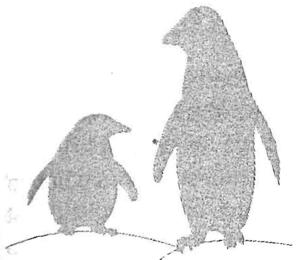

4

Introduction to Hypothesis Testing

Chapter Outline

hypothesis testing

Tip for Success

Before beginning this chapter, be sure you have mastered Chapters 1, 2, and 3.

In this chapter, we introduce the crucial topic of **hypothesis testing.** Hypothesis testing is a systematic procedure for deciding whether the results of a research study, which examines a sample, support a particular theory or practical innovation, which applies to a population. Hypothesis testing is the central theme in all the remaining chapters of this book, as it is in most psychology research.

Many students find the most difficult part of the course to be mastering the basic logic of this chapter and the next two. This chapter in particular requires some mental gymnastics. Even if you follow everything the first time through, you will be wise to review it thoroughly. Hypothesis testing involves grasping ideas that make little sense covered separately, so in this chapter you learn several new ideas all at once. However, once you understand the material in this chapter and the two that follow, your mind will be used to this sort of thing, and the rest of the course should seem easier.

At the same time, we have kept this introduction to hypothesis testing as simple as possible, putting off what we could for later chapters. For example, real-life psychology research involves samples of many individuals. However, to simplify how much you have to learn at one time, this chapter's examples are about studies in which the sample is a single individual. To do this, we use some odd examples. Just remember that you are building a foundation that will, by Chapter 7, prepare you to understand hypothesis testing as it is actually carried out.

A HYPOTHESIS-TESTING EXAMPLE

Here is our first necessarily odd example that we made up to keep this introduction to hypothesis testing as straightforward as possible. A large research project has been going on for several years. In this project, new babies are given a particular vitamin and then the research team follows their development during the first 2 years of life. So far, the vitamin has not speeded up the development of the babies. The ages at which these and all other babies start to walk is shown in Figure 4–1. The mean is 14 months ($\mu = 14$), the standard deviation is 3 months ($\sigma = 3$), and the ages follow a normal curve. Based on the normal curve percentages, you can figure that less than 2% of babies start walking before 8 months of age; these are the babies who are 2 standard deviations below the mean. (This fictional distribution is close to the true distribution psychologists have found for European babies, although that true distribution is slightly skewed to the right [Hindley et al., 1966].)

One of the researchers working on the project has an idea. If the vitamin the babies are taking could be more highly refined, perhaps the effect of the vitamin would be dramatically greater: Babies taking the highly purified version should start walking much earlier than other babies. (We will assume that the purification process could not possibly make the vitamin harmful.) However, refining the vitamin in this way is extremely expensive for each dose, so the research team decides to try the procedure with just enough purified doses for one baby. A newborn in the project is then randomly selected to take the highly purified version of the vitamin, and the researchers then follow this baby's progress for 2 years. What kind of result should lead the researchers to conclude that the highly purified vitamin allows babies to walk earlier?

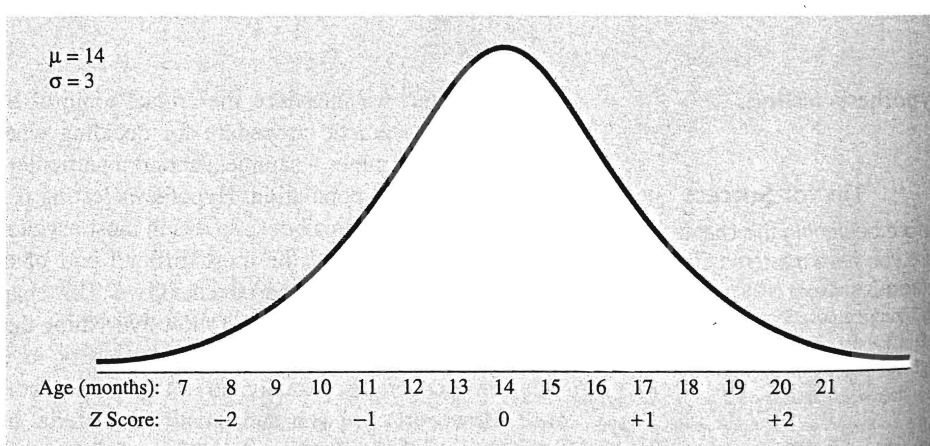

FIGURE 4–1 *Distribution of when babies begin to walk (fictional data).*

This is a hypothesis-testing problem. The researchers want to draw a general conclusion about whether the purified vitamin allows babies in general to walk earlier. The conclusion about the babies in general (a population of babies), however, will be based on results of studying only a sample. In this example, the sample is of a single baby.

THE CORE LOGIC OF HYPOTHESIS TESTING

There is a standard kind of reasoning researchers use for any hypothesis-testing problem. For this example, it works as follows. Ordinarily, among the population of babies that are not given the specially purified vitamin, the chance of a baby's starting to walk at age 8 months or earlier would be less than 2%. Thus, walking at 8 months or earlier is highly unlikely among such babies. But what if the randomly selected sample of one baby in our study does start walking by 8 months? If the specially purified vitamin had no effect on this particular baby's walking age (which means that the baby's walking age should be similar to that of babies that were not given the vitamin), it is highly unlikely (less than a 2% chance) that the particular baby we selected at random would start walking by 8 months. So, if the baby in our study does in fact start walking by 8 months, that allows us to *reject* the idea that the specially purified vitamin has *no* effect. And if we reject the idea that the specially purified vitamin has no effect, then we must also *accept* the idea that the specially purified vitamin *does* have an effect. Using the same reasoning, if the baby starts walking by 8 months, we can reject the idea that this baby comes from a population of babies with a mean walking age of 14 months. We therefore conclude that babies given the specially purified vitamin will start to walk before 14 months. Our explanation for the baby's early walking age in the study is that the specially purified vitamin speeded up the baby's development.

The researchers first spelled out what would have to happen for them to conclude that the special purification procedure makes a difference. Having laid this out in advance, the researchers could then go on to carry out their study. In this example, carrying out the study means giving the specially purified vitamin to a randomly selected baby and watching to see how early that baby walks. Suppose the result of the study is that the baby starts walking before 8 months. The researchers would then conclude that it is unlikely the specially purified vitamin makes *no* difference and thus also conclude that it *does* make a difference.

This kind of testing the opposite-of-what-you-predict, roundabout reasoning, is at the heart of inferential statistics in psychology. It is something like a double negative. One reason for this approach is that we have the information to figure the probability of getting a particular experimental result if the situation of there being *no difference* is true. In the purified vitamin example, the researchers know what the probabilities are of babies walking at different ages if the specially purified vitamin does not have any effect. It is the probability of babies walking at various ages that is already known from studies of babies in general—that is, babies who have not received the specially purified vitamin. (Suppose the specially purified vitamin has no effect. In that situation, the age at which babies start walking is the same whether or not they receive the specially purified vitamin. Thus, the distribution is that shown in Figure 4–1, based on ages at which babies start walking in general.)

Without such a tortuous way of going at the problem, in most cases you could just not do hypothesis testing at all. In almost all psychology research, we base our conclusions on this question: What is the probability of getting our research results

Tip for Success

This section, The Core Logic of Hypothesis Testing, is central to everything else we do in the book. Thus, you may want to read it a few times. You should also be certain that you understand the logic of hypothesis testing before reading later chapters.

if the opposite of what we are predicting were true? That is, we are usually predicting an effect of some kind. However, we decide on whether there *is* such an effect by seeing if it is unlikely that there is *not* such an effect. If it is highly unlikely that we would get our research results if the opposite of what we are predicting were true, that allows us to reject that opposite prediction. If we reject that opposite prediction, we are able to accept our prediction. However, if it is likely that we would get our research results if the opposite of what we are predicting were true, we are not able to reject that opposite prediction. If we are not able to reject that opposite prediction, we are not able to accept our prediction.

THE HYPOTHESIS-TESTING PROCESS

Let's look at our example again, this time going over each step in some detail. Along the way, we cover the special terminology of hypothesis-testing. Most important, we introduce five steps of hypothesis testing you use for the rest of this book.

STEP 1: RESTATE THE QUESTION AS A RESEARCH HYPOTHESIS AND A NULL HYPOTHESIS ABOUT THE POPULATIONS

Our researchers are interested in the effects on babies in general (not just this particular baby). That is, the purpose of studying samples is to know about populations. Thus, it is useful to restate the research question in terms of populations. In our example, we can think of two populations of babies:

Population 1: Babies who take the specially purified vitamin.
Population 2: Babies who do not take the specially purified vitamin.

Population 1 comprise those babies who receive the experimental treatment. In our example, we use a sample of one baby to draw a conclusion about the age that babies in Population 1 start to walk. Population 2 is a kind of comparison baseline of what is already known.

The prediction of our research team is that Population 1 babies (those who take the specially purified vitamin) will on the average walk earlier than Population 2 babies (those who do not take the specially purified vitamin). This prediction is based on the researchers' theory of how these vitamins work. A prediction like this about the difference between populations is called a **research hypothesis.** Put more formally, the prediction is that the mean of Population 1 is lower (babies receiving the special vitamin walk earlier) than the mean of Population 2. In symbols, the research hypothesis for this example is $\mu_1 < \mu_2$.

research hypothesis

The opposite of the research hypothesis is that the populations are not different in the way predicted. Under this scenario, Population 1 babies (those who take the specially purified vitamin) will on the average *not* walk earlier than Population 2 babies (those who do not take the specially purified vitamin). That is, this prediction is that there is no difference in when Population 1 and Population 2 babies start walking. They start at the same time. A statement like this, about a lack of difference between populations, is the crucial *opposite* of the research hypothesis. It is called a **null hypothesis.** It has this name because it states the situation in which

null hypothesis

there is no difference (the difference is "null") between the populations. In symbols, the null hypothesis is $\mu_1 = \mu_2$.[1]

The research hypothesis and the null hypothesis are complete opposites: If one is true, the other cannot be. In fact, the research hypothesis is sometimes called the *alternative hypothesis*—that is, it is the alternative to the null hypothesis. This is a bit ironic. As researchers, we care most about the research hypothesis. But when doing the steps of hypothesis testing, we use this roundabout method of seeing whether or not we can reject the null hypothesis so that we can decide about its alternative (the research hypothesis).

STEP 2: DETERMINE THE CHARACTERISTICS OF THE COMPARISON DISTRIBUTION

Recall that the overall logic of hypothesis testing involves figuring out the probability of getting a particular result if the null hypothesis is true. Thus, you need to know what the situation would be if the null hypothesis were true. In our example, we start out knowing the key information about Population 2 (see Figure 4–1)—we know $\mu = 14$, $\sigma = 3$, and it is normally distributed. If the null hypothesis is true, Population 1 and Population 2 are the same—in our example, this would mean Populations 1 and 2 both follow a normal curve, $\mu = 14$, and $\sigma = 3$.

In the hypothesis-testing process, you want to find out the probability that you could have gotten a sample score as extreme as what you got (say, a baby walking very early) if your sample were from a population with a distribution of the sort you would have if the null hypothesis were true. Thus, in this book we call this distribution a **comparison distribution.** (The comparison distribution is sometimes called a *statistical model* or a *sampling distribution*—an idea we discuss in Chapter 5.) That is, in the hypothesis-testing process, you compare the actual sample's score to this comparison distribution.

comparison distribution

In our vitamin example, the null hypothesis is that there is no difference in walking age between babies that take the specially purified vitamin (Population 1) and babies that do not take the specially purified vitamin (Population 2). The comparison distribution is the distribution for Population 2, since this population represents the walking age of babies if the null hypothesis is true. In later chapters, you will learn about different types of comparison distributions, but the same principle applies in all cases: The comparison distribution is the distribution that represents the population situation if the null hypothesis is true.

STEP 3: DETERMINE THE CUTOFF SAMPLE SCORE ON THE COMPARISON DISTRIBUTION AT WHICH THE NULL HYPOTHESIS SHOULD BE REJECTED

Ideally, before conducting a study, researchers set a target against which they will compare their result—how extreme a sample score they would need to decide against the null hypothesis: that is, how extreme the sample score would have to be

[1]We are oversimplifying a bit to make the initial learning easier. The research hypothesis is that one population will walk earlier than the other, $\mu_1 < \mu_2$. Thus, to be precise, its opposite is that the other group will either walk at the same time or later. That is, the opposite of the research hypothesis in this example includes both no difference and a difference in the direction opposite to what we predicted. In terms of symbols, if our research hypothesis is $\mu_1 < \mu_2$, then its opposite is $\mu_1 \geq \mu_2$ (the symbol $\geq$ means "greater than or equal to"). We discuss this issue in some detail later in the chapter.

cutoff sample score

for it to be too unlikely that they could get such an extreme score if the null hypothesis were true. This is called the **cutoff sample score.** (The cutoff sample score is also known as the *critical value.*)

Consider our purified vitamin example, in which the null hypothesis is that walking age is not influenced by whether babies take the specially purified vitamin. The researchers might decide that if the null hypothesis were true, a randomly selected baby walking before 8 months would be very unlikely. With a normal distribution, being 2 or more standard deviations below the mean (walking by 8 months) could occur less than 2% of the time. Thus, based on the comparison distribution, the researchers set their cutoff sample score even before doing the study. They decide in advance that *if* the result of their study is a baby who walks by 8 months, they will reject the null hypothesis.

But, what if the baby does not start walking until after 8 months? If that happens, the researchers will not be able to reject the null hypothesis.

When setting in advance how extreme a sample's score needs to be to reject the null hypothesis, researchers use Z scores and percentages. In our purified vitamin example, the researchers might decide that if a result were less likely than 2%, they would reject the null hypothesis. Being in the bottom 2% of a normal curve means having a Z score of about -2 or lower. Thus, the researchers would set -2 as their Z-score cutoff point on the comparison distribution for deciding that a result is extreme enough to reject the null hypothesis. So, if the actual sample Z score is -2 or lower, the researchers will reject the null hypothesis. However, if the actual sample Z score is greater than -2, the researchers will not reject the null hypothesis.

Suppose that the researchers are even more cautious about too easily rejecting the null hypothesis. They might decide that they will reject the null hypothesis only if they get a result that could occur by chance 1% of the time or less. They could then figure out the Z-score cutoff for 1%. Using the normal curve table, to have a score in the lower 1% of a normal curve, you need a Z score of -2.33 or less. (In our example, a Z score of -2.33 means 7 months.) In Figure 4–2, we have shaded the 1% of the comparison distribution in which a sample would be considered so extreme that the possibility that it came from a distribution like this would be rejected. So, now the researchers will only reject the null hypothesis if the actual

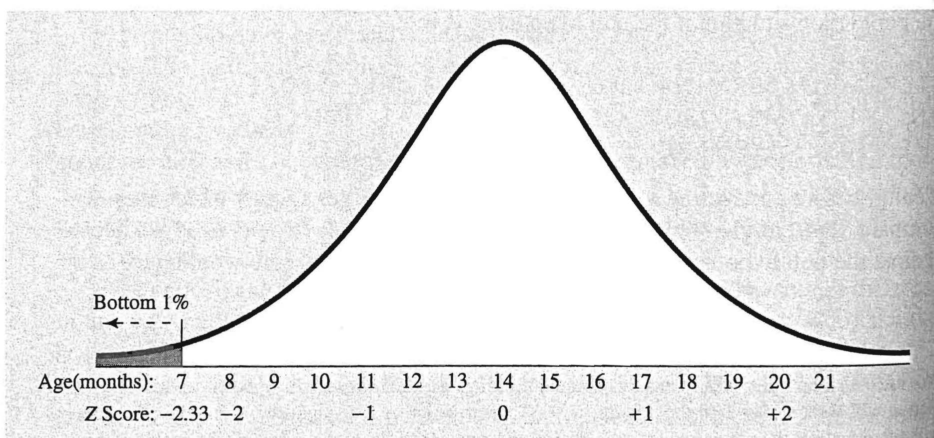

FIGURE 4–2 *Distribution of when babies begin to walk, with bottom 1% shaded (fictional data).*

sample Z score is -2.33 or lower—that is, if it falls in the shaded area in Figure 4–2. If the sample Z score falls outside of the shaded area in Figure 4–2, the researchers will *not* reject the null hypothesis.

In general, psychology researchers use a cutoff on the comparison distribution with a probability of 5% that a score will be at least that extreme if the null hypothesis were true. That is, researchers reject the null hypothesis if the probability of getting a sample score this extreme (if the null hypothesis were true) is less than 5%. This probability is usually written as $p < .05$. However, in some areas of research, or when researchers want to be especially cautious, they use a cutoff of 1% ($p < .01$).[2] These are called **conventional levels of significance.** They are described as the *.05 significance level* and the *.01 significance level.* We also refer to them as the 5% significance level and the 1% significance level. (We discuss in more detail in Chapter 6 the issues in deciding on the significance level to use.) When a sample score is so extreme that researchers reject the null hypothesis, the result is said to be **statistically significant** (or *significant,* as it is often abbreviated).

conventional levels of significance

statistically significant

STEP 4: DETERMINE YOUR SAMPLE'S SCORE ON THE COMPARISON DISTRIBUTION

The next step is to carry out the study and get the actual result for your sample. Once you have the results for your sample, you figure the Z score for the sample's raw score based on the population mean and standard deviation of the comparison distribution.

Assume that the researchers did the study and the baby who was given the specially purified vitamin started walking at 6 months. The mean of the comparison distribution to which we are comparing these results is 14 months and the standard deviation is 3 months. That is, $\mu = 14$ and $\sigma = 3$. Thus, a baby who walks at 6 months is 8 months below the population mean. This puts this baby $2\frac{2}{3}$ standard deviations below the population mean. The Z score for this sample baby on the comparison distribution is thus -2.67 ($Z = [6 - 14]/3 = -2.67$). Figure 4–3 shows the score of our sample baby on the comparison distribution.

Tip for Success

If you are unsure about these symbols for population parameters, be sure to review Table 3–2 on p. 95.

STEP 5: DECIDE WHETHER TO REJECT THE NULL HYPOTHESIS

To decide whether to reject the null hypothesis, you compare your actual sample's Z score (from Step 4) to the cutoff Z score (from Step 3). In our example, the actual result was -2.67. Let's suppose the researchers had decided in advance that they would reject the null hypothesis if the sample's Z score was below -2. Since -2.67 is below -2, the researchers would reject the null hypothesis.

Or, suppose the researchers had used the more conservative 1% significance level. The needed Z score to reject the null hypothesis would then have been -2.33 or lower. But, again, the actual Z for the randomly selected baby was -2.67 (a more extreme score than -2.33). Thus, even with this more conservative cutoff, they would still reject the null hypothesis. This situation is shown in Figure 4–3. As you can see in the figure, the bottom 1% of the distribution is shaded. We recommend

[2] These days, when hypothesis testing is usually done on a computer, you have to decide in advance only on the cutoff probability. The computer prints out the exact probability of getting your result if the null hypothesis were true. You then just compare the printed-out probability to see if it is less than the cutoff probability level you set in advance. However, to *understand* what these probability levels mean, you need to learn the entire process, including how to figure the Z score for a particular cutoff probability.

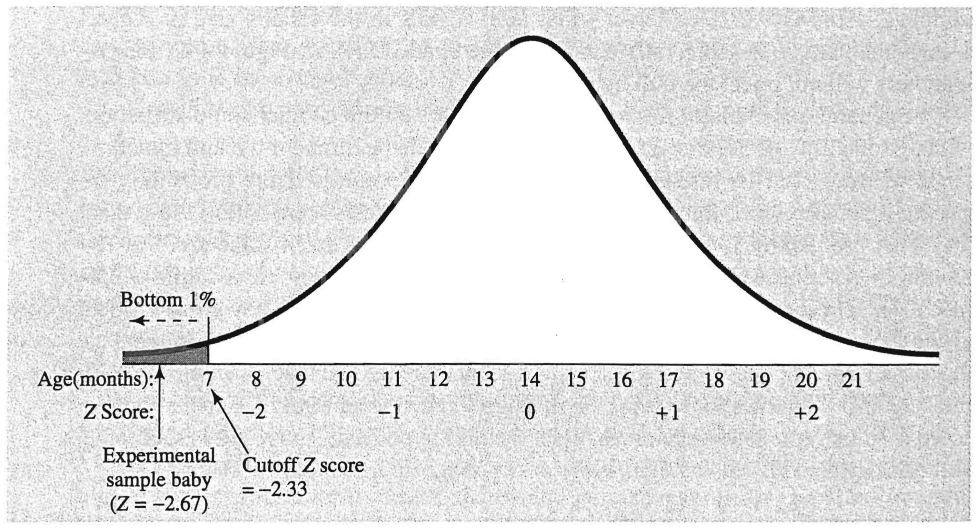

FIGURE 4–3 *Distribution of when babies begin to walk, showing both the bottom 1% and the single baby that is the sample studied (fictional data).*

that you always draw such a picture of the distribution. Be sure to shade in the part of the distribution that is *more extreme* (that is, further out in the tail) than the cutoff sample score. If your actual sample Z score falls within the shaded region, you can reject the null hypothesis. Since the sample Z score (of −2.67) in this example falls within the shaded tail region, the researchers can reject the null hypothesis.

If the researchers reject the null hypothesis, what remains is the research hypothesis. In this example, the research team can conclude that the results of their study support the research hypothesis, that babies who take the specially purified vitamin walk earlier than other babies.

IMPLICATIONS OF REJECTING OR FAILING TO REJECT THE NULL HYPOTHESIS

It is important to emphasize two points about the conclusions you can make from the hypothesis-testing process. First, suppose you reject the null hypothesis. Therefore, your results support the research hypothesis (as in our example). You would still not say that the results *prove* the research hypothesis or that the results show that the research hypothesis is *true.* This would be too strong because the results of research studies are based on probabilities. Specifically, they are based on the probability being low of getting your result if the null hypothesis were true. *Proven* and *true* are okay in logic and mathematics, but to use these words in conclusions from scientific research is quite unprofessional. (It is okay to use *true* when speaking hypothetically—for example, "*if* this hypothesis *were* true, then ..."—but not when speaking of conclusions about an actual result.) What you do say when you reject the null hypothesis is that the results are *statistically significant.*

Second, when a result is not extreme enough to reject the null hypothesis, you do not say that the result *supports* the null hypothesis. You simply say the result is *not statistically significant.*

A result that is not strong enough to reject the null hypothesis means the study was inconclusive. The results may not be extreme enough to reject the null hypothesis, but the null hypothesis might still be false (and the research hypothesis true).

Suppose in our example that the specially purified vitamin had only a slight but still real effect. In that case, we would not expect to find a baby given the purified vitamin to be walking a lot earlier than babies in general. Thus, we would not be able to reject the null hypothesis, even though it is false. (You will learn more about such situations in the Decision Errors section later in this chapter.)

Showing the null hypothesis to be true would mean showing that there is absolutely no difference between the populations. It is always possible that there is a difference between the populations, but that the difference is much smaller than what the particular study was able to detect. Therefore, when a result is not extreme enough to reject the null hypothesis, the results are *inconclusive.* Sometimes, however, if studies have been done using large samples and accurate measuring procedures, evidence may build up in support of something close to the null hypothesis—that there is at most very little difference between the populations. (We have more to say on this important issue later in this chapter and in Chapter 6.)

SUMMARY OF STEPS OF HYPOTHESIS TESTING

Here is a summary of the five steps of hypothesis testing.

❶ **Restate the question as a research hypothesis and a null hypothesis about the populations.**

❷ **Determine the characteristics of the comparison distribution.**

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.**

❹ **Determine your sample's score on the comparison distribution.**

❺ **Decide whether to reject the null hypothesis.**

A SECOND EXAMPLE

Here is another fictional example. Two happy-go-lucky personality psychologists are examining the theory that happiness comes from positive experiences. In particular, these researchers argue that if people have something very fortunate happen to them, they will become very happy and will still be happy 6 months later. So the researchers plan the following experiment: A person will be randomly selected from the North American adult public and given $10 million. Six months later, this person's happiness will be measured. It is already known (in this fictional example) what the distribution of happiness is like in the general population of North American adults, and this is shown in Figure 4–4. On the test being used, the mean happiness score is 70, the standard deviation is 10, and the distribution is approximately normal.

The psychologists now carry out the hypothesis-testing procedure. That is, the researchers consider how happy the person would have to be before they can confidently reject the null hypothesis that receiving that much money does *not* make people happier 6 months later. If the researchers' result shows a very high level of happiness, the psychologists will *reject* the null hypothesis and conclude that getting $10 million probably *does* make people happier 6 months later. But if the result is not very extreme, these researchers would conclude that there is not sufficient evidence to reject the null hypothesis, and the results of the experiment are inconclusive.

Now let us consider the hypothesis-testing procedure in more detail in this example, following the five steps.

❶ **Restate the question as a research hypothesis and a null hypothesis about the populations.** There are two populations of interest:

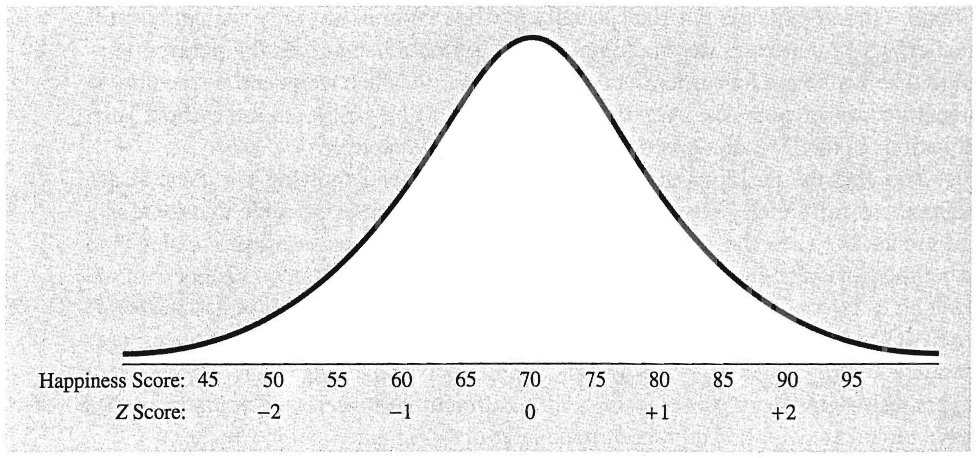

FIGURE 4–4 *Distribution of happiness scores (fictional data).*

Population 1: People who 6 months ago received \$10 million.
Population 2: People who 6 months ago did not receive \$10 million.

The prediction of the personality psychologists, based on their theory of happiness, is that Population 1 people will on the average be happier than Population 2 people: In symbols, $\mu_1 > \mu_2$. The null hypothesis is that Population 1 people (those who get \$10 million) will not be happier than Population 2 people (those who do not get \$10 million).

❷ **Determine the characteristics of the comparison distribution.** The comparison distribution is the distribution that represents the population situation if the null hypothesis is true. If the null hypothesis is true, the distributions of Populations 1 and 2 are the same. We know Population 2's distribution (it is normally distributed with $\mu = 70$ and $\sigma = 10$), so we can use it as the comparison distribution.

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** What kind of result would be extreme enough to convince us to reject the null hypothesis? In this example, assume that the researchers decided the following in advance: They will reject the null hypothesis as too unlikely if the results would occur less than 5% of the time if this null hypothesis were true. We know that the comparison distribution is a normal curve. Thus, we can figure that the top 5% of scores from the normal curve table begin at a *Z* score of about 1.64. This means the researchers would set as the cutoff point for rejecting the null hypothesis to be a result in which the sample's *Z* score on the comparison distribution is at or above 1.64. (The mean of the comparison distribution is 70 and the standard deviation is 10. Therefore, the null hypothesis would be rejected if the sample result was at or above 86.4.)

❹ **Determine your sample's score on the comparison distribution.** Now for the results: Six months after giving this randomly selected person \$10 million, the now very wealthy research participant takes the happiness test. The person's score is 80. As you can see from Figure 4–4, a score of 80 has a *Z* score of +1 on the comparison distribution.

❺ **Decide whether to reject the null hypothesis.** The *Z* score of the sample individual is +1. The researchers set the minimum *Z* score to reject the null hypothesis at +1.64. Thus, the sample score is not extreme enough to reject the null hypothesis. The experiment is inconclusive; researchers would say the results are "not

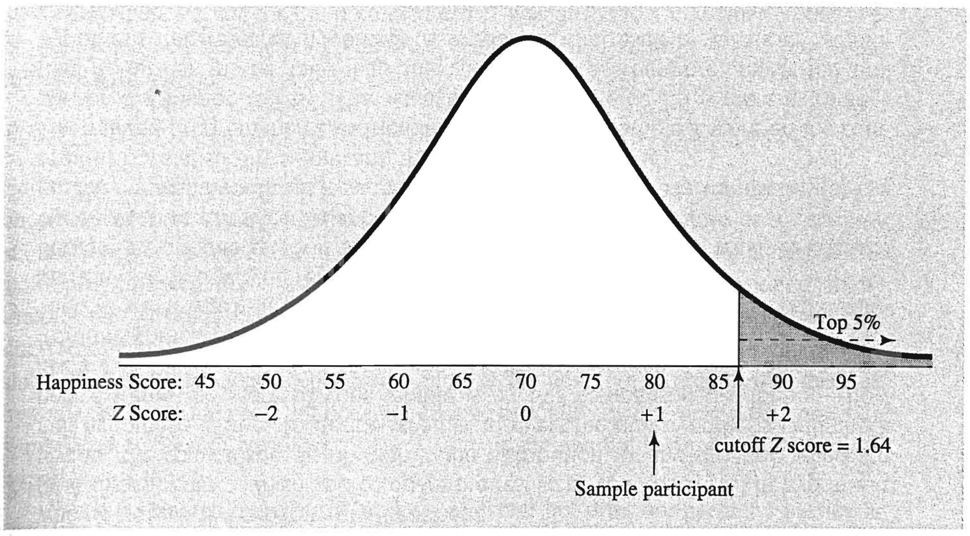

FIGURE 4–5 *Distribution of happiness scores with upper 5% shaded and showing the location of the sample participant (fictional data).*

statistically significant." Figure 4–5 shows the comparison distribution with the top 5% shaded and the location of the sample participant who received $10 million.

You may be interested to know that Brickman et al. (1978) carried out a more elaborate study based on the same question. They studied lottery winners as examples of people suddenly having a very positive event happen to them. Their results were similar to those in our fictional example: those who won the lottery were not much happier 6 months later than people who did not win the lottery. Also, another group they studied, people who had become paraplegics through a random accident, were not much less happy than other people 6 months later. These researchers studied fairly large numbers of individuals and explored the issue in several different ways. Their conclusion was that if a major event does have a lasting effect on happiness, it is probably not a very big one. So it looks like the lottery isn't the answer. (This pattern has also been found in other studies, e.g., Suh et al., 1996.)

HOW ARE YOU DOING?

1. A sample is given an experimental treatment that is predicted to make them score higher than the general public on a standard memory test. State (a) the null hypothesis and (b) the research hypothesis.
2. (a) What is a comparison distribution? (b) What role does it play in hypothesis testing?
3. What is the cutoff sample score?
4. Why do we say that hypothesis testing involves a double negative logic?
5. What can you conclude when (a) a result is so extreme that you reject the null hypothesis, and (b) a result is not very extreme so you cannot reject the null hypothesis?
6. A training program to increase friendliness is tried on one individual randomly selected from the general public. Among the general public (which does not get this training program) the mean on the friendliness measure is 30 with a standard deviation of 4. The researchers want to test their hypothesis at the 5% significance level. After going through the training program, this individual takes the friendliness measure and gets a score of 40. What should the researchers conclude?

Answers

1. (a) The population of people like those who get the experimental treatment score the same on the memory test as the population of individuals who do not get the experimental treatment. (b) The population of people like those who get the experimental treatment score higher on the memory test than the population of people who do not get the experimental treatment.
2. (a) A distribution to which you compare the results of your study. (b) In hypothesis testing, the comparison distribution is the distribution for the situation when the null hypothesis is true. To decide whether to reject the null hypothesis, you check how extreme the score of your sample is on this comparison distribution—how likely it would be to get a sample with a score this extreme if your sample came from this comparison distribution.
3. The Z score at which, if the sample's Z score is more extreme than it is on the comparison distribution, you reject the null hypothesis.
4. Because we are interested in the research hypothesis, but we test whether it is true by seeing if we can reject its opposite, the null hypothesis.
5. (a) The research hypothesis is supported; the result is statistically significant. (b) The result is not statistically significant; the result is inconclusive.
6. The training program increases friendliness. (The cutoff sample Z score on the comparison distribution is 1.64. The actual sample's Z score of 2.5 is more extreme—that is, further in the tail—than the cutoff Z score. Therefore, reject the null hypothesis; the research hypothesis is supported; the result is statistically significant.)

ONE-TAILED AND TWO-TAILED HYPOTHESIS TESTS

In our examples so far, the researchers were interested in only one direction of result. In our first example, researchers tested whether babies given the specially purified vitamin would walk *earlier* than babies in general. In the happiness example, the personality psychologists predicted the person who received $10 million would be *happier* than other people. The researchers in these studies were not interested in the possibility that giving the specially purified vitamin would cause babies to start walking *later* or that people getting $10 million might become *less* happy.

DIRECTIONAL HYPOTHESES AND ONE-TAILED TESTS

directional hypotheses

The purified vitamin and happiness studies are examples of testing **directional hypotheses.** Both studies focused on a specific direction of effect. When a researcher makes a directional hypothesis, the null hypothesis is also, in a sense, directional. Suppose the research hypothesis is that getting $10 million will make a person happier. The null hypothesis, then, is that the money will either have no effect or make the person less happy. (In symbols, if the research hypothesis is $\mu_1 > \mu_2$, then the null hypothesis is $\mu_1 \leq \mu_2$; $\leq$ is the symbol for less than or equal to.) Thus, in Figure 4–5, to reject the null hypothesis, the sample had to have a score in one particular tail of the comparison distribution—the upper extreme or tail (in this example, the top 5%) of the comparison distribution. (When it comes to rejecting the null hypothesis with a directional hypothesis, a score at the other tail would be the same as a score in the middle—that is, it would not allow you to reject the null hypothesis.) For this reason, the test of a directional hypothesis is called a **one-tailed test.** A one-tailed test can be one-tailed in either direction. In the happiness study example, the tail for the predicted effect was at the high end. In the baby

one-tailed test

study example, the tail for the predicted effect was at the low end (that is, the prediction tested was that babies given the specially purified vitamin would start walking unusually *early*).

NONDIRECTIONAL HYPOTHESES AND TWO-TAILED TESTS

Sometimes, a research hypothesis states that an experimental procedure will have an effect, without saying whether it will produce a very high score or a very low score. Suppose an organizational psychologist is interested in how a new social skills program will affect productivity. The program could improve productivity by making the working environment more pleasant. Or, the program could hurt productivity by encouraging people to socialize instead of work. The research hypothesis is that the social skills program *changes* the level of productivity; the null hypothesis is that the program does not change productivity one way or the other. In symbols, the research hypothesis is $\mu_1 \neq \mu_2$ ($\neq$ is the symbol for not equal); the null hypothesis is $\mu_1 = \mu_2$.

When a research hypothesis predicts an effect but does not predict a particular direction for the effect, it is called a **nondirectional hypothesis.** To test the significance of a nondirectional hypothesis, you have to take into account the possibility that the sample could be extreme at either tail of the comparison distribution. Thus, this is called a **two-tailed test.**

nondirectional hypothesis

two-tailed test

DETERMINING CUTOFF SCORES WITH TWO-TAILED TESTS

There is a special complication in a two-tailed test. You have to divide up the significance percentage between the two tails. For example, with a 5% significance level, you reject a null hypothesis only if the sample is so extreme that it is in either the top 2.5% or the bottom 2.5%. This keeps the overall level of significance at a total of 5%.

Note that a two-tailed test makes the cutoff Z scores for the 5% level +1.96 and −1.96. For a one-tailed test at the 5% level, the cutoff was not so extreme—only +1.64 or −1.64, but only one side of the distribution was considered. These situations are shown in Figure 4–6a.

Using the 1% significance level, a two-tailed test (.5% at each tail) has cutoffs of +2.58 and −2.58, while a one-tailed test's cutoff is either +2.33 or −2.33. These situations are shown in Figure 4–6b. The Z score cutoffs for one-tailed and two-tailed tests for the .05 and .01 significance levels are also summarized in Table 4–1.

WHEN TO USE ONE-TAILED OR TWO-TAILED TESTS

If the researcher decides in advance to use a one-tailed test, then the sample's score does not need to be so extreme to be significant as it would need to be with a two-tailed test. Yet there is a price: With a one-tailed test, if the result is extreme in the direction opposite to what was predicted, no matter how extreme, the result cannot be considered statistically significant.

In principle, you plan to use a one-tailed test when you have a clearly directional hypothesis and a two-tailed test when you have a clearly nondirectional hypothesis. In practice, it is not so simple. Even when a theory clearly predicts a particular result, the actual result may come out opposite to what you expected. Sometimes, this opposite may be more interesting than what you had predicted. (What if, as in all the fairy tales about wish-granting genies and fish, receiving $10 million and being able to fulfill almost any wish had made that one individual miserable?) By using one-tailed tests, we risk having to ignore possibly important results.

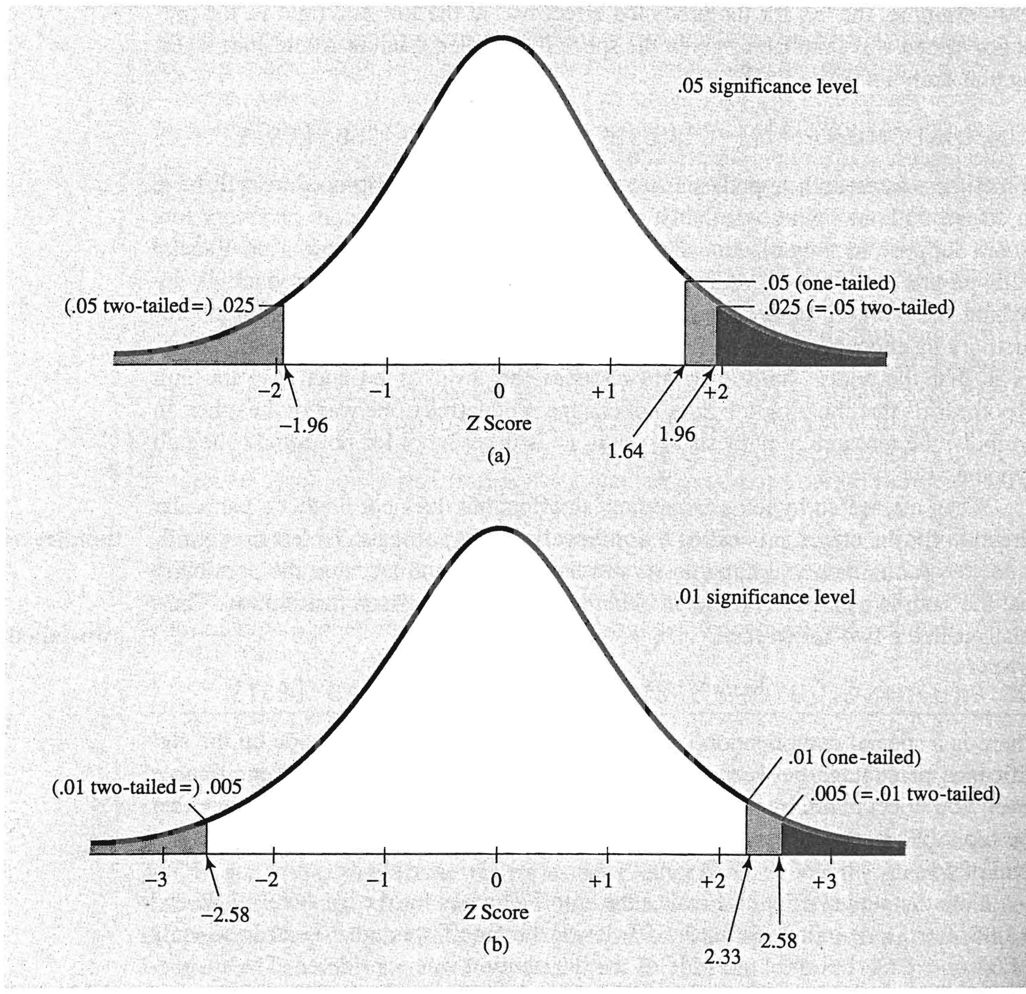

FIGURE 4–6 *Significance level cutoffs for one-tailed and two-tailed tests: (a) .05 significance level; (b) .01 significance level. (The one-tailed tests in these examples assume the prediction was for a high score. You could instead have a one-tailed test where the prediction is for the lower, left tail.)*

For these reasons, researchers disagree about whether one-tailed tests should be used, even when there is a clearly directional hypothesis. To be safe, many researchers use two-tailed tests for both nondirectional and directional hypotheses. If the two-tailed test is significant, then the researcher looks at the result to see the direction and considers the study significant in that direction.[3] In practice, always using two-tailed tests is a conservative procedure. This is because the cutoff scores are more extreme for a two-tailed test, so it is less likely that a two-tailed test will

[3]Leventhal and Huynh (1996) argue that this procedure is technically incorrect. If you are testing a nondirectional hypothesis, you should only make nondirectional conclusions. A better procedure, they suggest, is to use a "directional two-tailed test"—what amounts to two simultaneous one-tailed tests (one in each direction). Thus, if you want an overall significance level of .05, you use a directional two-tailed test in which the two one-tailed subparts each use the .025 level. (See Jones & Tukey, 2000, for a related approach.) Leventhal and Huynh's way of thinking about two-tailed tests does seem to be more logical and to have some technical advantages. However, researchers have not yet adopted this approach, and for most purposes the result is the same. Thus, in this book we stick to the more traditional approach.

TABLE 4–1 One-Tailed and Two-Tailed Cutoff Z Scores for the .05 and .01 Significance Levels

		Type of Test	
		One-Tailed	*Two-Tailed*
Significance	*.05*	−1.64 *or* 1.64	−1.96 *and* 1.96
Level	*.01*	−2.33 *or* 2.33	−2.58 *and* 2.58

give a significant result. Thus, if you do get a significant result with a two-tailed test, you are more confident about the conclusion. In fact, in most psychology research articles, unless the researcher specifically states that a one-tailed test was used, it is assumed that it was a two-tailed test.

In practice, however, it is our experience that most research results are either so extreme that they will be significant whether you use a one-tailed or two-tailed test or so far from extreme that they would not be significant no matter what you use. But what happens when a result is less certain? The researcher's decision about one-tailed or two-tailed tests now can make a big difference. In this situation the researcher tries to use the type of test that will give the most accurate and noncontroversial conclusion. The idea is to let nature—and not a researcher's decisions—determine the conclusion as much as possible. Further, whenever a result is less than completely clear one way or the other, most researchers will not be comfortable drawing strong conclusions until more research is done.

EXAMPLE OF HYPOTHESIS TESTING WITH A TWO-TAILED TEST

Here is one more fictional example, this time using a two-tailed test. Clinical psychologists at a residential treatment center have developed a new type of therapy to reduce depression that they believe is more effective than the therapy now given. However, as with any treatment, it is also possible that it could make patients do worse. Thus, the clinical psychologists make a nondirectional hypothesis.

The psychologists randomly select an incoming patient to receive the new form of therapy instead of the usual therapy. (In a real study, of course, more than one patient would be selected; but let's assume that only one person has been trained to do the new therapy and she has time to treat only one patient.) After 4 weeks, the patient fills out a standard depression scale that is given automatically to all patients after 4 weeks. The standard scale has been given at this treatment center for a long time. Thus, the psychologists know in advance the distribution of depression scores at 4 weeks for those who receive the usual therapy: It follows a normal curve with a mean of 69.5 and a standard deviation of 14.1. (These figures correspond roughly to the depression scores found in a national survey of 75,000 psychiatric patients given a widely used standard test; Dahlstrom et al., 1986.) This distribution is shown in Figure 4–7.

The clinical psychologists then carry out the five steps of hypothesis-testing.

❶ **Restate the question as a research hypothesis and a null hypothesis about the populations.** There are two populations of interest:

Population 1: Patients diagnosed as depressed who receive the new therapy.
Population 2: Patients diagnosed as depressed who receive the usual therapy.

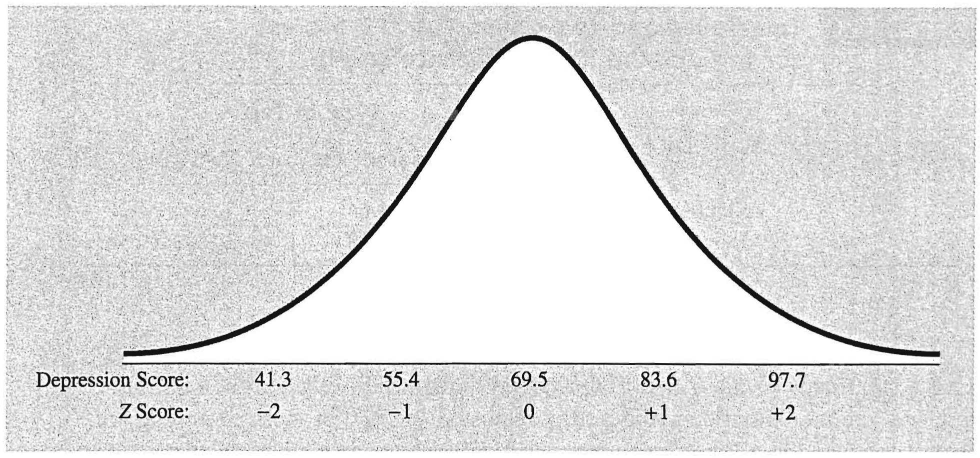

FIGURE 4–7 *Distribution of depression scores at 4 weeks after admission for diagnosed depressed psychiatric patients receiving the standard therapy (fictional data).*

Tip for Success

Remember that the research hypothesis and null hypothesis must always be complete opposites. Researchers specify the research hypothesis and this determines the corresponding null hypothesis.

The research hypothesis is that when measured on depression 4 weeks after admission, patients who receive the new therapy (Population 1) will on the average score differently from patients who receive the current therapy (Population 2). In symbols, the research hypothesis is $\mu_1 \neq \mu_2$. The opposite of the research hypothesis, the null hypothesis, is that patients who receive the new therapy will have the same average depression level as the patients who receive the usual therapy. (That is, the depression level measured after 4 weeks will have the same mean for Populations 1 and 2.) In symbols, the null hypothesis is $\mu_1 = \mu_2$.

❷ **Determine the characteristics of the comparison distribution.** If the null hypothesis is true, the distributions of Populations 1 and 2 are the same. We know the distribution of Population 2 (it is the one shown in Figure 4–7). Thus, we can use Population 2 as our comparison distribution. As noted, it follows a normal curve, with $\mu = 69.5$ and $\sigma = 14.1$.

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** The clinical psychologists select the 5% significance level. They have made a nondirectional hypothesis and will therefore use a two-tailed test. Thus, they will reject the null hypothesis only if the patient's depression score is in either the top or bottom 2.5% of the comparison distribution. In terms of Z scores, these cutoffs are +1.96 and −1.96 (see Figure 4–8 and Table 4–1).

❹ **Determine your sample's score on the comparison distribution.** The patient who received the new therapy was measured 4 weeks after admission. The patient's score on the depression scale was 41, which is a Z score on the comparison distribution of −2.02. That is, $Z = (X - M)/SD = (41 - 69.5)/14.1 = -2.02$.

❺ **Decide whether to reject the null hypothesis.** A Z score of −2.02 is slightly more extreme than a Z score of −1.96, which is where the lower 2.5% of the comparison distribution begins. Notice in Figure 4–8 that the Z score of −2.02 falls within the shaded area in the left tail of the comparison distribution. This Z score of −2.02 is a result so extreme that it is unlikely to have occurred if this patient were from a population no different from Population 2. Therefore, the clinical psychologists reject the null hypothesis. The result is statistically significant and it supports the research hypothesis that depressed patients receiving the new therapy have different depression levels than depressed patients that receive the usual therapy.

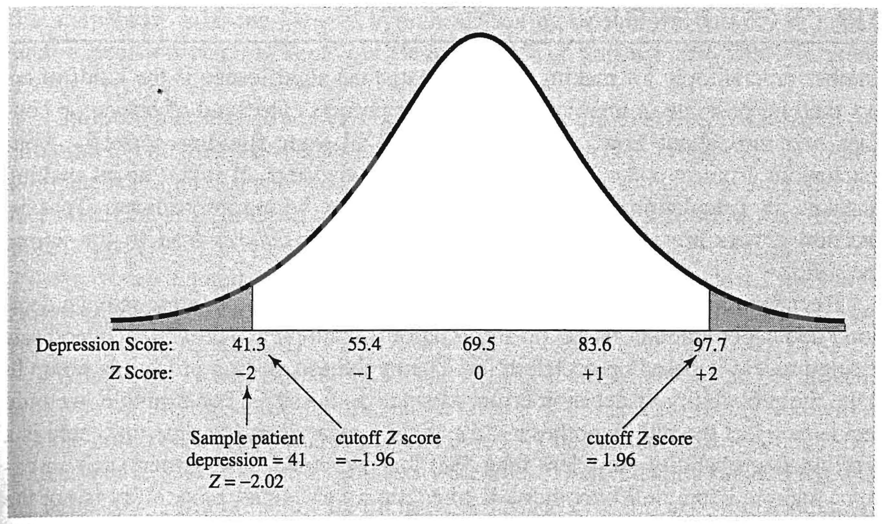

FIGURE 4–8 *Distribution of depression scores with upper and lower 2.5% shaded and showing the sample patient who received the new therapy (fictional data).*

Tip for Success

When carrying out the five steps of hypothesis testing, always draw a figure like Figure 4–8. Be sure to include the cutoff score(s) and shade the appropriate tail(s). If the sample score falls within a shaded tail region, the null hypothesis can be rejected and the result is statistically significant. If the sample score does not fall within a shaded tail region, the null hypothesis cannot be rejected.

HOW ARE YOU DOING?

1. What is a nondirectional hypothesis test?
2. What is a two-tailed test?
3. Why do you use a two-tailed test when testing a nondirectional hypothesis?
4. What is the advantage of using a one-tailed test when your theory predicts a particular direction of result?
5. Why might you use a two-tailed test even when your theory predicts a particular direction of result?
6. A researcher predicts that making a person hungry will affect how he or she does on a coordination test. A randomly selected person agrees not to eat for 24 hours before taking a standard coordination test and gets a score of 400. For people in general of this age group tested under normal conditions, coordination scores are normally distributed with a mean of 500 and a standard deviation of 40. Using the .01 significance level, what should the researcher conclude?

ANSWERS

1. A nondirectional hypothesis test is a hypothesis test in which you do not predict a particular direction of difference.
2. A two-tailed test is one in which the overall percentage for the cutoff is divided between the two tails of the comparison distribution. A two-tailed test is used to test the significance of a nondirectional hypothesis.
3. You use a two-tailed test when testing a nondirectional hypothesis because an extreme result in either direction would support the research hypothesis.
4. The cutoff for a one-tailed test is not so extreme; thus, if your result comes out in the predicted direction, it is more likely to be significant. The cutoff is not so extreme because the entire percentage (say 5%) is put in one tail instead of being divided between two tails.
5. It lets you count as significant an extreme result in either direction; if you used a one-tailed test and the result came out opposite to the prediction, it could not be called statistically significant.
6. The cutoffs are +2.58 and −2.58. The sample person's Z score is $(400 - 500)/40 = -2.5$. The result is not significant; the study is inconclusive.

DECISION ERRORS

Another crucial topic for making sense of statistical significance is the kinds of errors that are possible in the hypothesis-testing process. The kind of errors we consider here are about how, in spite of doing all your figuring correctly, your conclusions from hypothesis-testing can still be incorrect. It is *not* about making mistakes in calculations or even about using the wrong procedures. That is, **decision errors** are situations in which the *right procedures* lead to the *wrong decisions.*

decision errors

Decision errors are possible in hypothesis testing because you are making decisions about populations based on information in samples. The whole hypothesis testing process is based on probabilities. The hypothesis-testing process is set up to make the probability of decision errors as small as possible. For example, we only decide to reject the null hypothesis if a sample's mean is so extreme that there is a very small probability (say, less than 5%) that we could have gotten such an extreme sample if the null hypothesis is true. But a very small probability is not the same as a zero probability! Thus, in spite of your best intentions, decision errors are always possible.

There are two kinds of decision errors in hypothesis testing: Type I error and Type II error.[4]

TYPE I ERROR

Type I error

You make a **Type I error** if you reject the null hypothesis when in fact the null hypothesis is true. Or, to put it in terms of the research hypothesis, you make a Type I error when you conclude that the study supports the research hypothesis when in reality the research hypothesis is false.

Suppose you carried out a study in which you had set the significance level cutoff at a very lenient probability level, such as 20%. This would mean that it would not take a very extreme result to reject the null hypothesis. If you did many studies like this, you would often (about 20% of the time) be deciding to consider the research hypothesis supported when you should not. That is, you would have a 20% chance of making a Type I error.

Even when you set the probability at the conventional .05 or .01 levels, you will still make a Type I error sometimes (5% or 1% of the time). Consider again the example of giving the new therapy to a depressed patient. Suppose the new therapy is not more effective than the usual therapy. However, in randomly picking a sample of one depressed patient to study, the clinical psychologists might just happen to pick a patient whose depression would respond equally well to the new therapy and the usual therapy. Randomly selecting a sample patient like this is unlikely, but such extreme samples are possible, and should this happen, the clinical psychologists would reject the null hypothesis and conclude that the new therapy is different than the usual therapy. Their decision to reject the null hypothesis would be wrong—a Type I error. Of course, the researchers could not know they had made a decision error of this kind. What reassures researchers is

[4]You may also occasionally hear about a Type III error. This is concluding there is a significant result in a particular direction, when the true effect is in the opposite direction.

that they know from the logic of hypothesis testing that the probability of making such a decision error is kept low (less than 5% if you use the .05 significance level).

Still, the fact that Type I errors can happen at all is of serious concern to psychologists, who might construct entire theories and research programs, not to mention practical applications, based on a conclusion from hypothesis testing that is in fact mistaken. It is because these errors are of such serious concern that they are called Type I.

As we have noted, researchers cannot tell when they have made a Type I error. However, they can try to carry out studies so that the chance of making a Type I error is as small as possible.

What is the chance of making a Type I error? It is the same as the significance level you set. If you set the significance level at $p < .05$, you are saying you will reject the null hypothesis if there is less than a 5% (.05) chance that you could have gotten your result if the null hypothesis were true. When rejecting the null hypothesis in this way, you are allowing up to a 5% chance that you got your results even though the null hypothesis was actually true. That is, you are allowing a 5% chance of a Type I error.

The significance level, which is the chance of making a Type I error, is called **alpha** (the Greek letter α). The lower the alpha, the smaller the chance of a Type I error. Researchers who do not want to take a lot of risk set alpha lower than .05, such as $p < .001$. In this way the result of a study has to be very extreme in order for the hypothesis testing process to reject the null hypothesis.

alpha

Using a .001 significance level is like buying insurance against making a Type I error. However, as when buying insurance, the better the protection, the higher the cost. There is a cost in setting the significance level at too extreme a level. We turn to that cost next.

TYPE II ERROR

If you set a very stringent significance level, such as .001, you run a different kind of risk. With a very stringent significance level, you may carry out a study in which in reality the research hypothesis is true, but the result does not come out extreme enough to reject the null hypothesis. Thus, the decision error you would make is in *not* rejecting the null hypothesis when in reality the null hypothesis is false. To put this in terms of the research hypothesis, you make this kind of decision error when the hypothesis-testing procedure leads you to decide that the results of the study are inconclusive when in reality the research hypothesis is true. This is called a **Type II error.** The probability of making a Type II error is called **beta** (the Greek letter β). (Do not confuse this beta with the standardized regression coefficient that you will learn about in Chapter 12, which is also called beta.)

Type II error
beta

Consider again our depression therapy example. Suppose that, in truth, the new therapy is better at treating depression than the usual therapy. However, in conducting your particular study, the results for the sample patient are not strong enough to allow you to reject the null hypothesis. Perhaps the random sample patient that you selected to try out the new therapy happened to be a person who would not respond to either the new therapy or the usual therapy. The results would not be significant. Having decided not to reject the null hypothesis, and thus refusing to draw a conclusion, would be a Type II error.

Tip for Success

It is very easy to get confused between a Type I error and a Type II error. Be sure you understand each type of error (and the difference between them) before reading on in this chapter.

Type II errors especially concern psychologists interested in practical applications, because a Type II error could mean that a valuable practical procedure is not used.

As with a Type I error, you cannot know when you have made a Type II error. But researchers can try to carry out studies so as to reduce the probability of making one. One way of buying insurance against a Type II error is to set a very lenient significance level, such as $p < .10$ or even $p < .20$. In this way, even if a study produces only a very small effect, this effect has a good chance of being significant. There is a cost to this insurance policy too.

RELATION OF TYPE I AND TYPE II ERRORS

When it comes to setting significance levels, protecting against one kind of decision error increases the chance of making the other. The insurance policy against Type I error (setting a significance level of, say, .001) has the cost of increasing the chance of making a Type II error. (This is because with a stringent significance level like .001, even if the research hypothesis is true, the results have to be quite strong to be extreme enough to reject the null hypothesis.) The insurance policy against Type II error (setting a significance level of, say, .20) has the cost of increasing the chance of making a Type I error. (This is because with a level of significance like .20, even if the null hypothesis is true, it is fairly easy to get a significant result just by accidentally getting a sample that is higher or lower than the general population before doing the study.)

The trade-off between these two conflicting concerns usually is worked out by compromise—thus the standard 5% and 1% significance levels.

SUMMARY OF POSSIBLE OUTCOMES OF HYPOTHESIS TESTING

The entire issue of possibly correct or mistaken conclusions in hypothesis testing is shown in Table 4–2. Along the top of this table are the two possibilities about whether the null hypothesis or the research hypothesis is really true. (Remember, you never actually know this.) Along the side is whether, after hypothesis testing, you decide that the research hypothesis is supported (reject the null hypothesis) or decide that the results are inconclusive (do not reject the null hypothesis). Table 4–2

TABLE 4–2 Possible Correct and Incorrect Decisions in Hypothesis Testing

		Real Situation (in practice, unknown)	
		Null Hypothesis True	*Research Hypothesis True*
Conclusion Using Hypothesis-testing Procedure	*Research hypothesis supported (reject null hypothesis)*	Error (Type I) α	Correct decision
	Study is inconclusive (do not reject null hypothesis)	Correct decision	Error (Type II) β

shows that there are two ways to be correct and two ways to be in error in any hypothesis testing situation. You will learn more about these possibilities in Chapter 6.

HOW ARE YOU DOING?

1. What is a decision error?
2. (a) What is a Type I error? (b) Why is it possible? (c) What is its probability? (d) What is this probability called?
3. (a) What is a Type II error? (b) Why is it possible? (c) What is its probability called?
4. If you set a lenient alpha level (say .25), what is the effect on the probability of (a) Type I error and (b) Type II error?
5. If you set a stringent alpha level (say .001), what is the effect on the probability of (a) Type I error and (b) Type II error?

ANSWERS

1. A decision error is a conclusion from hypothesis testing that does not match reality.
2. (a) Rejecting the null hypothesis (and thus supporting the research hypothesis) when the null hypothesis is actually true (and the research hypothesis false) is a Type I error. (b) You reject the null hypothesis when a sample's result is so extreme that it is unlikely you would have gotten that result if the null hypothesis is true. However, even though it is unlikely, it is still possible that the null hypothesis is true. (c) Its probability is the significance level (such as .05). (d) Alpha.
3. (a) Failing to reject the null hypothesis (and thus failing to support the research hypothesis) when the null hypothesis is actually false (and the research hypothesis true) is a Type II error. (b) You reject the null hypothesis when a sample's result is so extreme it is unlikely you would have gotten that result if the null hypothesis is true. However, the null hypothesis could be false, but the sample mean may not be extreme enough to reject the null hypothesis. (c) Beta.
4. (a) It is high; (b) it is low.
5. (a) It is low; (b) it is high.

CONTROVERSY: SHOULD SIGNIFICANCE TESTS BE BANNED?

In recent years, there has been a major controversy about significance testing itself, with a concerted movement on the part of a small but vocal group of psychologists to ban significance tests completely! This is a radical suggestion with far-reaching implications (for at least half a century, nearly every research study in psychology has used significance tests). There probably has been more written in the major psychology journals in the last 10 years about this controversy than ever before in history about any issue having to do with statistics.

The discussion has gotten so heated that one article began as follows:

> It is not true that a group of radical activists held 10 statisticians and six editors hostage at the 1996 convention of the American Psychological Society and chanted, "Support the total test ban!" and "Nix the null!" (Abelson, 1997, p. 12).

Since this is by far the most important controversy in years regarding statistics as used in psychology, we discuss the issues in at least three different places. In this chapter we focus on some basic challenges to hypothesis testing. In Chapters 5 and

6, we cover other topics that relate to aspects of hypothesis testing that you will learn about in those chapters.

Before discussing this controversy, you should be reassured that you are not learning about hypothesis testing for nothing. Whatever happens in the future, you absolutely have to understand hypothesis testing to make sense of virtually every research article published in the past. Further, in spite of the controversy that has raged for the last decade, it is extremely rare to see new articles that do not use significance testing. Thus, it is doubtful that there will be any major shifts in the near future. Finally, even if hypothesis testing is completely abandoned, the alternatives (which involve procedures you will learn about in Chapters 5 and 6) require understanding virtually all of the logic and procedures we are covering here.

So, what is the big controversy? Some of the debate concerns subtle points of logic. For example, one issue relates to whether it makes sense to worry about rejecting the null hypothesis when a hypothesis of no effect whatsoever is extremely unlikely to be true. We discuss this issue briefly in Box 4–1. Another issue is about the foundation of hypothesis testing in terms of populations and samples, since in most experiments the samples we use are not randomly selected from any definable population. We discussed some points relating to this issue in Chapter 3. Finally, some have questioned the appropriateness of concluding that if the data are inconsistent with the null hypothesis, this should be counted as evidence for the research hypothesis. This controversy becomes rather technical, but our own view is that given recent considerations of the issues, the way researchers in psychology use hypothesis testing is reasonable (Nickerson, 2000).

However, the biggest complaint against significance tests, and the one that has received almost universal agreement, is that they are misused. In fact, opponents of significance tests argue that even if there were no other problems with the tests, they should be banned simply because they are so often and so badly misused. There are two main ways in which they are misused; one we can consider now, the other must wait until we have covered a topic you learn in Chapter 6.

A major misuse of significance tests is the tendency for researchers to decide that if a result is not significant, the null hypothesis is shown to be true. We have emphasized that when the null hypothesis is not rejected, the results are inconclusive. The error of concluding the null hypothesis is true from failing to reject it is extremely serious, because important theories and methods may be considered false just because a particular study did not get strong enough results. (You learn in Chapter 6 that it is quite easy for a true research hypothesis not to come out significant just because there were too few people in the study or the measures were not very accurate. In fact, Hunter [1997] argues that in about 60% of psychology studies, we are likely to get nonsignificant results even when the research hypothesis is actually true.)

What should be done? The general consensus seems to be that we should keep significance tests, but better train our students not to misuse them (hence, the emphasis on these points in this book). We should not, as it were, throw the baby out with bathwater. To address this controversy, the American Psychological Association (APA) established a committee of eminent psychologists renowned for their statistical expertise. The committee met over a 2-year period, circulated a preliminary report, and considered reactions to it from a large number of researchers. In the end, they strongly condemned various misuses of significance testing of the kind we have been discussing, but they left its use up to the decision of each researcher. In their report they concluded:

> Some had hoped that this task force would vote to recommend an outright ban on the use of significance tests in psychology journals. Although this might eliminate some abuses, the committee thought there were enough counterexamples (e.g., Abelson, 1997) to justify forbearance (Wilkinson & Task Force on Statistical Inference, 1999, pp. 602–603).

A few yeas ago, Nickerson (2000) systematically reviewed more than 400 articles on this controversy. His conclusion, with which we agree (as do probably most psychology researchers), is that significance testing "is easily misunderstood and misused but that when applied with good judgment it can be an effective aid in the interpretation of experimental data" (p. 241).

BOX 4-1 To Be or Not to Be—But Can Not Being Be? The Problem of Whether and When to Accept the Null Hypothesis

The null hypothesis states that there is no difference between populations represented by different groups or experimental conditions. As we have seen, the usual rule in statistics is that a study cannot find the null hypothesis to be true. A study can only tell you that you cannot reject the null hypothesis. That is, a study that fails to reject the null hypothesis is simply uninformative. Such studies tend not to be published, obviously. However, much work could be avoided if people knew what interventions, measures, or experiments had not worked. Indeed, Greenwald (1975) reports that sometimes ideas have been assumed too long to be true just because a few studies found results supporting them, while many more, unreported, had not.

Frick (1995) has pointed out yet another serious problem with being rigidly uninterested in the null hypothesis: Sometimes it may be true that one thing has no effect on another. This does not mean that there would be a zero relationship of no correlation or no difference at all—a result that is almost impossible in many situations. It would only mean that the effect was so small that it probably represented no real, or at least no important, relationship or difference.

The problem is knowing when to conclude that the null hypothesis (or something close to it) might be true. Frick (1995) gives three criteria. First, the null hypothesis should seem possible. Second, the results in the study should be consistent with the null hypothesis and not easily interpreted any other way. Third, and most important, the researcher has to have made a strong effort to find the effect that he or she wants to conclude is not there. Among other things, this means studying a large sample and having very thorough and sensitive measurement. If the study is an experiment, the experimenter should have tried to produce the difference by using a strong manipulation and rigorous conditions of testing.

Frick (1995) points out that all of this leaves a subjective element to the acceptance of the null hypothesis. Who decides when a researcher's effort was strong enough? Subjective judgments are a part of science, like it or not. For example, reviewers of articles submitted for publication in scientific journals have to decide if a topic is important enough to compete for limited space in those journals. Further, the null hypothesis is being accepted all the time anyway. (For example, many psychologists accept the null hypothesis about the effect of extrasensory perception.) It is better to discuss our basis for accepting the null hypothesis than just to accept it.

What are we to make of all this? It is clear that just failing to reject the null hypothesis is not the same as supporting it. Indeed, equating these is a serious mistake. But Frick (1995) reminds us that there are situations in which the evidence ought to convince us that something like the null hypothesis is likely to be the case.

HYPOTHESIS TESTS IN RESEARCH ARTICLES

In general, hypothesis testing is reported in research articles as part of one of the specific methods you learn in later chapters. For each result of interest, the researcher usually first indicates whether the result was statistically significant. (Note that, as with the first example below, the researcher will not necessarily use the word "significant," so look out for other indicators, such as reporting that scores on a variable decreased, increased, or were associated with scores on another variable.) Next, the researcher usually gives the symbol associated with the specific method used in figuring the probabilities, such as t, F, or χ^2 (see Chapters 7 to 13). Finally, there will be an indication of the significance level, such as $p < .05$ or $p < .01$. (The researcher will usually also provide other information, such as the mean and standard deviation of sample scores.) For example, Carver (2004) reported: "Frustration increased considerably from the start of the session ($M = 2.35$, $SD = 1.60$) to the end of the session ($M = 5.36$, $SD = 2.25$), $t(65) = 11.22$, $p < .001$." There is a lot here that you will learn about in later chapters, but the key thing to understand now about this result is the "$p < .001$." This means that the probability of the results if the null hypothesis (of no difference between the populations the groups represent) were true is less than .001 (.1%).

When a result is close, but does not reach the significance level chosen, it may be reported anyway as a "near significant trend," or as having "approached significance," with $p < .10$, for example. When a result is not even close to being extreme enough to reject the null hypothesis, it may be reported as "not significant" or the abbreviation *ns* will be used. Finally, whether or not a result is significant, it is increasingly common for researchers to report the exact p level—such as $p = .03$ or $p = .27$. The p reported here is based on the proportion of the comparison distribution that is more extreme than the sample score information that you could figure from the Z score for your sample and a normal curve table.

A researcher will usually note if a one-tailed test is used. When reading research articles, assume a two-tailed test was used if nothing is said otherwise. Even though a researcher has chosen a significance level in advance, such as .05, results that meet more rigorous standards may be noted as such. Thus, in the same article, you may see some results noted as "$p < .05$," others as "$p < .01$," and still others as "$p < .001$," for example.

Finally, the results of hypothesis testing may be shown only as asterisks in a table of results. In such tables, a result with an asterisk is significant, while a result without one is not. For example, Table 4–3 shows results of part of a study by Stipek and Ryan (1997) comparing economically disadvantaged and advantaged preschoolers. This table gives figures for variables measured by observing the children in the classroom, including means, standard deviations, and F statistics (an indication of the procedure used in this study to test significance, a procedure you will learn in Chapters 9 and 10). The important thing to look at for purposes of the present discussion are the asterisks (and the notes at the bottom of the table that go with them) telling you the significance levels for the various measures. For example, for calling attention to their accomplishments, disadvantaged children ($M = .20$) scored significantly higher than advantaged children ($M = .04$). The reverse pattern was seen for "Smiles after completing the task."

On the other hand, making positive social comparisons did not differ significantly between the groups (the means were .71 and .64, but these were not different enough to be significant in this study). Thus, we cannot conclude that for preschoolers, being disadvantaged has any relation to making positive social comparisons. It