estimation) the value of σ^2, the population variance, and N, the sample size. But again, those are specifics we can deal with when the time comes. The logic of the approach is representative of the logic of most, if not all, statistical tests:

1. Begin with a research hypothesis.
2. Set up the null hypothesis.
3. Construct the sampling distribution of the particular statistic on the assumption that H_0 is true.
4. Collect some data.
5. Compare the sample statistic to that distribution.
6. Reject or retain H_0, depending on the probability, under H_0, of a sample statistic as extreme as the one we have obtained.

4·4 The Null Hypothesis

As we have seen, the concept of the null hypothesis plays a crucial role in the testing of hypotheses. People frequently are puzzled by the fact that we set up a hypothesis that is directly counter to what we hope to show. For example, if we hope to demonstrate the research hypothesis that college students do not come from a population with a mean self-confidence score of 100, we immediately set up the null hypothesis that they do. Or if we hope to demonstrate the validity of a research hypothesis that the means (μ_1 and μ_2) of the population from which two samples are drawn are different, we state the null hypothesis that the population means are the same (or, equivalently $\mu_1 - \mu_2 = 0$). (The term "null hypothesis" is most easily seen in this second example, in which it refers to the hypothesis that the difference between the two population means is zero, or *null*.) We use the null hypothesis for several reasons. The philosophical argument, put forth by Fisher when he first introduced the concept, is that we can never prove something to be true, but we can prove something to be false. Observing 3000 people with two arms does not prove the statement "Everyone has two arms." However, finding one person with three arms does disprove the original statement beyond any shadow of a doubt. While one might argue with Fisher's basic position—and many people have—the null hypothesis retains its dominant place in statistics.

A second and more practical reason for employing the null hypothesis is that it provides us with the starting point for any statistical test. Consider the case in which you want to show that the mean self-confidence score of college students is greater than 100. Suppose further that you were granted the privilege of proving the truth of some hypothesis. What hypothesis are you going to test? Should you test the hypothesis that $\mu = 101$, or maybe the hypothesis that $\mu = 112$, or how about $\mu = 113$? The point is that you do not have a *specific* alternative (research) hypothesis in mind, and without one you cannot construct the sampling distribution you need. However, if you start off by assuming H_0: $\mu = 100$, you can immediately set about obtaining the sampling distribution for $\mu = 100$ and then, with luck, reject that hypothesis and conclude that the mean score of college students is greater than 100, which is what you wanted to show in the first place.

Statistical Conclusions

When the data differ markedly from what we would expect if the null hypothesis were true, we simply reject the null hypothesis and there is no particular disagreement about what our conclusions mean—we conclude that the null hypothesis is false. The interpretation is murkier and more problematic, however, when the data do not lead us to reject the null hypothesis. How are we to interpret a nonrejection? Shall we say that we have "proved" the null hypothesis to be true? Or shall we claim that we can "accept" the null, or that we shall "retain" it, or that we shall "withhold judgment"?

The problem of how to interpret a nonrejected null hypothesis has plagued students in statistics courses for over 50 years, and it will probably continue to do so. The idea that if something is not false then it must be true is too deeply ingrained in common sense to be dismissed lightly.

The one thing on which all statisticians agree is that we can never claim to have "proved" the null hypothesis. As was pointed out, the fact that the next 3000 people we meet all have two arms certainly does not prove the null hypothesis that all people have two arms. In fact, we know that many perfectly normal people have fewer than two arms. Failure to reject the null hypothesis often means that we have not collected enough data.

The issue is easier to understand if we use a concrete example. Wagner, Compas, and Howell (1988) conducted a study to evaluate the effectiveness of a program for teaching high-school students to deal with stress. If this study found that students who participate in such a program had significantly fewer stress-related problems than did students in a control group who did not have the program, then we could, without much debate, conclude that the program was effective. However, if the groups did not differ at some predetermined level of statistical significance, what could we conclude?

We know we cannot conclude from a nonsignificant difference that we have proved that the mean of a population of scores of treatment subjects is the same as the mean of a population of scores of control subjects. The two treatments may in fact lead to subtle differences that we were not able to identify conclusively with our relatively small sample of observations.

Fisher's position was that a nonsignificant result is an inconclusive result. For Fisher, the choice was between rejecting a null hypothesis and suspending judgment. He would have argued that a failure to find a significant difference between conditions could result from the fact that the students who participated in the program handled stress only *slightly* better than did control subjects, or that they handled it only slightly less well, or that there was no difference between the groups. For Fisher, a failure to reject H_0 merely meant that our data are insufficient to allow us to choose among these three alternatives; therefore, we must suspend judgment.

A slightly different approach was taken by Neyman and Pearson (1933), who took a much more pragmatic view of the results of an experiment. In our example, Neyman and Pearson would be concerned with the problem faced by the school board, who must decide whether to continue spending money on this stress-management program. The school board would probably not be impressed if we told them that our study was inconclusive and then asked them to give us money to continue operating the program until we had sufficient data to state confidently whether or not the program was beneficial (or harmful). In the Neyman–Pearson position, one either rejects or *accepts* the null hypothesis. When we say that we "accept" a null hypothesis, however, we do not mean that we take it to be proven as true. We simply mean that we will *act as if* it is true, at least until we have more adequate data. Whereas given a nonsignificant result, the ideal school board from Fisher's point of view would continue to support the program until we finally were able to make up our minds, the school board with a Neyman–Pearson perspective would conclude that the available evidence is that the program is not worth supporting and would cut off our funding.

This discussion of the Neyman–Pearson position has been much oversimplified, but it does contain the central issue of their point of view. The debate between Fisher on the one hand and Neyman and Pearson on the other was a lively (and rarely civil) one, and present practice contains elements of both viewpoints. Most statisticians prefer to use phrases such as "retain the null hypothesis" and "fail to reject the null

hypothesis" because these make clear the tentative nature of a nonrejection. These phrases have a certain Fisherian ring to them. On the other hand, the emphasis on Type II errors (discussed in Section 4.7) is clearly an essential feature of the Neyman–Pearson school. If you are going to choose between two alternatives (accept or reject), then you have to be concerned with the probability of falsely accepting as well as that of falsely rejecting the null hypothesis. Since Fisher would never accept a null hypothesis in the first place, he did not need to worry much about the probability of accepting a false one.[1]

One of the questions I often hear is "How could anyone be dumb enough to think that the null hypothesis is ever true?" That's a good point. If you measured precisely enough I'm sure that you'd find that people who live west of the Mississippi are taller (or shorter) on the average than people who live east of the Mississippi. I find it impossible to believe that your (population) means would be exactly the same, even to three or four decimal places. BUT, I still think that it would be meaningful to test the null hypothesis ($\mu_{West} = \mu_{East}$), not because I think it might be true, but because I want to know *which* side of the country has the higher mean. When you retain H_0 you are not saying that the two means are equal, but only that you don't have enough data to reliably tell which mean is larger. When you reject H_0 you are in a position to say that $\mu_{West} > \mu_{East}$ (or $\mu_{West} < \mu_{East}$). And that is *not* a trivial conclusion. Moreover, you are also in a position to make a statement about how large the difference actually is, and to put some sort of limits on your confidence about the size of that difference.

4·5 Test Statistics and Their Sampling Distributions

We have been discussing the sampling distribution of the mean, but the discussion would have been essentially the same had we dealt instead with the median, the variance, the range, the correlation coefficient (as in our course evaluation example), proportions (as in our horn-honking example), or any other statistic you care to consider. (Technically the shapes of these distributions would be different, but I am deliberately ignoring such issues in this chapter.) The statistics just mentioned usually are referred to as **sample statistics** because they describe samples. There is a whole different class of statistics called **test statistics,** which are associated with specific statistical procedures and which have their own sampling distributions. Test statistics are statistics such as t, F, and χ^2, which you may have run across in the past. If you are not familiar with them, don't worry—we will consider them separately in later chapters. This is not the place to go into a detailed explanation of any test statistics (I put this chapter where it is because I didn't want readers to think that they were supposed to worry about technical issues). This chapter is the place, however, to point out that

sample statistics
test statistics

[1]Excellent discussions of the differences between the theories of Fisher on the one hand, and Neyman and Pearson on the other can be found in Gigerenzer et al. (1989), Lehmann (1993), and Oakes (1990). The central issues involve the concept of probability, the idea of an infinite population or infinite resampling, and the choice of a critical value, among other things. The controversy is far from a simple one.

the sampling distributions for test statistics are obtained and used in essentially the same way as the sampling distribution of the mean.

As an illustration, consider the sampling distribution of the statistic t, which will be discussed in Chapter 7. For those who have never heard of the t test, it is sufficient to say that the t test is often used, among other things, to determine whether two samples were drawn from populations with the same means. Let μ_1 and μ_2 represent the means of the populations from which the two samples were drawn. The null hypothesis is the hypothesis that the two population means are equal, in other words, H_0: $\mu_1 = \mu_2$ (or $\mu_1 - \mu_2 = 0$). If we were extremely patient, we could empirically obtain the sampling distribution of t when H_0 is true by drawing an infinite number of pairs of samples, all from one population, calculating t for each pair of samples (by methods to be discussed later), and plotting the resulting values of t. In that case H_0 must be true because the samples came from the same population. The resulting distribution is the sampling distribution of t when H_0 is true. If we later had two samples that produced a particular value of t, we would test the null hypothesis by comparing our sample t to the sampling distribution of t. We would reject the null hypothesis if our obtained t did not look like the kinds of t values that the sampling distribution told us to expect when the null hypothesis is true.

I could rewrite the preceding paragraph, substituting χ^2, or F, or any other test statistic in place of t, with only minor changes dealing with how the statistic is calculated. Thus, you can see that all sampling distributions can be obtained in basically the same way (calculate and plot an infinite number of statistics by sampling from a known population). Once you understand that fact, much of the remainder of the book is an elaboration of methods for calculating the desired statistic and a description of characteristics of the appropriate sampling distribution.

4.6 Using the Normal Distribution to Test Hypotheses

Much of the discussion so far has dealt with statistical procedures that you do not yet know how to use. I did this deliberately to emphasize the point that the logic and the calculations behind a test are two separate issues. However, we now can use what you already know about the normal distribution to test some simple hypotheses. In the process we can deal with several fundamental issues that are more easily seen by use of a concrete example.

An important use of the normal distribution is to test hypotheses, either about individual observations or about sample statistics such as the mean. In this chapter we will deal with individual observations, leaving the question of testing sample statistics until later chapters. Note, however, that in the general case we test hypotheses about sample statistics such as the mean rather than about individual observations. I am starting with an example of an individual observation because the explanation is somewhat clearer. Since we are dealing with only single observations, the sampling distribution invoked here will be the distribution of individual scores (rather than the distribution of means). The basic logic is the same, and we are using an example of

individual scores only because it simplifies the explanation and is something with which you have had experience.

Psychologists who study neurological functioning have a battery of tests at their disposal. A common test is simple finger tapping speed, which is useful for diagnosing hidden brain damage. (For example, people with damage to the dorsal lateral frontal lobes are especially slow in terms of speed of finger tapping, but are often unaware of their loss of behavioral competency.) For a simple example, assume we know that the mean rate of finger tapping of normal healthy adults is 100 taps in 20 seconds, with a standard deviation of 20, and that tapping speeds are normally distributed in the population. We already know that the tapping rate is slower among people with dorsal lateral frontal lobe damage. Finally, suppose that an individual has just been sent to us who taps at a rate of 70 taps in 20 seconds. Is his score sufficiently below the mean for us to assume that he did not come from a population of neurologically healthy people? This situation is diagrammed in Figure 4.2, in which the arrow indicates the location of our piece of data (the person's score).

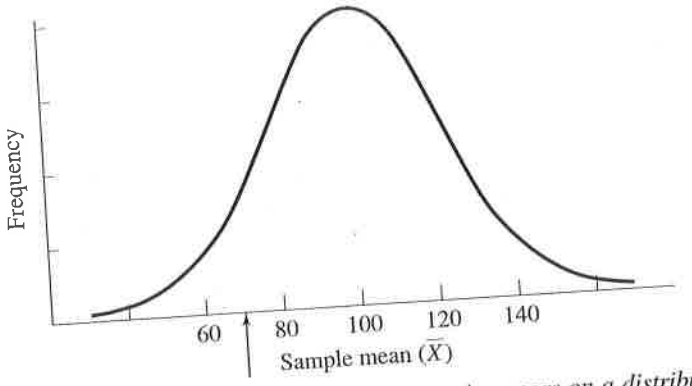

FIGURE 4.2 *Location of a person's tapping score on a distribution of scores of neurologically healthy people*

The logic of the solution to this problem is the same as the logic of hypothesis testing in general. We begin by assuming that the individual's score *does* come from the population of healthy scores. This is the null hypothesis (H_0). If H_0 is true, we automatically know the mean and the standard deviation of the population from which he was supposedly drawn (100 and 20, respectively). With this information we are in a position to calculate the probability that a score *as low as* his would be obtained from this population. If the probability is very low, we can reject H_0 and conclude that he did not come from the healthy population. Conversely, if the probability is not particularly low, then the data represent a reasonable result under H_0, and we would have no reason to doubt its validity and thus no reason to doubt that the person is healthy. Keep in mind that we are not interested in the probability of a score *equal* to 70 (which, because the distribution is continuous, would be infinitely small) but rather in the probability that the score would be at least as low as (i.e., less than or equal to) 70.

The individual had a score of 70. We want to know the probability of obtaining a score *at least as low as* 70 if H_0 is true. We already know how to find this—it is the area below 70 in Figure 4.2. All we have to do is convert 70 to a z score and then refer to Appendix z.

$$z = \frac{X - \mu}{\sigma} = \frac{70 - 100}{20} = \frac{-30}{20} = -1.5$$

From Appendix z, we can see that the probability of a z score of -1.5 or below is 0.0668. (Locate $z = 1.50$ in the table and then read across to the column headed "Smaller Portion.")

decision making

At this point we have to become involved in the **decision-making** aspects of hypothesis testing. We must decide whether an event with a probability of 0.0668 is sufficiently unlikely to cause us to reject H_0. Here we will fall back on arbitrary conventions that have been established over the years. The rationale for these conventions will become clearer as we go along, but for the time being keep in mind that they are merely conventions. One convention calls for rejecting H_0 if the probability under H_0 is less than or equal to 0.05 ($p \leq 0.05$), while another convention—one that is more conservative with respect to the probability of rejecting H_0—calls for rejecting H_0 whenever the probability under H_0 is less than or equal to 0.01. These values of 0.05 and 0.01 are often referred to as the **rejection level,** or **significance level,** of the test. Whenever the probability obtained under H_0 is less than or equal to our predetermined significance level, we will reject H_0. Another way of stating this is to say that any outcome whose probability under H_0 is less than or equal to the significance level falls in the **rejection region,** since such an outcome leads us to reject H_0.

rejection level
significance level

rejection region

For the purpose of setting a standard level of rejection for this book, we will use the 0.05 level of significance, keeping in mind that some people would consider this level to be too lenient.[2] For our particular example we have obtained a probability value of $p = 0.0668$, which obviously is greater than 0.05. Because we have specified that we will not reject H_0 unless the probability of the data under H_0 is less than 0.05, we must conclude that we have no reason to decide that the person did not come from a population of healthy people.

More specifically, we conclude that a finger-tapping rate of 70 reasonably could have come from a population of scores with a mean equal to 100 and a standard deviation equal to 20. It is important to note that we have not shown that this person is healthy, but only that we have insufficient reason to believe that he is not. It may be that he is just acquiring the disease and therefore is not quite as different from normal

[2]The particular view of hypothesis testing described here is the classical one that a null hypothesis is rejected if the probability of obtaining the data when the null hypothesis is true is less than the predefined significance level, and not rejected if that probability is greater than the significance level. Currently a substantial body of opinion holds that such cut-and-dried rules are inappropriate and that more attention should be paid to the probability value itself. In other words, the classical approach (using a 0.05 rejection level) would declare $p = 0.051$ and $p = 0.150$ to be (equally) "nonsignificant" and $p = 0.048$ and $p = 0.0003$ to be (equally) "significant." The alternative view would think of $p = 0.051$ as "nearly significant" and $p = 0.0003$ as "very significant." While this view has much to recommend it, it will not be wholeheartedly adopted here. Most computer programs do print out exact probability levels, and those values, when interpreted judiciously, can be useful. The difficulty comes in defining what is meant by "interpreted judiciously."

as is usual for his condition. Or maybe he has the disease at an advanced stage but just happens to be an unusually fast tapper. This is an example of the fact that we can never say that we have proved the null hypothesis. We can conclude only that this person does not tap sufficiently slowly for an illness, if any, to be statistically detectable.

As I mentioned earlier, the theory of significance testing as just outlined was popularized by R. A. Fisher in the first third of the 20th century. The theory was expanded and cast in more of a decision framework by Jerzy Neyman and Egon Pearson between 1928 and 1938, often against the loud and abusive objections of Fisher. Current statistical practice more closely follows the Neyman–Pearson approach, which emphasizes more than did Fisher the fact that we also have an **alternative hypothesis** (H_1) that is contradictory to the null hypothesis (H_0). Thus if the null hypothesis is

alternative hypothesis

$$H_0: \mu = 100$$

then the alternative hypothesis could be

$$H_1: \mu \neq 100$$

or

$$H_1: \mu < 100$$

or

$$H_1: \mu > 100$$

We will discuss alternative hypotheses in more detail shortly.

4.7 Type I and Type II Errors

Whenever we reach a decision with a statistical test, there is always a chance that our decision is the wrong one. While this is true of almost all decisions, statistical or otherwise, the statistician has one point in her favor that other decision makers normally lack. She not only makes a decision by some rational process, but she can also specify the conditional probabilities of a decision's being in error. In everyday life we make decisions with only subjective feelings about what is probably the right choice. (I had one excellent student who went around truly believing that whatever his choice, it was probably the wrong one.) The statistician, however, can state quite precisely the probability that she erroneously rejected H_0 in favor of the alternative (H_1). This ability to specify the probability of error follows directly from the logic of hypothesis testing.

Consider the finger-tapping example, this time ignoring the score of the individual sent to us. The situation is diagrammed in Figure 4.3, in which the distribution is the distribution of scores from healthy subjects, and the shaded portion represents the lowest 5% of the distribution. The actual score that cuts off the lowest 5% is called the **critical value.** Critical values are those values of X (the variable) that describe the boundary or boundaries of the rejection region(s). For this particular example the critical value is 67.

critical value

If we have a decision rule that says to reject H_0 whenever an outcome falls in the lowest 5% of the distribution, we will reject H_0 whenever an individual's score falls in the shaded area—that is, whenever a score as low as his has a probability of 0.05 or less of coming from the population of healthy scores. Yet by the very nature of our

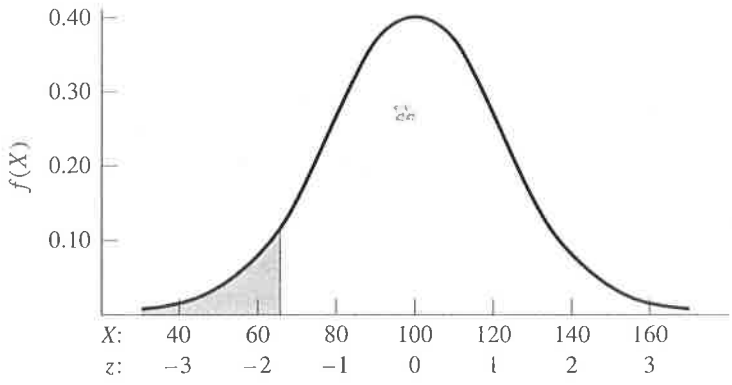

FIGURE 4.3 *Lowest 5% of scores from clinically healthy people*

procedure, 5% of the scores from perfectly healthy people will themselves fall in the shaded portion. Thus if we actually have sampled a person who is healthy, we stand a 5% chance of his score being in the shaded tail of the distribution, causing us erroneously to reject the null hypothesis. This kind of error (rejecting H_0 when in fact it is true) is called a **Type I error,** and its conditional probability (the probability of rejecting the null hypothesis given that it is true) is designated as α **(alpha),** the size of the rejection region. In the future, whenever we represent a probability by α, we will be referring to the probability of a Type I error.

Type I error
α (alpha)

Keep in mind the "conditional" nature of the probability of a Type I error. I know that sounds like jargon, but what it means is that you should be sure you understand that when we speak of a Type I error we mean the probability of rejecting H_0 *given that it is true.* We are not saying that we will reject H_0 on 5% of the hypotheses we test. We would hope to run experiments on important and meaningful variables and, therefore, to reject H_0 often. But when we speak of a Type I error, we are speaking only about rejecting H_0 in those situations in which the null hypothesis happens to be true.

You might feel that a 5% chance of making an error is too great a risk to take and suggest that we make our criterion much more stringent, by rejecting, for example, only the lowest 1% of the distribution. This procedure is perfectly legitimate, but realize that the more stringent you make your criterion, the more likely you are to make another kind of error—failing to reject H_0 when it is in fact false and H_1 is true. This type of error is called a **Type II error,** and its probability is symbolized by β **(beta).**

Type II error
β (beta)

The major difficulty in terms of Type II errors stems from the fact that if H_0 is false, we almost never know what the true distribution (the distribution under H_1) would look like for the population from which our data came. We know only the distribution of scores under H_0. Put in the present context, we know the distribution of scores from healthy people but not from nonhealthy people. It may be that people suffering from some neurological disease tap, on average, considerably more slowly than healthy people, or it may be that they tap, on average, only a little more slowly. This situation is illustrated in Figure 4.4 on page 106, in which the distribution labeled H_0 represents the distribution of scores from healthy people (the set of observations expected under the null hypothesis), and the distribution labeled H_1 represents our hypothetical

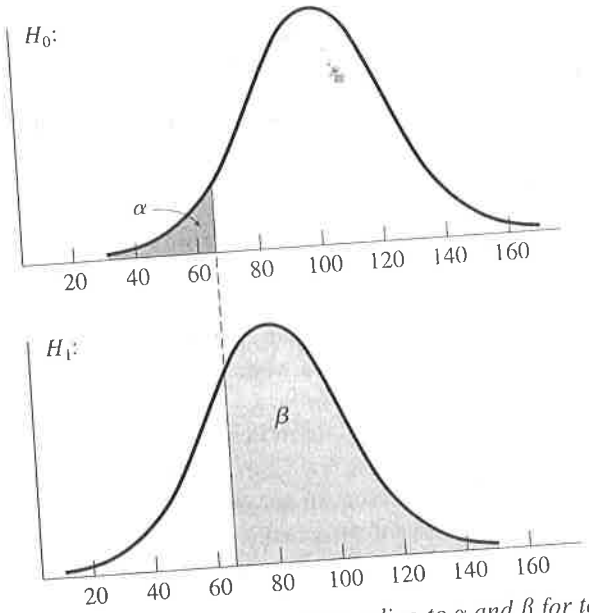

FIGURE 4·4 *Areas corresponding to α and β for tapping speed example*

distribution of nonhealthy scores (the distribution under H_1). Remember that the curve H_1 is only hypothetical. We really do not know the location of the nonhealthy distribution, other than that it is lower (slower speeds) than the distribution of H_0. (I have arbitrarily drawn that distribution with a mean of 80 and a standard deviation of 20.)

The darkly shaded portion in the top half of Figure 4.4 represents the rejection region. Any observation falling in that area (i.e., to the left of about 67) would lead to rejection of the null hypothesis. If the null hypothesis is true, we know that our observation will fall in this area 5% of the time. Thus, we will make a Type I error 5% of the time.

The lightly shaded portion in the bottom half of Figure 4.4 represents the probability (β) of a Type II error. This is the situation of a person who was actually drawn from the nonhealthy population but whose score was not sufficiently low to cause us to reject H_0.

In the particular situation illustrated in Figure 4.4, we can in fact calculate β by using the normal distribution to calculate the probability of obtaining a score *greater than* 67 (the critical value) if $\mu = 80$ and $\sigma = 20$. The actual calculation is not important for your understanding of β; because this chapter was designed specifically to avoid calculation, I will simply state that this probability (i.e., the area labeled β) is 0.74. Thus for this example, 74% of the time when we have a person who is actually nonhealthy (i.e., H_1 is actually true), we will make a Type II error by failing to reject H_0 when it is false (as medical diagnosticians, we leave a lot to be desired).

From Figure 4.4 you can see that if we were to reduce the level of α (the probability of a Type I error) from 0.05 to 0.01 by moving the rejection region to the left, it would reduce the probability of Type I errors but would increase the probability of Type II errors. Setting α at 0.01 would mean that $\beta = 0.908$. Obviously there is room for debate over what level of significance to use. The decision rests primarily on your

opinion concerning the relative importance of Type I and Type II errors for the kind of study you are conducting. If it were important to avoid Type I errors (such as telling someone that he has a disease when he does not), then you would set a stringent (i.e., small) level of α. If, on the other hand, you want to avoid Type II errors (telling someone to go home and take an aspirin when in fact he needs immediate treatment), you might set a fairly high level of α. (Setting $\alpha = 0.20$ in this example would reduce β to 0.44.) Unfortunately, in practice most people choose an arbitrary level of α, such as 0.05 or 0.01, and simply ignore β. In many cases this may be all you can do. (In fact you will probably use the alpha level that your instructor recommends.) In other cases, however, there is much more you can do, as you will see in Chapter 8.

I should stress again that Figure 4.4 is purely hypothetical. I was able to draw the figure only because I arbitrarily decided that speeds of nonhealthy people were normally distributed with a mean of 80 and a standard deviation of 20. The calculated answers would be different if I had chosen to draw it with a mean of 70 and/or a standard deviation of 10. In most everyday situations we do not know the mean and the variance of that distribution and can only make educated guesses, thus providing only crude estimates of β. In practice we can select a value of μ under H_1 that represents the *minimum* difference we would like to be able to detect, since larger differences will have even smaller βs.

From this discussion of Type I and Type II errors we can summarize the decision-making process with a simple table. Table 4.1 presents the four possible outcomes of an experiment. The items in this table should be self-explanatory, but there is one concept—power—that we have not yet discussed. The **power** of a test is the probability of rejecting H_0 when it is actually false. Because the probability of *failing* to reject a false H_0 is β, then power must equal $1 - \beta$. Those who want to know more about power and its calculation will find the material in Chapter 8 relevant.

power

TABLE 4.1 Possible Outcomes of the Decision-Making Process

	True State of the World	
Decision	**H_0 True**	**H_0 False**
Reject H_0	Type I error $p = \alpha$	Correct decision $p = 1 - \beta$ = Power
Don't reject H_0	Correct decision $p = 1 - \alpha$	Type II error $p = \beta$

4.8 One- and Two-Tailed Tests

The preceding discussion brings us to a consideration of one- and two-tailed tests. In our tapping example we knew that nonhealthy subjects tapped more slowly than healthy subjects; therefore, we decided to reject H_0 only if a subject tapped too slowly. However, suppose our subject had tapped 180 times in 20 seconds. Although this is an exceedingly unlikely event to observe from a healthy subject, it did not fall in the rejection region, which consisted *solely* of low rates. As a result we find

ourselves in the position of not rejecting H_0 in the face of a piece of data that is very unlikely, but not in the direction expected.

The question then arises as to how we can protect ourselves against this type of situation (if protection is thought necessary). The answer is to specify before we run the experiment that we are going to reject a given percentage (say 5%) of the *extreme* outcomes, both those that are extremely high and those that are extremely low. But if we reject the lowest 5% and the highest 5%, then we would in fact reject H_0 a total of 10% of the time when it is actually true, that is, $\alpha = 0.10$. We are rarely willing to work with α as high as 0.10 and prefer to see it set no higher than 0.05. The way to accomplish this is to reject the lowest 2.5% and the highest 2.5%, making a total of 5%.

one-tailed test (directional test)

The situation in which we reject H_0 for only the lowest (or only the highest) tapping speeds is referred to as a **one-tailed, or directional, test.** We make a prediction of the direction in which the individual will differ from the mean and our rejection region is located in only one tail of the distribution. (That makes sense when we know that brain damage is only associated with slow tapping speeds.) When we reject extremes in both tails, we have what is called a **two-tailed, or nondirectional, test.** It is important to keep in mind that while we gain something with a two-tailed test (the ability to reject the null hypothesis for extreme scores in either direction), we also lose something. A score that would fall in the 5% rejection region of a one-tailed test may not fall in the rejection region of the corresponding two-tailed test, because now we reject only 2.5% in each tail.

two-tailed test (nondirectional test)

In the finger-tapping example, the decision between a one- and a two-tailed test might seem reasonably clear-cut. We know that people with a given disease tap more slowly; therefore we care only about rejecting H_0 for low scores—high scores have no diagnostic importance. In other situations, however, we do not know which tail of the distribution is important (or both are), and we need to guard against extremes in either tail. The situation might arise when we are considering a campaign to persuade children not to start smoking. We might find that the campaign leads to a decrease in the incidence of smoking. Or, we might find that campaigns run by adults to persuade children not to smoke simply make smoking more attractive and exciting, leading to an increase is the number of children smoking. In either case we would want to reject H_0.

In general, two-tailed tests are far more common than one-tailed tests for several reasons. First, the investigator may have no idea what the data will look like and therefore has to be prepared for any eventuality. Although this situation is rare, it does occur in some exploratory work.

Another common reason for preferring two-tailed tests is that the investigators are reasonably sure the data will come out one way but want to cover themselves in the event that they are wrong. This type of situation arises more often than you might think. (Carefully formed hypotheses have an annoying habit of being phrased in the wrong direction, for reasons that seem so obvious after the event.) The smoking example is a case in point, where there is some evidence that poorly contrived anti-smoking campaigns actually do more harm than good. A frequent question that arises when the data may come out the other way around is, "Why not plan to run a one-tailed test and then, if the data come out the other way, just change the test to a two-tailed test?" This kind of question comes from people who have no intention of being devious but who just do not fully understand the logic of hypothesis testing. If you start an experiment with the extreme 5% of the left-hand tail as your rejection re-

gion and then turn around and reject any outcome that happens to fall in the extreme 2.5% of the right-hand tail, you are working at the 7.5% level. In that situation you will reject 5% of the outcomes in one direction (assuming that the data fall in the desired tail), and you are willing also to reject 2.5% of the outcomes in the other direction (when the data are in the unexpected direction). There is no denying that 5% + 2.5% = 7.5%. To put it another way, would you be willing to flip a coin for an ice cream cone if I have chosen "heads" but also reserve the right to switch to "tails" after I see how the coin lands? Or would you think it fair of me to shout, "Two out of three!" when the coin toss comes up in your favor? You would object to both of these strategies, and you should. For the same reason, the choice between a one-tailed test and a two-tailed one is made *before* the data are collected. It is also one of the reasons that two-tailed tests are usually chosen.

Although the preceding discussion argues in favor of two-tailed tests, and although in this book we generally confine ourselves to such procedures, there are no hard-and-fast rules. The final decision depends on what you already know about the relative severity of different kinds of errors. It is important to keep in mind that with respect to a given tail of a distribution, the difference between a one-tailed test and a two-tailed test is that the latter just uses a different cutoff. A two-tailed test at $\alpha = 0.05$ is more liberal than a one-tailed test at $\alpha = 0.01$.[3]

If you have a sound grasp of the logic of testing hypotheses by use of sampling distributions, the remainder of this course will be relatively simple. For any new statistic you encounter, you will need to ask only two basic questions:

1. How and with which assumptions is the statistic calculated?
2. What does the statistic's sampling distribution look like under H_0?

If you know the answers to these two questions, your test is accomplished by calculating the test statistic for the data at hand and comparing the statistic to the sampling distribution. Because the relevant sampling distributions are tabled in the appendices, all you really need to know is which test is appropriate for a particular situation and how to calculate its test statistic. (Keep in mind, however, there is a great deal more to understanding the field of statistics than how to calculate, and evaluate, a specific statistical test. Calculation is the easy part, especially with modern computer software.)

---

[3]One of the reviewers of an earlier edition of this book made the case for two-tailed tests even more strongly: "It is my (minority) belief that what an investigator *expects to be true* has absolutely no bearing *whatsoever* on the issue of one- versus two-tailed tests. Nature couldn't care less what psychologists' theories predict, and will often show patterns/trends in the opposite direction. Since our goal is to know the truth (not to prove we are astute at predicting), our tests must always allow for testing *both* directions. I say *always* do two-tailed tests, and if you are worried about β, jack the sample size up a bit to offset the loss in power" (D. Bradley, personal communication, 1983). I am personally inclined toward this point of view. Nature is notoriously fickle, or else we are notoriously inept at prediction. On the other hand, a second reviewer (J. Rodgers, personal communication, 1986) takes exception to this position. While acknowledging that Bradley's point is well considered, Rodgers, engaging in a bit of hyperbole, argues, "To generate a theory about how the world works that implies an expected direction of an effect, but then to hedge one's bet by putting some (up to ½) of the rejection region in the tail other than that predicted by the theory, strikes me as both scientifically dumb and slightly unethical. . . . Theory generation and theory testing are much closer to the proper goal of science than truth searching, and running one-tailed tests is quite consistent with those goals." Neither Bradley nor I would accept the judgment of being "scientifically dumb and slightly unethical," but I presented the two positions in juxtaposition because doing so gives you a flavor of the debate. Obviously there is room for disagreement on this issue.

4.9 What Does It Mean to Reject the Null Hypothesis?

One of the common problems that even well-trained researchers have with the null hypothesis is the confusion over what rejection really means. Suppose that we test a null hypothesis about the difference between two population means and reject it at $p = 0.045$. There is a temptation to say that such a result means that the probability of the null being true is 0.045. But that is *not* what this probability means. What we have shown is that *if the null hypothesis were true,* the probability of obtaining a difference between means as great as the difference we found is only 0.045. That is quite different from saying that the probability that the null is true is 0.045. What we are doing here is confusing the probability of the hypothesis given the data, and the probability of the data given the hypothesis. These are called **conditional probabilities,** and will be discussed in Chapter 5. The probability of 0.045 that we have here is the probability of the data given that H_0 is true [written $p(\text{D} \mid H_0)$]. It is not the probability of H_0 true given the data [written $p(H_0 \mid \text{D})$]. The best discussion of this issue that I have read is a paper by Nickerson (2000).

As an example of what is going on here, suppose that I create a computer-generated example where I know for a fact that the data for one sample came from a population with a mean of 54.28, and the data for a second sample came from a population with a mean of 54.25. Here I *know for a fact* that the null hypothesis is false. In other words, $p(H_0 \text{ is true}) = 0.00$. However, if I have two small samples I might happen to get results such as 54.26 and 54.36 that would have a very high probability of occurring if the null were true and both means were, say, 54.25. Thus the probability of the data given a true null might be 0.75, and yet we know that the probability that the null is really true is exactly 0.00. Alternatively, assume that I created a situation where I know that the null is true. For example, I set up populations where both means are 54.00. It is easy to imagine getting samples with means of 53 and 54.5. If the null is really true, the probability of getting means this different may be 0.33. Thus the probability that the null is true is fixed, by me, at 1.00, yet the probability of the data when the null is true is 0.33. Notice that in both of these cases there is a serious discrepancy between the probability of the null being true and the probability of the data given the null. You will see several instances like this throughout the book whenever I sample data from known populations. Never confuse the probability value associated with a test of significance with the probability that the null hypothesis is true. They are very different things.

4.10 Effect Size

Near the beginning of the chapter I mentioned that there was a movement afoot to go beyond simple significance testing to report some measure of the size of an effect. I will expand on this topic in some detail later, but it is worth noting here that I have already sneaked a measure of effect size past you, and I'll bet that nobody noticed. I

wrote "These five children have a mean of 56 on the Achenbach Youth Self-Report scale of Total Behavior Problems. This mean is *over half a standard deviation* (italics added) above the mean (50) in the general population, and we want to know whether this finding is sufficiently deviant for us to conclude that the stress associated with divorce tends to elicit behavior problems in children at higher than normal levels." The fact that I expressed the difference between these children and the mean of normal children as being "over half a standard deviation" is itself a measure of how large the effect was—and a very reputable one. We are saying that for that group of children the effect of family divorce was to increase their mean by over half a standard deviation. There is much more to be said about effect sizes, but at least this gives you some idea of what we are talking about.

4.11 A Final Worked Example

A number of years ago the mean on the verbal section of the Graduate Record Exam (GRE) was 489 with a standard deviation of 126. The statistics were based on all students taking the exam in that year, the vast majority of whom were native speakers of English. Suppose we have an application from an individual with a Chinese name who scored particularly low (e.g., 220). If this individual is a native speaker of English, that score would be sufficiently low for us to question his suitability for graduate school unless the rest of the documentation is considerably better. If, however, this student is not a native speaker of English, we would probably disregard the low score entirely, on the grounds that it is a poor reflection of his abilities.

We have two possible choices here—namely, that the individual is or is not a native speaker of English. If he is a native speaker, we know the mean and the standard deviation of the population from which his score was sampled: 489 and 126, respectively. If he is not a native speaker, we have no idea what the mean and the standard deviation are for the population from which his score was sampled. To help us to draw a reasonable conclusion about this person's status, we will set up the null hypothesis that this individual is a native speaker, or, more precisely, he was drawn from a population with a mean of 489, H_0: $\mu = 489$. We will identify H_1 with the hypothesis that the individual is not a native speaker ($\mu \neq 489$).

We now need to choose between a one-tailed and a two-tailed test. In this particular case we will choose a one-tailed test on the grounds that the GRE is given in English, and it is difficult to imagine that a population of nonnative speakers would have a mean higher than the mean of native speakers of English on a test that is given in English. (*Note:* This does not mean that non-English speakers may not, singly or as a population, outscore English speakers on a fairly administered test. It just means that they are unlikely to do so, especially as a group, when both groups take the test in English.) Because we have chosen a one-tailed test, we have set up the alternative hypothesis as H_1: $\mu < 489$.

Before we can apply our statistical procedures to the data at hand, we must make one additional decision. We have to decide on a level of significance for our test. In this case I have chosen to run the test at the 5% level, instead of at the 1% level,