Introduction to Hypothesis Testing

Chapter Outline

In this chapter, we introduce the crucial topic of **hypothesis testing.** A **hypothesis** is a prediction intended to be tested in a research study. The prediction may be based on informal observation (as in clinical or applied settings), on related results of previous studies, or on a broader *theory* about what is being studied. You can think of a **theory** as a set of principles that attempt to explain one or more facts, relationships, or events. A theory usually gives rise to various specific hypotheses that can be tested in research studies.

This chapter focuses on the basic logic for analyzing results of a research study to test a hypothesis. The central theme of hypothesis testing has to do with the important distinction between sample and population. Hypothesis testing is a systematic procedure for deciding whether the results of a research study, which examines a sample, support a hypothesis that applies to a population. Hypothesis testing is the central theme in most behavioral and social science research.

hypothesis testing Procedure for deciding whether the outcome of a study (results for a sample) supports a particular theory or practical innovation (which is thought to apply to a population).

hypothesis A prediction, often based on observation, previous research, or theory, that is tested in a research study.

theory A set of principles that attempt to explain one or more facts, relationships, or events; behavioral and social scientists often derive specific predictions (hypotheses) from theories that are then tested in research studies.

From Chapter 5 of *Statistics for the Behavioral & Social Sciences: A Brief Course*, Fifth Edition. Arthur Aron, Elliot J. Coups, Elaine N. Aron.

Many students find the most difficult part of the course to be mastering the basic logic of this chapter. This chapter requires some mental gymnastics. Even if you follow everything the first time through, you will be wise to review the chapter 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, a real-life behavioral and social science research study involves a sample of many individuals. However, to minimize how much you have to learn at one time, the examples in this chapter 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 prepare you to understand hypothesis testing as it is actually done in real research.

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 special 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 1. Notice that the mean is 14 months (Population $M = 14$), the standard deviation is 3 months (Population $SD = 3$), and the ages follow a normal curve. Based on the normal curve percentages, you can figure that fewer than 2% of babies start walking before 8 months of age; these are the babies who are 2 standard deviations or more below the mean. (This fictional distribution actually is close to the true distribution researchers have found for European babies, although that true distribution is slightly skewed to the right [Hindley, Filliozat, Klackenberg, Nicolet-Meister, & Sand, 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 its effect would be dramatically increased: babies taking the highly purified version should start walking much earlier than other babies. (We will assume that the purification process could not possibly

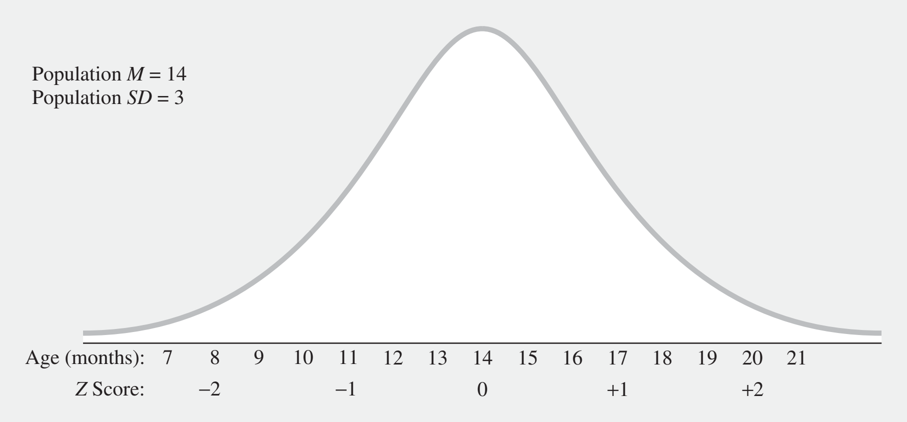


Figure 1 Distribution of when babies begin to walk (fictional data).

make the vitamin harmful.) However, refining the vitamin in this way is extremely expensive for each dose. Thus, the research team decides to try the procedure with just enough highly purified doses for only one baby. A newborn in the project is then randomly selected to take the highly purified version of the vitamin, and the researchers 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?

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

The Core Logic of Hypothesis Testing

There is a standard way researchers approach any hypothesis-testing problem. For this example, it works as follows. Consider first the population of babies in general (those who are not given the specially purified vitamin). In this population, 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, this would mean that the baby's walking age should be similar to that of babies that were not given the vitamin. In that case, it is highly unlikely (less than a 2% chance) that the particular baby we selected at random would start walking by 8 months. But what if the baby in our study does in fact start walking by 8 months? If that happened, we could *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 like that of a general population with a mean walking age of 14 months. We therefore conclude that babies given the specially purified vitamin will on the average start to walk before 14 months. Our explanation for the baby's early walking age in the study would be that the specially purified vitamin speeded up the baby's development.

In this example, the researchers first spelled out what would have to happen to conclude that the special purification procedure makes a difference. Having laid this out in advance, the researchers then conducted their study. Conducting the study in this case meant giving the specially purified vitamin to a randomly selected baby and watching to see how early that baby walked. We supposed that the result of the study is that the baby started walking before 8 months. The researchers then concluded that it is unlikely the specially purified vitamin makes *no* difference, and thus also that it *does* make a difference.

This kind of testing, with its opposite-of-what-you-predict, roundabout reasoning, is at the heart of inferential statistics in behavioral and social sciences. 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. The probabilities of babies walking at various ages are already known from studies of babies in general—that is, babies who have not received the specially purified vitamin. If the specially purified vitamin has no effect, then the ages at which babies start walking are the same with or without the specially purified vitamin. Thus, the distribution is that shown in Figure 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 not test hypotheses scientifically at all. In almost all behavioral and social sciences research, whether involving experiments, surveys, or whatever, we base our conclusions on the question, “What is the probability of getting our research results if the opposite of what we are predicting were true?” That is, we usually predict 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 finding allows us to reject the opposite prediction. If we reject the 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 the opposite prediction. If we are not able to reject the opposite prediction, we are not able to accept our prediction.

TIP FOR SUCCESS

This section, “The Core Logic of Hypothesis Testing,” is central to everything else you will learn in this course. So, we recommend that you read it a few times to ensure that you fully understand the logic of hypothesis testing.

The Hypothesis-Testing Process

Let’s look at our example, this time going over each step in some detail. Along the way, we cover the special terminology of hypothesis testing. Most important, we introduce the five steps of hypothesis testing.

Step ❶: 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 in general (that is, babies who do not take the specially purified vitamin).

Population 1 consists of babies who receive the experimental treatment (the specially purified vitamins). In our example, we use a sample of one baby to draw a conclusion about the age at which babies in Population 1 start to walk. Population 2 is a kind of comparison baseline of what is already known about babies in general.

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 (babies in general 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 in this example is that the mean of Population 1 is lower (babies receiving the special vitamin walk earlier) than the mean of Population 2.

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 (babies in general who do not take the specially purified vitamin). That is, the prediction is that there is no difference in the ages at which Population 1 and Population 2 babies start walking. On the average, 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

research hypothesis Statement in hypothesis testing about the predicted relation between populations (usually a prediction of a difference between population means).

is called a **null hypothesis.** It has this name because it states the situation in which there is no difference (the difference is "null") between the populations.[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 term 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 we can reject the null hypothesis so that we can decide about its alternative (the research hypothesis).

Step ❷: 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 about what the situation would be if the null hypothesis were true. In our example, we start out knowing the key information about Population 2, babies in the general population (see Figure 1): We know that it follows a normal curve, Population $M = 14$, and Population $SD = 3$. (We use the term "Population" before *M* and *SD* because we are referring to the mean and standard deviation of a population.) 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 and have a mean of 14 months and a standard deviation of 3 months.

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, we call this distribution a **comparison distribution.** (The comparison distribution is sometimes called a *sampling distribution*.) That is, in the hypothesis-testing process, you compare the actual sample's score to this comparison distribution.

In our vitamin example, the null hypothesis is that there is no difference in walking age between babies who take the specially purified vitamin (Population 1) and babies in general who 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. 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.

null hypothesis Statement about a relationship between populations that is the opposite of the research hypothesis; a statement that in the population there is no difference (or a difference opposite to that predicted) between populations; a contrived statement set up to examine whether it can be rejected as part of hypothesis testing.

comparison distribution Distribution used in hypothesis testing. It represents the population situation if the null hypothesis is true. It is the distribution to which you compare the score based on your sample's results.

Step ❸: 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 for it to

[1] We are oversimplifying a bit here to make the initial learning easier. The research hypothesis is that one population will walk earlier than the other. Thus, to be precise, its opposite is that the other group will either walk at the same time or walk 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. We discuss this issue in some detail later in the chapter.

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 by 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 (or *critical value*) 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 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. Now the researchers will reject the null hypothesis only if the actual sample *Z* score is −2.33 or lower—that is, if it falls in the shaded area in Figure 2. If the sample *Z* score falls above the shaded area in Figure 2, the researchers will *not* reject the null hypothesis.

cutoff sample score In hypothesis testing, the point on the comparison distribution at which, if reached or exceeded by the sample score, you reject the null hypothesis; also called *critical value.*

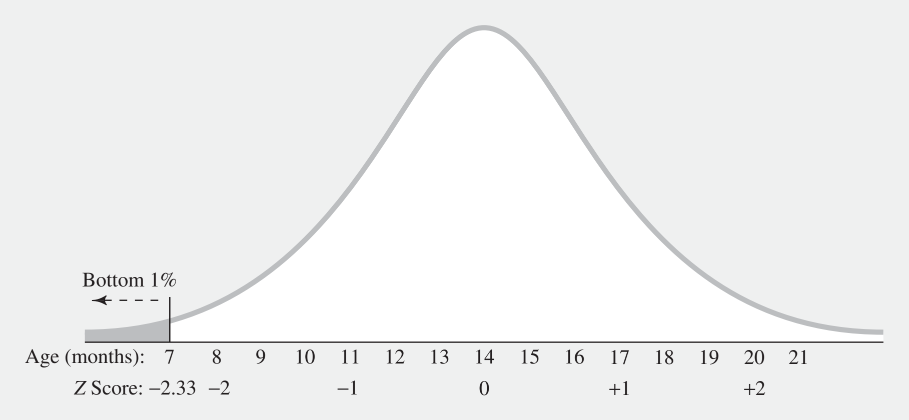


Figure 2 Distribution of when babies begin to walk, with bottom 1% shaded (fictional data).

In general, behavioral and social science 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. 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).

Step ❹: 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. You figure this *Z* score based on the population mean and standard deviation of the comparison distribution.

Let's 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, Population $M = 14$ and Population $SD = 3$. 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 $[Z = (6 - 14)/3]$. Figure 3 shows the score of our sample baby on the comparison distribution.

Step ❺: 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 ❹) to the cutoff *Z* score (from Step ❸). In our example, the actual sample's *Z* score 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.

Alternatively, 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 3. As you can see in the figure, the bottom 1% of the distribution is shaded. We recommend 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, farther out in the tail) than the cutoff sample score (critical value). If your actual sample *Z* score falls within the shaded region, you can reject the null hypothesis. Since the sample *Z* score (-2.67)

conventional levels of significance $(p < .05, p < .01)$ The levels of significance widely used in the behavioral and social sciences.

statistically significant Conclusion that the results of a study would be unlikely if in fact the sample studied represents a population that is no different from the population in general; an outcome of hypothesis testing in which the null hypothesis is rejected.

[2] In practice, since hypothesis testing is usually done using statistical software, you have to decide in advance only on the cutoff probability. The output of results usually includes the exact probability of getting your result if the null hypothesis were true. You then just compare the probability shown in the output 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.

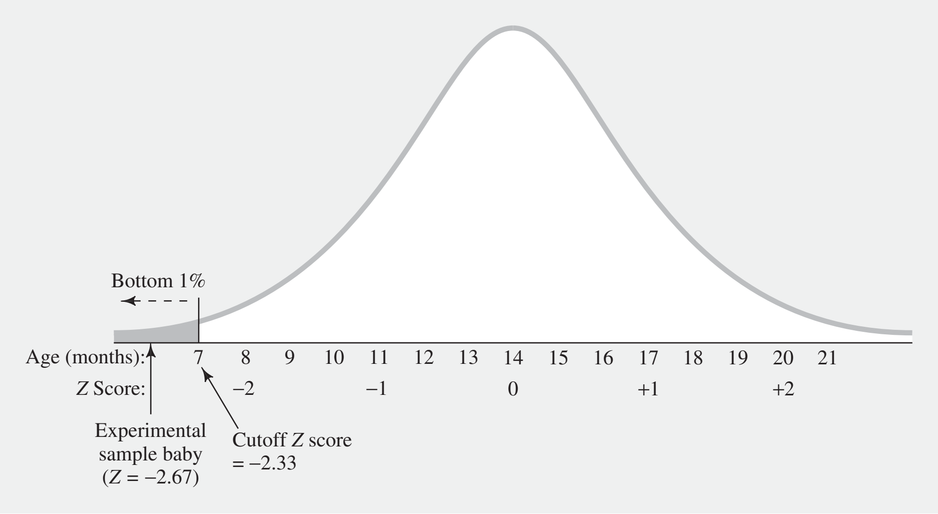


Figure 3 Distribution of when babies begin to walk, showing both the bottom 1% and the single baby that is the sample studied (fictional data).

in this example falls within the shaded 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 begin walking earlier than other babies in general.

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, when you reject the null hypothesis, all you are saying is that your results *support* the research hypothesis (as in our example). You would not go on to say that the results *prove* the research hypothesis or that the results show that the research hypothesis is *true.* Terms such as *prove* or *true* are 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 terms in logic and mathematics, but to use these words in conclusions from scientific research is inappropriate. (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.* You can also say that the results "support" or "provide evidence for" the research hypothesis.

Second, when a result is not extreme enough to reject the null hypothesis, you do not say that the result *supports* the null hypothesis. (And certainly you don't say the result proves the null hypothesis or shows the null hypothesis is true.) 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 who is 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 the particular study was able to detect. Therefore, when a result is not extreme enough to reject the null hypothesis, the results are said to be *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 in Box 1.) The core logic of hypothesis testing is summarized in Table 1, which also includes the logic for our example of a baby who is given a specially purified vitamin.

It is also important to bear in mind that just because a result is statistically significant, it does not necessarily mean that it is important or has practical or theoretical implications. As Frick (1995) puts it, statistical significance is about the strength of the evidence that we have a nonzero effect. Thus, statistical significance alone does not automatically indicate that an effect is of practical importance or that it advances a particular theory. Several researchers have observed that the word *significant* is the cause of confusion. Fidler, Cumming, Thomason, Pannuzzo, Smith et al. (2005) noted: "Confusion of clinical and statistical significance often manifests itself in ambiguous language: Researchers describe their results as *significant* or *nonsignificant* without distinguishing whether they are speaking statistically or substantively" (p. 137).

Table 1 The Basic Logic of Hypothesis Testing, Including the Logic for the Example of the Effect of a Specially Purified Vitamin on the Age That Babies Begin to Walk

	Basic Logic		**Baby Example**
Focus of Research	Sample is studied		Baby given specially purified vitamin and age of walking observed
Question	Is the sample typical of the general population?		Is this baby's walking age typical of babies in general?
Answer	Very unlikely	Could be	Very unlikely
	¶	¶	¶
Conclusion	The sample is probably not from the general population; it is probably from a different population.	Inconclusive	This baby is probably not from the general population of babies, because its walking age is much lower than for babies in general. Therefore, babies who take the specially purified vitamin will probably begin walking at an earlier age than babies in the general population.

Summary of the 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.

How are you doing?

1. A sample of rats in a laboratory is given an experimental treatment intended to make them learn a maze faster than other rats. 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 that 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 (who do 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?

7. How is it possible that a result can be statistically significant but be of little practical importance?

Answers

1. (a) Null hypothesis: The population of rats like those that get the experimental treatment score the same on the time to learn the maze as the population of rats that do not get the experimental treatment. (b) Research hypothesis: The population of rats like those that get the experimental treatment learn the maze faster than the population of rats in general that do not get the experimental treatment.

2. (a) A comparison distribution is 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, 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 cutoff sample *Z* score is the *Z* score at which, if the sample's *Z* score is more extreme than it on the comparison distribution, you reject the null hypothesis.

4. We say that hypothesis testing involves a double negative logic 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 when a result is so extreme that you reject the null hypothesis; the result is statistically significant. (b) The result is not statistically significant when a result is not very extreme; 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 2.50 is more extreme (that is, farther in the tail) than the cutoff *Z* score. Therefore, reject the null hypothesis; the research hypothesis is supported; the result is statistically significant.
7. Statistical significance is about whether it is likely there is a nonzero effect. It does not provide a test of whether such a nonzero effect is large enough to have important real-world implications.

One-Tailed and Two-Tailed Hypothesis Tests

In the baby-walking example, the researchers were interested in only one direction of result. Specifically, they tested whether babies given the specially purified vitamin would walk *earlier* than babies in general. The researchers in this study were not really interested in the possibility that giving the specially purified vitamins would cause babies to start walking *later*.

Directional Hypotheses and One-Tailed Tests

The baby-walking study is an example of testing a **directional hypothesis.** The study 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 taking the specially purified vitamin will make babies walk earlier. The null hypothesis, then, is that the specially purified vitamin will either have no effect or make babies walk later. Thus, in Figure 2, for the null hypothesis to be rejected, the sample has to have a score in one particular tail of the comparison distribution—the lower extreme or tail (in this example, the bottom 1%) of the comparison distribution. (When it comes to rejecting the null hypothesis with a directional hypothesis, a score at the other tail is the same as a score in the middle of the distribution; that is, such a score does 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 this example, the prediction was that the baby given the specially purified vitamin would start walking especially early—a prediction in the direction of a low score on months before walking. Thus, the cutoff region was at the low end (left side) of the comparison distribution. In other research situations with a directional hypothesis, the cutoff may be at the high end (right side) of the comparison distribution. That is, in these situations, the researchers would be predicting that the experimental procedure will produce a high score.

Nondirectional Hypotheses and Two-Tailed Tests

Sometimes, a research hypothesis is that an experimental procedure will have an effect, without saying whether it will produce a very high score or a very low score. Suppose a researcher is interested in whether a new social skills program will affect worker productivity. The program could improve productivity by making the working

directional hypothesis Research hypothesis predicting a particular direction of difference between populations—for example, a prediction that the population like the sample studied has a higher mean than the population in general.

one-tailed test Hypothesis-testing procedure for a directional hypothesis; situation in which the region of the comparison distribution in which the null hypothesis would be rejected is all on one side (or tail) of the distribution.

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.

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

Determining Cutoff Scores with Two-Tailed Tests. There is a special complication in a two-tailed test. You have to divide the significance percentage between the two tails. For example, with a 5% significance level, you would reject a null hypothesis only if the sample was so extreme that it was in either the top 2.5% or the bottom 2.5% of the comparison distribution. 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 is not so extreme: only +1.64 or −1.64. But with a one-tailed test, only one side of the distribution is considered. These situations are shown in Figure 4a.

Using the 1% significance level, a two-tailed test (0.5% at each tail) has cutoffs of +2.58 and −2.58. With a one-tailed test, the cutoff is either +2.33 or −2.33. These situations are shown in Figure 4b. The *Z* score cutoffs for one-tailed and two-tailed tests for the .05 and .01 significance levels are also summarized in Table 2.

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 compared to what would be needed with a two-tailed test. Yet there is a price. 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. You plan to use a two-tailed test when you have a clearly nondirectional hypothesis. In practice, the decision 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, the opposite result may be more interesting than what you had predicted. By using one-tailed tests, we risk having to ignore possibly important results.

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. 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 a two-tailed test will 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 behavioral and social sciences research articles, unless the researcher specifically states that a one-tailed test was used, it is assumed that the test was two-tailed.

In practice, however, our experience is that most research results are either so extreme that they will be significant whether you use a one-tailed or two-tailed test, or they are so far from extreme that they would not be significant no matter what you use.

nondirectional hypothesis Research hypothesis that does not predict a particular direction of difference between the population like the sample studied and the population in general.

two-tailed test Hypothesis-testing procedure for a nondirectional hypothesis; the situation in which the region of the comparison distribution in which the null hypothesis would be rejected is divided between the two sides (tails) of the distribution.

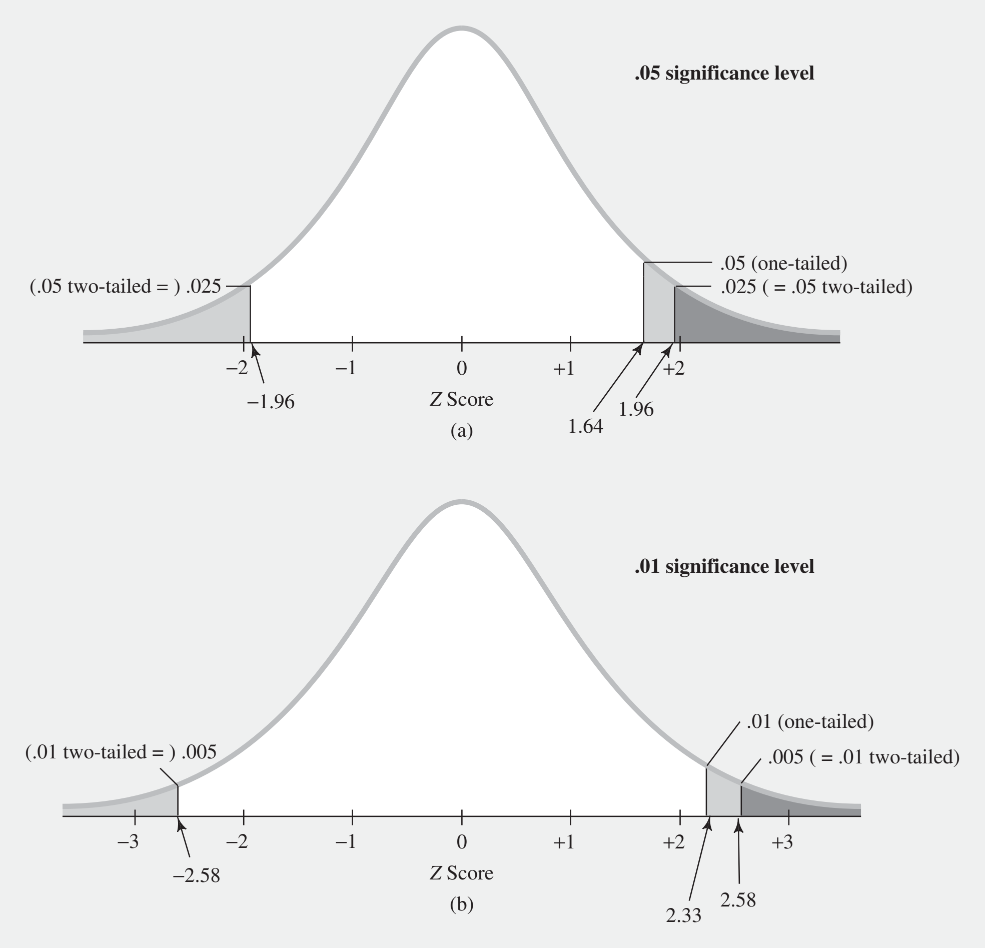


Figure 4 Significance level cutoffs for one-tailed and two-tailed tests: (a) .05 significance level; (b) .01 significance level. (The one-tailed tests in these examples assume the prediction was for a high score. You could instead have a one-tailed test where the prediction is for a lower score, which would be on the left tail of the distribution.)

But what happens when a result is less certain? The researcher's decision about one- 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—not a researcher's decisions—determine the conclusion as much as possible. Furthermore, whenever a result is less than completely clear one way

Table 2 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

or the other, most researchers are not comfortable drawing strong conclusions until more research is done.

An Example of Hypothesis Testing with a Two-Tailed Test

Here is another made-up example, this time using a two-tailed test. A researcher is interested in the effect of going through a natural disaster on the attitude of police chiefs about the goodness of the people in their city. The researcher believes that after a disaster, the police chief is likely to have a more positive attitude about the people of the city (because the chief will have seen many acts of heroism and helping of neighbors after the event). However, it is also possible that a disaster will lead to police chiefs having more negative attitudes, because there may be looting and other dishonest behavior after the disaster. Thus, because there could be an effect either way, the researcher will make a nondirectional hypothesis.

Let us assume that there is lots of previous research on the attitudes of police chiefs about the goodness of the people in their cities. And also let's assume that this previous research shows that on a standard questionnaire, the mean attitude rating is 69.5 with a standard deviation of 14.1, and the attitude scores follow a normal curve. Finally, let's assume that a major earthquake has just occurred in an isolated city, and shortly afterward the researcher is able to give the standard questionnaire to the police chief of that city and that chief's score is 35. Remember that in this chapter we are considering the special situation in which the sample is a single individual. The researcher then carries out the five steps of hypothesis testing.

BOX 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 that sometimes it may be true that one thing has so little effect on another that it probably represented no real, or at least no important, relationship or difference. The problem is knowing when to conclude that. Frick 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 sensitive measurement, 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. But subjective judgments are a part of science. For example, reviewers of articles submitted for publication in the top scientific journals have to decide if a topic is important enough to compete for limited space in those journals. Furthermore, the null hypothesis is being accepted all the time anyway. (For example, many behavioral and social scientists 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. Later in this chapter, you will learn about decision errors that can be made during the process of significance testing. As you will see, such errors are also relevant to the issue of when to accept the null hypothesis.

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

Population 1: Police chiefs whose city has just been through a disaster.
Population 2: Police chiefs in general.

The research hypothesis is that when measured on their attitude toward the goodness of the people of their city, police chiefs whose city has just been through a disaster (Population 1) score differently from police chiefs in general (Population 2). The opposite of the research hypothesis, the null hypothesis, is this: Police chiefs whose city has just been through a disaster have the same attitude as police chiefs in general. That is, the null hypothesis is that the attitudes of Populations 1 and 2 are the same.

TIP FOR SUCCESS

Remember that the research hypothesis and null hypothesis must always be *complete opposites.* Researchers specify the research hypothesis, and this determines what the null hypothesis that goes with it will be.

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, so we can use it as our comparison distribution. As noted, it follows a normal curve, with Population $M = 69.5$ and Population $SD = 14.1$.

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** The researcher selects the 5% significance level. The researcher has made a nondirectional hypothesis and will therefore use a two-tailed test. Thus, the researcher will reject the null hypothesis only if the police chief's attitude 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 4a and Table 2).

TIP FOR SUCCESS

When carrying out the five steps of hypothesis testing, always draw a figure like Figure 5. Be sure to include the cutoff score(s) and shade the appropriate tail(s). If the sample score falls within a shaded tail region, you reject the null hypothesis and the result is statistically significant. If the sample score does not fall within a shaded tail region, you do not reject the null hypothesis.

❹ **Determine your sample's score on the comparison distribution.** The police chief whose city went through the earthquake took the standard attitude questionnaire and had a score of 35. This corresponds to a *Z* score on the comparison distribution of −2.45. That is, $Z = (X - M)/SD = (35 - 69.5)/14.1 = -2.45$.

❺ **Decide whether to reject the null hypothesis.** A *Z* score of −2.02 is more extreme than the *Z* score of −1.96, which is where the lower 2.5% of the comparison distribution begins. Notice in Figure 5 that the *Z* score of −2.45

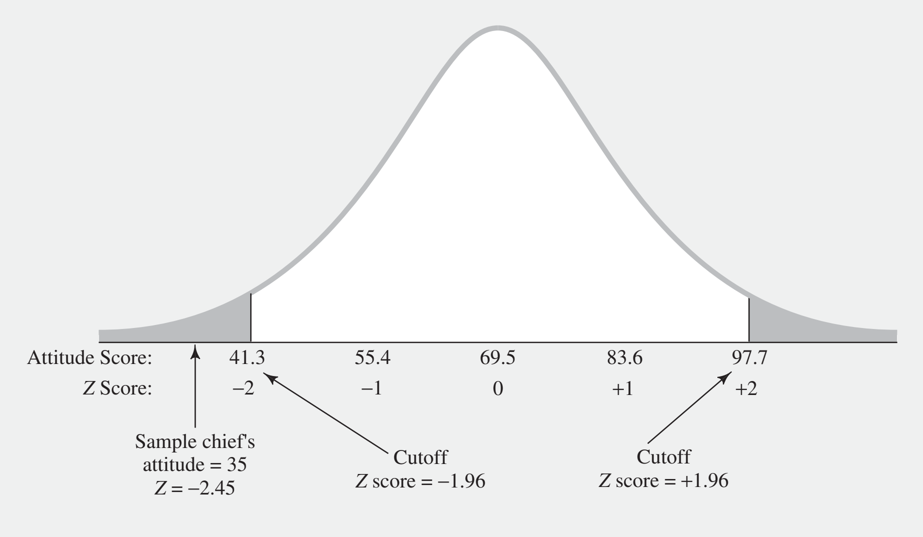


Figure 5 Distribution of attitudes of police chiefs toward goodness of people in their cities with upper and lower 2.5% shaded and showing the sample police chief whose city has just been through a disaster (fictional data).

falls within the shaded area in the left tail of the comparison distribution. This *Z* score of −2.45 is a result so extreme that it is unlikely to have occurred if this police chief were from a population no different than Population 2. Therefore, the researcher rejects the null hypothesis. The result is statistically significant. It supports the research hypothesis that going through a disaster does indeed change police chiefs' attitudes toward their cities. In this case, the results would mean that the effect is one of making chiefs less positive about their cities' people. (Remember, however, that this is a fictional study.)

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 people hungry will affect how they do on a coordination test. A randomly selected person is asked 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 and gender, 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 evenly 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 supports 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. You might use a two-tailed test even when your theory predicts a particular direction of result because 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 kind 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 a situation in which the *right procedures* lead to the *wrong decisions.*

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: it 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.

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 did 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, you will still sometimes make a Type I error (to be precise, 5% or 1% of the time). Consider again the example of giving a baby a new purified vitamin and examining the age at which the baby starts walking. Suppose the special new vitamin in reality has no effect whatsoever on the age at which babies start walking. However, in randomly picking a sample of one baby to study, the researchers might just happen to pick a baby who would have started walking at a very young age, regardless of whether it was given the new vitamin. Randomly selecting a sample baby like this is *unlikely.* But such extreme samples are *possible.* Should this happen, the researchers would reject the null hypothesis and conclude that the new vitamin does make a difference. 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 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 behavioral and social scientists. It is of serious concern because they 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.

decision error Incorrect conclusion in hypothesis testing in relation to the real (but unknown) situation, such as deciding the null hypothesis is false when it is really true.

Type I error Rejecting the null hypothesis when in fact it is true; getting a statistically significant result when in fact the research hypothesis is not true.

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 we 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. (You will sometimes see the significance level, the chance of making a Type I error, referred to as *alpha,* the Greek letter α.)

Again, the significance level is the same as the chance of making a Type I error. Thus, the lower probability we set for the significance level, the smaller the chance of a Type I error. Researchers who do not want to take a lot of risk set the significance level lower than .05, such as $p < .001$. In this way the result of a study has to be very extreme for the hypothesis-testing process to reject the null hypothesis.

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 extreme significance level, such as $p < .001$, you run a different kind of risk. With a very extreme 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 fact 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.**

Consider again our example of the new purified vitamin and the age at which babies begin walking. Suppose that, in truth, the new vitamin does cause babies to begin walking at an earlier age than normal. However, in conducting your study, the results for the sample baby are not strong enough to allow you to reject the null hypothesis. Perhaps the random sample baby that you selected to try out the new vitamin happened to be a baby who would not respond to the new vitamin. 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.

Type II errors especially concern behavioral and social scientists interested in practical applications. This is 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 chance 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 results in only a very small effect, the results have a good chance of being significant. There is a cost to this insurance policy too.

Type II error Failing to reject the null hypothesis when in fact it is false; failing to get a statistically significant result when in fact the research hypothesis is true.

Relationship Between 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 an extreme significance level like .001, even if the research hypothesis is true, the results have to be quite strong for you to reject the null hypothesis.) The insurance policy against Type II error (setting a significance level of, say, .20) has the cost that you increase 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 tradeoff 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 possible correct and mistaken conclusions in hypothesis testing is shown in Table 3. 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 2 shows that there are two ways to be correct and two ways to be in error in any hypothesis-testing situation.

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.

Table 3 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)	Correct decision
	Study is inconclusive (do not reject null hypothesis)	Correct decision	Error (Type II)

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?
3. (a) What is a Type II error? (b) Why is it possible?
4. If you set a lenient significance level (say, .25), what is the effect on the probability of (a) Type I error and (b) Type II error?
5. If you set an extreme significance 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) A Type I error is rejecting the null hypothesis (and thus supporting the research hypothesis) when the null hypothesis is actually true (and the research hypothesis false). (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, even though it is unlikely, it is still possible that the null hypothesis is true. (c) The probability of a Type I error is the significance level (such as .05).
3. (a) A Type II error is 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). (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.
4. (a) The probability is high; (b) the probability is low.
5. (a) The probability is low; (b) the probability is high.

Hypothesis Tests as Reported in Research Articles

In general, hypothesis testing is reported in research articles using one of the specific methods of hypothesis testing you learn in later chapters. For each result of interest, the researcher usually first says whether the result was statistically significant. Next, the researcher usually gives the symbol for the specific method used in figuring the probabilities, such as *t, F,* or χ^2. Finally, there will be an indication of the significance level, such as $p < .05$ or $p < .01$. (The researcher will usually also provide much other information, such as sample means and standard deviations.) For example, Gentile (2009) conducted a survey study of video-game use in a random sample of 1,178 American youth aged 8 to 18 years. The vast majority (88%) of the participants reported playing video games at least occasionally. The survey included an 11-item scale of symptoms of pathological use of video games (e.g., "Have you ever done poorly on a school assignment or test because you spent too much time playing video games?"). Gentile was interested in potential differences between boys and girls in these symptoms. Here is what he reported: "Boys exhibited a greater number of symptoms $(M = 2.8, SD = 2.2)$ than girls $(M = 1.3, SD = 1.7)$, $t\ (1175) = 12.5$, $p < .001$. . . . Note that the average number of symptoms reported was not high for either group." 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 their groups represent) were true is less than .001 (.1%). Thus it is very highly unlikely in the population of American youth aged 8 to 18 years that boys and girls do not differ in their number of symptoms of pathological use of video games. Also, since it was not specified as a one-tailed test, you can assume this was a two-tailed test.

When a result is close, but does not reach the significance level chosen, it may be reported 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. Regardless of whether a result is significant, it is increasingly common for researchers to report the exact *p* level—such as $p = .03$ or $p = .27$ (these are given in computer outputs of statistical tests). In addition, if a one-tailed test was used, that usually will be noted. Again, when reading research articles in most areas of the behavioral and social sciences, assume a two-tailed test 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 will likely 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 "$p < .001$."

Finally, often the outcomes of hypothesis testing are shown simply as asterisks in a table of results. In such tables, a result with an asterisk is significant, while a result without one is not.

In reporting results of significance testing, researchers rarely make explicit the research hypothesis or the null hypothesis, or describe any of the other steps of the process in any detail. It is assumed that the reader understands all of this very well. Decision errors are rarely mentioned in research articles.

Learning Aids

1. Hypothesis testing considers the probability that the result of a study could have come about even if the experimental procedure had no effect. If this probability is low, the scenario of no effect is rejected and the hypothesis behind the experimental procedure is supported.
2. The expectation of an effect is the research hypothesis, and the hypothetical situation of no effect is the null hypothesis.
3. When a result (that is, a sample score) is so extreme that the result would be very unlikely if the null hypothesis were true, the researcher rejects the null hypothesis and describes the research hypothesis as supported. If the result is not that extreme, the researcher does not reject the null hypothesis, and the study is inconclusive.
4. Behavioral and social scientists usually consider a result too extreme if it is less likely than 5% (that is, a significance level of $p < .05$) to have come about if the null hypothesis were true. Sometimes a more extreme 1% ($p < .01$ significance level), or even .1% ($p < .001$ significance level), cutoff is used.
5. The cutoff percentage is the probability of the result being extreme in a predicted direction in a directional or one-tailed test. The cutoff percentages are the probability of the result being extreme in either direction in a nondirectional or two-tailed test.
6. The five steps of hypothesis testing are:
   - 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.**

7. There are two kinds of decision errors one can make in hypothesis testing. A Type I error is when a researcher rejects the null hypothesis, but the null hypothesis is actually true. A Type II error is when a researcher does not reject the null hypothesis, but the null hypothesis is actually false.
8. Research articles typically report the results of hypothesis testing by saying a result was or was not significant and giving the probability level cutoff (usually 5% or 1%) on which the decision was based. Research articles rarely mention decision errors.

Key Terms

hypothesis testing
hypothesis
theory
research hypothesis
null hypothesis
comparison distribution
cutoff sample score
conventional levels of significance ($p < .05, p < .01$)
statistically significant
directional hypothesis
one-tailed test
nondirectional hypothesis
two-tailed test
decision errors
Type I error
Type II error

After going through an experimental treatment, a randomly selected individual has a score of 27 on a particular measure. The scores of people in general on this measure are normally distributed with a mean of 19 and a standard deviation of 4. The researcher predicts an effect, but does not predict a particular direction of effect. Using the 5% significance level, what should you conclude? Solve this problem explicitly using all five steps of hypothesis testing and illustrate your answer with a sketch showing the comparison distribution, the cutoff (or cutoffs), and the score of the sample on this distribution.

Answer

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

Population 1: People who go through the experimental treatment.
Population 2: People in general (that is, people who do not go through the experimental treatment).

The research hypothesis is that Population 1 will score differently than Population 2 on the particular measure. The null hypothesis is that the two populations are not different on the measure.

❷ **Determine the characteristics of the comparison distribution:** Population $M = 19$, Population $SD = 4$, normally distributed.

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** For a two-tailed test at the 5% level (2.5% at each tail), the cutoff sample scores are $+1.96$ and -1.96 (see Figure 4 or Table 2).

❹ **Determine your sample's score on the comparison distribution.** $Z = (27 - 19)/4 = 2$.

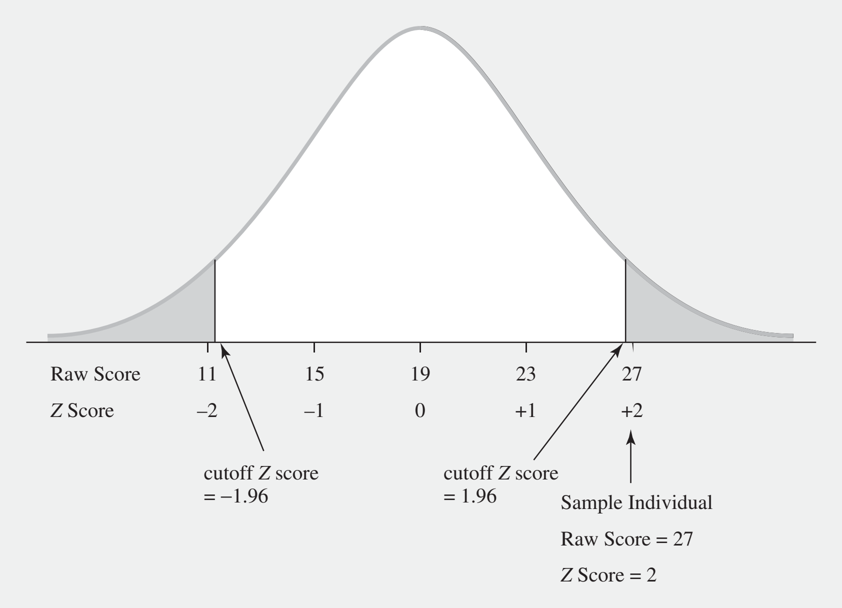


Figure 6 Diagram for Example Worked-Out Problem showing comparison distribution, cutoffs (2.5% shaded area in each tail), and sample score.

❺ **Decide whether to reject the null hypothesis.** A *Z* score of 2 is more extreme than the cutoff *Z* of +1.96. Reject the null hypothesis; the result is significant. The experimental treatment affects scores on this measure. The diagram is shown in Figure 6.

Outline for Writing Essays for Hypothesis-Testing Problems Involving a Sample of One Participant and a Known Population

1. Describe the core logic of hypothesis testing. Be sure to explain terminology such as research hypothesis and null hypothesis, and explain the concept of providing support for the research hypothesis when the study results are strong enough to reject the null hypothesis.
2. Explain the concept of the comparison distribution. Be sure to mention that it is the distribution that represents the population situation if the null hypothesis is true. Note that the key characteristics of the comparison distribution are its mean, standard deviation, and shape.
3. Describe the logic and process for determining (using the normal curve) the cutoff sample scores on the comparison distribution at which you should reject the null hypothesis.
4. Describe how to figure the sample's score on the comparison distribution.
5. Explain how and why the scores from Steps ❸ and ❹ of the hypothesis-testing process are compared. Explain the meaning of the result of this comparison with regard to the specific research and null hypotheses being tested.

Practice Problems

These problems involve figuring. Most real-life statistics problems are done on a computer with special statistical software. Even if you have such software, do these problems by hand to ingrain the method in your mind.

All data are fictional unless an actual citation is given.

Set I (for answers, see the end of this chapter)

1. Define the following terms in your own words: (a) hypothesis-testing procedure, (b) .05 significance level, and (c) two-tailed test.
2. When a result is not extreme enough to reject the null hypothesis, explain why it is wrong to conclude that your result supports the null hypothesis.
3. For each of the following, (a) say which two populations are being compared, (b) state the research hypothesis, (c) state the null hypothesis, and (d) say whether you should use a one-tailed or two-tailed test and why.
   i. Do Canadian children whose parents are librarians score higher than Canadian children in general on reading ability?
   ii. Is the level of income for residents of a particular city different from the level of income for people in the region?
   iii. Do people who have experienced an earthquake have more or less self-confidence than the general population?
4. Based on the information given for each of the following studies, decide whether to reject the null hypothesis. For each, give (a) the *Z*-score cutoff (or cutoffs) on the comparison distribution at which the null hypothesis should be rejected, (b) the *Z* score on the comparison distribution for the sample score, and (c) your conclusion. Assume that all populations are normally distributed.

	Population				
Study	*M*	*SD*	*Sample Score*	*p*	*Tails of Test*
A	10	2	14	.05	1 (high predicted)
B	10	2	14	.05	2
C	10	2	14	.01	1 (high predicted)
D	10	2	14	.01	2
E	10	4	14	.05	1 (high predicted)

5. Based on the information given for each of the following studies, decide whether to reject the null hypothesis. For each, give (a) the *Z*-score cutoff (or cutoffs) on the comparison distribution at which the null hypothesis should be rejected, (b) the *Z* score on the comparison distribution for the sample score, and (c) your conclusion. Assume that all populations are normally distributed.

	Population				
Study	*M*	*SD*	*Sample Score*	*p*	*Tails of Test*
A	70	4	74	.05	1 (high predicted)
B	70	1	74	.01	2
C	70	2	76	.01	2
D	70	2	77	.01	2
E	70	2	68	.05	1 (high predicted)

6. A researcher studying the senses of taste and smell has carried out many studies in which students are given each of 20 different foods (apricot, chocolate, cherry, coffee, garlic, etc.). She administers each food by dropping a liquid on the tongue. Based on her past research, she knows that for students overall at the university, the mean number of the 20 foods that students can identify correctly is 14, with a standard deviation of 4, and the distribution of scores follows a normal curve. The researcher wants to know whether people's accuracy on this task has more to do with smell than with taste. In other words, she wants to test whether people do *worse* on the task when they are only able to taste the liquid compared to when they can both taste and smell it (note that this is a directional hypothesis). Thus, she sets up special procedures that keep a person from being able to use the sense of smell during the task. The researcher then tries the procedure on one randomly selected student. This student is able to identify only five correctly. (a) Using the .05 significance level, what should the researcher conclude? Solve this problem explicitly using all five steps of hypothesis testing and illustrate your answer with a sketch showing the comparison distribution, the cutoff (or cutoffs), and the score of the sample on this distribution. (b) Explain your answer to someone who has never had a course in statistics (but who is familiar with mean, standard deviation, and *Z* scores).
7. A nursing researcher is working with people who have had a particular type of major surgery. This researcher proposes that people will recover from the operation more quickly if friends and family are in the room with them for the first 48 hours after the operation. It is known that time to recover from this kind of surgery is normally distributed with a mean of 12 days and a standard deviation of 5 days. The procedure of having friends and family in the room for the period after the surgery is tried with a randomly selected patient. This patient recovers in 4 days. (a) Using the .01 significance level, what should the researcher conclude? Solve this problem explicitly using all five steps of hypothesis testing and illustrate your answer with a sketch showing the comparison distribution, the cutoff (or cutoffs), and the score of the sample on this distribution. (b) Explain your answer to someone who has never had a course in statistics (but who is familiar with mean, standard deviation, and *Z* scores).
8. Robins and John (1997) carried out a study on narcissism (extreme self-love), comparing people who scored high versus low on a narcissism questionnaire. (An example item was "If I ruled the world, it would be a better place.") They also had other questionnaires, including one that had an item about how many times the participant looked in the mirror on a typical day. In their results section, the researchers noted ". . . as predicted, high-narcissism individuals reported looking at themselves in the mirror more frequently than did low narcissism individuals ($Ms = 5.7$ vs. 4.8), . . . $p < .05$" (p. 39). Explain this result to a person who has never had a course in statistics. (Focus on the meaning of this result in terms of the general logic of hypothesis testing and statistical significance.)
9. Reber and Kotovsky (1997), in a study of problem solving, described one of their results comparing a specific group of participants within their overall control condition as follows: "This group took an average of 179 moves to solve the puzzle, whereas the rest of the control participants took an average of 74 moves, $t(19) = 3.31$, $p < .01$" (p. 183). Explain this result to a person who has never had a course in statistics. (Focus on the meaning of this result in terms of the general logic of hypothesis testing and statistical significance.)
10. For each of the following studies, make a chart of the four possible correct and incorrect decisions, and explain what each would mean. Each chart should be

laid out like Table 3, but you should put into the boxes the possible results, using the names of the variables involved in the study.

(a) A study of whether increasing the amount of recess time improves schoolchildren's in-class behavior.

(b) A study of whether color-blind individuals can distinguish gray shades better than the population at large.

(c) A study comparing individuals who have ever been in psychotherapy to the general public to see if they are more tolerant of other people's upsets than is the general population.

11. You conduct a research study. How can you know (a) if you have made a Type I error? and (b) if you have made a Type II error?

Set II

12. List the five steps of hypothesis testing and explain the procedure and logic of each.
13. When a result is significant, explain why it is wrong to say the result "proves" the research hypothesis.
14. For each of the following, (a) state which two populations are being compared, (b) state the research hypothesis, (c) state the null hypothesis, and (d) say whether you should use a one-tailed or two-tailed test and why.

    i. In an experiment, people are told to solve a problem by focusing on the details. Is the speed of solving the problem different for people who get such instructions compared to the speed for people who are given no special instructions?

    ii. Based on anthropological reports in which the status of women is scored on a 10-point scale, the mean and standard deviation across many cultures are known. A new culture is found in which there is an unusual family arrangement. The status of women is also rated in this culture. Do cultures with the unusual family arrangement provide higher status to women than cultures in general?

    iii. Do people who live in big cities develop more stress-related conditions than people in general?

15. Based on the information given for each of the following studies, decide whether to reject the null hypothesis. For each, give (a) the *Z*-score cutoff (or cutoffs) on the comparison distribution at which the null hypothesis should be rejected, (b) the *Z* score on the comparison distribution for the sample score, and (c) your conclusion. Assume that all populations are normally distributed.

	Population				
Study	*M*	*SD*	*Sample Score*	*p*	*Tails of Test*
A	5	1	7	.05	1 (high predicted)
B	5	1	7	.05	2
C	5	1	7	.01	1 (high predicted)
D	5	1	7	.01	2

16. Based on the information given for each of the following studies, decide whether to reject the null hypothesis. For each, give (a) the *Z*-score cutoff (or cutoffs) on the comparison distribution at which the null hypothesis should be rejected, (b) the *Z* score on the comparison distribution for the sample score, and (c) your conclusion. Assume that all populations are normally distributed.

	Population				
Study	*M*	*SD*	*Sample Score*	*p*	*Tails of Test*
A	100.0	10.0	80	.05	1 (low predicted)
B	100.0	20.0	80	.01	2
C	74.3	11.8	80	.01	2
D	76.9	1.2	80	.05	1 (low predicted)
E	88.1	12.7	80	.05	2

17. A researcher wants to test whether a certain sound will make rats do worse on learning tasks. It is known that an ordinary rat can learn to run a particular maze correctly in 18 trials, with a standard deviation of 6. (The number of trials to learn this maze is normally distributed. More trials mean worse performance on the task.) The researcher now tries an ordinary rat in the maze, but with the sound. The rat takes 38 trials to learn the maze. (a) Using the .05 level, what should the researcher conclude? Solve this problem explicitly using all five steps of hypothesis testing and illustrate your answer with a sketch showing the comparison distribution, the cutoff (or cutoffs), and the score of the sample on this distribution. (b) Then explain your answer to someone who has never had a course in statistics (but who is familiar with mean, standard deviation, and *Z* scores).
18. A researcher developed an elaborate training program to reduce the stress of childless men who marry women with adolescent children. It is known from previous research that such men, one month after moving in with their new wife and her children, have a stress level of 85 with a standard deviation of 15, and the stress levels are normally distributed. The training program is tried on one man randomly selected from all those in a particular city who during the preceding month have married a woman with an adolescent child. After the training program, this man's stress level is 60. (a) Using the .05 level, what should the researcher conclude? Solve this problem explicitly using all five steps of hypothesis testing and illustrate your answer with a sketch showing the comparison distribution, the cutoff (or cutoffs), and the score of the sample on this distribution. (b) Explain your answer to someone who has never had a course in statistics (but who is familiar with mean, standard deviation, and *Z* scores).
19. A researcher predicts that listening to classical music while solving math problems will make a particular brain area more active. To test this, a research participant has her brain scanned while listening to classical music and solving math problems, and the brain area of interest has a percent signal change of 5.8. From many previous studies with this same math-problems procedure (but not listening to music), it is known that the signal change in this brain area is normally distributed with a mean of 3.5 and a standard deviation of 1.0. Using the .01 level, what should the researcher conclude? (a) Solve this problem explicitly using all five steps of hypothesis testing and illustrate your answer with a sketch showing the comparison distribution, the cutoff (or cutoffs), and the score of the sample on this distribution. (b) Explain your answer to someone who has never had a course in statistics (but who is familiar with mean, standard deviation, and *Z* scores).
20. Earlier in the chapter, we described the results of a study conducted by Gentile (2009), in which American youth completed a survey about their use of video games. In his results section, Gentile reported, "[T]here was a sizable difference between boys' average playing time [of video games] $(M = 16.4$ hr/week, $SD = 14.1)$ and girls' average playing time $(M = 9.2$ hr/week, $SD = 10.2)$, $t(1034) = 9.2$, $p < .001 \ldots$" Explain this result to a person who has never had a course in

statistics. (Focus on the meaning of this result in terms of the general logic of hypothesis testing and statistical significance.)

21. In an article about anti-tobacco campaigns, Siegel and Biener (1997) discuss the results of a survey of tobacco usage and attitudes conducted in Massachusetts in 1993 and 1995; Table 4 shows the results of this survey. Focusing on just the first line (the percentage smoking $>$ 25 cigarettes daily), explain what this result means to a person who has never had a course in statistics. (Focus on the meaning of this result in terms of the general logic of hypothesis testing and statistical significance.)
22. For each of the following studies, make a chart of the four possible correct and incorrect decisions, and explain what each would mean. Each chart should be laid out like Table 3, but put into the boxes the possible results, using the names of the variables involved in the study.

    (a) A study of whether infants born prematurely begin to recognize faces later than do infants in general.
    (b) A study of whether high school students who receive an AIDS prevention program in their schools are more likely to practice safe sex than are other high school students.
    (c) A study of whether memory for abstract ideas is reduced if the information is presented in distracting colors.

Table 4 Selected Indicators of Change in Tobacco Use, ETS Exposure, and Public Attitudes toward Tobacco Control Policies—Massachusetts, 1993–1995

	1993	1995
Adult Smoking Behavior		
Percentage smoking >25 cigarettes daily	24	10*
Percentage smoking <15 cigarettes daily	31	49*
Percentage smoking within 30 minutes of waking	54	41
Environmental Tobacco Smoke Exposure		
Percentage of workers reporting a smoke free worksite	53	65*
Mean hours of ETS exposure at work during prior week	4.2	2.3*
Percentage of homes in which smoking is banned	41	51*
Attitudes Toward Tobacco Control Policies		
Percentage supporting further increase in tax on tobacco with funds earmarked for tobacco control	78	81
Percentage believing ETS is harmful	90	84
Percentage supporting ban on vending machines	54	64*
Percentage supporting ban on support of sports and cultural events by tobacco companies	59	53*

*$p < .05$

Source: Siegel, M., & Biener, L. (1997). Evaluating the impact of statewide anti-tobacco campaigns: The Massachusetts and California tobacco control programs. *Journal of Social Issues, 53,* 147–168. Copyright © 2002, 1997 The Society for the Psychological Study of Social Issues. Reprinted by permission of Blackwell Publishers Journals.

Answers to Set I Practice Problems

1. (a) The logical, statistical procedure for determining the likelihood of your study having gotten a particular pattern of results if the null hypothesis is true. (b) The situation in hypothesis testing in which you decide to reject the null hypothesis because the probability of getting your particular results if the null hypothesis were true is less than 5%. (c) A procedure used in hypothesis testing when the research hypothesis does not specify a particular direction of difference—it tests for extreme results that are either higher or lower than would be expected by chance.
2. It is possible that the research hypothesis is correct but the result in the particular sample was not extreme enough to be able to reject the null hypothesis.
3. (i) (a) Population 1: Canadian children of librarians; Population 2: All Canadian children. (b) Population 1 children have a higher average reading ability than Population 2 children. (c) Population 1's average reading ability is not higher than Population 2's. (d) One-tailed, because the question is whether they "score higher," so only one direction of difference is of interest.
   (ii) (a) Population 1: People who live in a particular city; Population 2: All people who live in the region. (b) Populations 1 and 2 have different mean incomes. (c) Populations 1 and 2 have the same mean income. (d) Two-tailed, because the question is whether the income of the people in the city is "different" from those in the region as a whole, so a difference in either direction would be of interest.
   (iii) (a) Population 1: People who have experienced an earthquake; Population 2: People in general. (b) Populations 1 and 2 have different mean levels of self-confidence. (c) Populations 1 and 2 have the same mean level of self-confidence. (d) Two-tailed, because the question specifies "more or less," so a difference in either direction would be of interest.
4.

Study	Cutoff	*Z* Score on Comparison Distribution	Decision
A	+1.64	2	Reject null hypothesis
B	±1.96	2	Reject null hypothesis
C	+2.33	2	Inconclusive
D	±2.57	2	Inconclusive
E	+1.64	1	Inconclusive

5.

Study	Cutoff	*Z* Score on the Comparison Distribution	Decision
A	+1.64	1	Inconclusive
B	±2.57	4	Reject null hypothesis
C	±2.57	3	Reject null hypothesis
D	±2.57	3.5	Reject null hypothesis
E	+1.64	−1	Inconclusive

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

Population 1: Students who are prevented from using their sense of smell.
Population 2: Students in general.

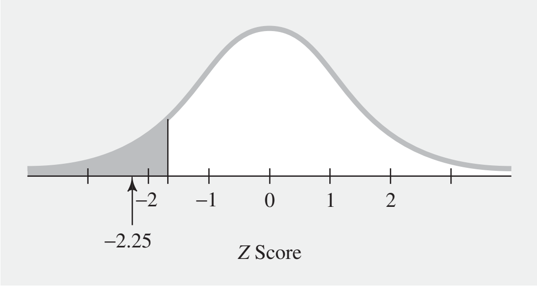
The research hypothesis is that students prevented from using their sense of smell (Population 1) will do worse on the taste task than students in general (Population 2). The null hypothesis is that students prevented from using their sense of smell (Population 1) will not do worse on the taste task than students in general (Population 2).
❷ **Determine the characteristics of the comparison distribution.** The comparison distribution will be the same as Population 2. As stated in the problem, Population $M = 14$ and Population $SD = 4$. We assume it follows a normal curve.
❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** For a one-tailed test at the .05 level, the cutoff is -1.64. (The cutoff is a *negative value* because the research hypothesis is that Population 1 will *do worse* on the task than Population 2—that is, Population 1 will have a *lower score* on the task than Population 2.)
❹ **Determine your sample's score on the comparison distribution.** The sample's score was 5. $Z = (5 - 14)/4 = -2.25$.
❺ **Decide whether to reject the null hypothesis.** A *Z* score of -2.25 is more extreme than the cutoff of -1.64. Thus, you can reject the null hypothesis. The research hypothesis is supported—not having a sense of smell makes for fewer correct identifications.
(b) In brief, you solve this problem by considering the likelihood that being without a sense of smell makes no difference. If the sense of smell made no difference, the probability of the student studied getting any particular number correct is simply the probability of students in general getting any particular number correct. We know the distribution of the number correct that students get in general. Thus, you can figure that probability. It turns out that it would be fairly unlikely to get only 5 correct—so the researcher concludes that not having the sense of smell does make a difference.

To go into the details a bit, the key issue is determining these probabilities. We assumed that the number correct for the students in general follows a normal curve—a specific bell-shaped mathematical pattern in which most of the scores are in the middle and there are fewer scores as the number correct gets higher or lower. There are tables showing exactly what proportions are between the middle and any particular *Z* score on the normal curve.

When considering what to conclude from a study, researchers often use a convention that if a result could have happened by chance less than 5% of the time under a particular scenario, that scenario will be considered unlikely. The normal curve tables show that the top 5% of the normal curve begins with a *Z* score of 1.64. The normal curve is completely symmetrical; thus, the bottom 5% includes all *Z* scores below -1.64. Therefore, the researcher would probably set the following rule:

The scenario in which being without the sense of smell makes no difference will be rejected as unlikely if the number correct (converted to a Z score using the mean and standard deviation for students in general) is less than -1.64.

The actual number correct for the student who could not use the sense of smell was 5. The normal curve for students in general had a mean of 14 and a standard deviation of 4. Getting 5 correct is 9 below the mean of 14; in terms of standard deviations of 4 each, it is 9/4 below the mean. A Z score of -2.25 is more extreme than -1.64. Thus, the researcher concludes that the scenario in which being without smell has no effect is unlikely. This is shown below:



7. Cutoff (.01 level, one-tailed) $= -2.33$; Z score on comparison distribution for patient studied $= -1.6$; the experiment is inconclusive. The hypothesis-testing steps, explanation, and sketch are similar to 6 above.
8. The two *M*s (5.7 and 4.8) and the $p < .05$ are crucial. *M* stands for *mean,* the average of the scores in a particular group. The average number of times per day the high-narcissism participants looked in the mirror was 5.7, while the average for the low-narcissism participants was only 4.8. The $p < .05$ tells us that this difference is statistically significant at the .05 level. This means that if a person's level of narcissism made no difference in how often the person looked in the mirror, the chances of getting two groups of participants who were this different on looking in the mirror just by chance would be 5%. Hence, you can reject that possibility as unlikely and conclude that the level of narcissism does make a difference in how often people look in the mirror.
9. Similar to 8 above.
10. (a)
    (b) and (c) are similar to (a) above.
11. (a) You can never know for certain if you have made a Type I error. (b) You can never know for certain if you have made a Type II error.

		Real Situation	
		Null Hypothesis True	**Research Hypothesis True**
Conclusion from Hypothesis Testing	**Research Hypothesis Supported (Reject null)**	*Type I Error* Decide more recess time improves behavior, but it really doesn't	*Correct Decision* Decide more recess time improves behavior, and it really does
	Study Inconclusive (Do not reject null)	*Correct Decision* Decide effect of recess time on behavior is not shown in this study; actually more recess time doesn't improve behavior	*Type II Error* Decide effect of recess time on behavior is not shown in this study; actually more recess time improves behavior