CHAPTER 5

Foundations of Inferential Statistics

In Chapters 2 through 4, we considered various statistical procedures that are used to organize and summarize data. At times, the researcher's sole objective is to describe the people (or things) in terms of the characteristic(s) associated with the data. When that is the case, the statistical task is finished as soon as the data are displayed in an organized picture, are reduced to compact indices (e.g., the mean and standard deviation), are described in terms of distributional shape, are evaluated relative to the concerns of reliability and validity, and, in the case of a bivariate concern, are examined to discern the strength and direction of a relationship.

In many instances, however, the researcher's primary objective is to draw conclusions that extend beyond the specific data that are collected. In this kind of study, the data are considered to represent a sample—and the goal of the investigation is to make one or more statements about the larger group of which the sample is only a part. Such statements, when based upon sample data but designed to extend beyond the sample, are called *statistical inferences*. Not surprisingly, the term **inferential statistics** is used to label the portion of statistics dealing with the principles and techniques that allow researchers to generalize their findings beyond the actual data sets obtained.

In this chapter, we will consider the basic principles of inferential statistics. We begin by considering the simple notions of sample, population, and scientific guess. Next, we take a look at eight of the main types of samples used by applied researchers. Then we consider certain problems that crop up to block a researcher's effort to generalize findings to the desired population. Finally, a few tips are offered concerning specific things to look for as you read professional research reports.

Statistical Inference

Whenever a statistical inference is made, a **sample** is first extracted (or is considered to have come from) a larger group called the **population.** Measurements are then taken on the people or objects that compose the sample. Once these measurements are summarized—for example, by computing a correlation coefficient—an educated guess is made as to the numerical value of the same statistical concept (which, in our example, would be the correlation coefficient) in the population. This educated guess as to the population's numerical characteristic is the **statistical inference.**

If measurements could be obtained on all people (or objects) contained in the population, statistical inference would be unnecessary. For instance, suppose the coach of the girls' basketball team at a local high school wants to know the median height of 12 varsity team members. It would be silly for the coach to use inferential statistics to answer this question. Instead of the coach making an educated guess as to the team's median height (after seeing how tall a few of the girls are), it would be easy to measure the height of each member of the varsity team and then obtain the precise answer to the question.

In many situations, researchers cannot answer their questions about their populations as easily as could the coach in the basketball example. Two reasons seem to account for the wide use of inferential statistics. One of these explanations concerns the measurement process while the other concerns the nature of the population. Because inferential statistics are used so often by applied researchers, it is worthwhile to pause for a moment and consider these two explanations as to why only portions of populations are measured, with educated guesses being made on the basis of the sample data.

First of all, it is sometimes too costly (in dollars and/or time) to measure every member of the population. For example, the intelligence of all students in a high school cannot be measured with an individual intelligence test because (1) teachers would be upset by having each student removed from classes for two consecutive periods to take the test and (2) the school's budget would not contain the funds needed to pay a psychologist to do this testing. In this situation, it would be better for the principal to make an educated guess about the average intelligence of the high school students than to have no data-based idea whatsoever as to the students' intellectual capabilities. The principal's guess about the average intelligence is based on a sample of students taken from the population made up of all students in the high school. In this example, the principal is sampling from a **tangible population** because each member of the student body could end up in the sample and be tested.

The second reason for using inferential statistics is even more compelling than the issue of limited funds and time. Often, the population of interest extends into the future. For example, the high school principal in the previous example probably would like to have information about the intellectual capabilities of the school's student body so improvements in the curriculum could be made. Such changes are made on the assumption that next year's students will not be

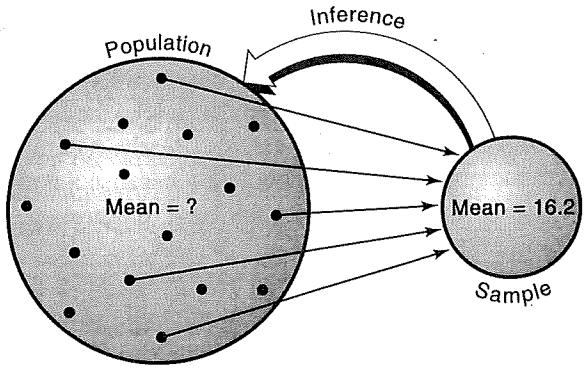
dissimilar from this year's students. Even if the funds and time could be found to administer an individual intelligence test to every student in the school, the obtained data would be viewed as coming from a *portion* of the population of interest. That population is made up of students who attend the school now *plus* students who will follow in their footsteps. Clearly, measurements cannot be obtained from all members of such a population because a portion of the population has not yet "arrived on the scene." In this case, the principal creates an **abstract population** to fit an existing sample.

Several years ago, I participated as a subject in a study to see if various levels of consumed oxygen have an effect, during strenuous exercise, on blood composition. The researcher who conducted this study was interested in what took place physiologically during exercise on a stationary bicycle among nonsedentary young men between the ages of 25 and 35. That researcher's population was not just active males who were 25–35 years old at the time of the investigation. The population was defined to include active males who *would be* in this age range at the time the research summary got published—approximately 18 months following the data collection. Inferential statistics were used because the subjects of the investigation were considered to be a representative sample of a population of similar individuals that extended into the future.

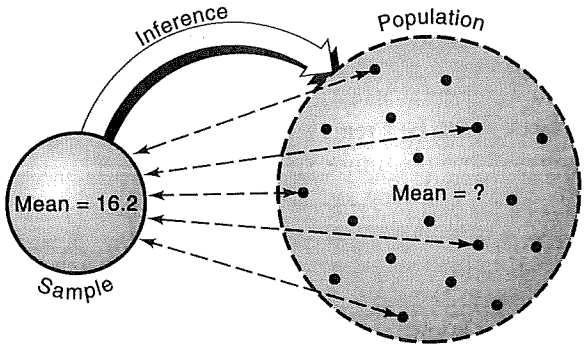
To clarify the way statistical inference works, consider the two pictures in Figure 5.1. These pictures are identical in that (1) measurements are taken only on the people (or objects) that compose the sample; (2) the educated guess, or inference, extends *from* the sample *to* the population; and (3) the value of the population characteristic is not known (nor ever can be known as a result of the inferential process). Although these illustrations show that the inference concerns the mean, the pictures could have been set up to show that the educated guess deals with the median, the variance, the product-moment correlation, or any other statistical concept.

As you can see, the only differences between the two pictures involve the solid versus dotted nature of the larger circle and the black arrows. In the top picture, the population is tangible in nature, with each member within the larger circle available for inclusion in the sample. When this is the case, the researcher actually begins with the population and then ends up with the sample. In Figure 5.1, the lower picture is meant to represent the inferential setup in which the sequence of events is reversed. Here, the researcher begins with the sample and then creates an abstract population that is considered to include people (or objects) like those included in the sample.

Excerpts 5.1 and 5.2 illustrate the distinction between tangible and abstract populations. In the first of these excerpts, the population was made up of 2,033 students in third, fourth, and fifth grade in a North Carolina community. This was a tangible population because (a) every student in the population had a unique name or ID number and (b) any of those individual students could have ended up in the sample. In Excerpt 5.1, you will see the term **sampling frame.** Generally speaking, a sampling frame is simply a list that enumerates the things—people, animals,



(a) Sampling from a tangible population



(b) Creation of an abstract population to fit an existing sample

FIGURE 5.1 *Two Kinds of Sample/Population Situations*

objects, or whatever—in the population. In a very real sense, there must be a sampling frame for any tangible population.

In Excerpt 5.2, we see a study in which the population was abstract. No sampling frame was referred to by the researchers because there wasn't one. The 90 students who composed the sample were not "pulled from" (i.e., drawn out of) a larger group; instead, they got into the sample because they voluntarily responded to a posted advertisement. Because the researchers associated with Excerpt 5.2 used inferential statistics with the data collected from those 90 students, it's clear that they wanted to generalize the study's findings beyond those specific students. The relevant population cannot be realistically defined as the full student body at the University of Chicago, for it's highly like that (a) only some of the University's students saw the posted information about the study and (b) only some of those who saw it volunteered to be in the study. Thus, the population in this study was abstract because it existed only hypothetically as a larger "mirror image" of the sample.

CHAPTER 7

Hypothesis Testing

In Chapter 6, we saw how the inferential techniques of estimation can assist researchers when they use sample data to make educated guesses about the unknown value of population parameters. Now, we turn our attention to a second way in which researchers engage in inferential thinking. This procedure is called **hypothesis testing.**

Before we turn our attention to the half-dozen elements of hypothesis testing, I'd like to reiterate something I said near the beginning of Chapter 5. In order for inferential statistics to begin, the researcher must first answer four preliminary questions: (1) What is/are the relevant population(s)? (2) How will a sample be extracted from the population(s) of interest? (3) What characteristic(s) of the sample people, animals, or objects will serve as the target of the measurement process? (4) What is the study's statistical focus—or stated differently, how will the sample data be summarized so as to obtain a statistic that can be used to make an inferential statement concerning the unknown parameter? In this chapter, I will assume that these four questions have been both raised and answered by the time the researcher starts to apply the hypothesis testing procedure.

To help you understand the six-step version of hypothesis testing, I first will simply list the various steps in their proper order (that is, the order in which a researcher ought to do things when engaged in this form of statistical inference). After presenting an ordered list of the six steps, I then will discuss the function and logic of each step.

An Ordered List of the Six Steps

Whenever researchers use the six-step version of the hypothesis testing procedure, they will

1. State the null hypothesis.
2. State the alternative hypothesis.
3. Select a level of significance.
4. Collect and summarize the sample data.
5. Refer to a criterion for evaluating the sample evidence.
6. Make a decision to discard/retain the null hypothesis.

assumptions?

It should be noted that there is no version of hypothesis testing that involves fewer than six steps. Stated differently, it is outright impossible to eliminate any of these six ingredients and have enough left to test a statistical hypothesis.

A Detailed Look at Each of the Six Steps

As indicated previously, the list of steps we just presented is arranged in an ordered fashion. In discussing these steps, however, we now will look at these six component parts in a somewhat jumbled order: 1, 6, 2, 4, 5, and then 3. My motivation in doing this is not related to sadistic tendencies! Rather, I am convinced that the function and logic of these six steps can be understood far more readily if we purposely chart an unusual path through the hypothesis testing procedure. Please note, however, that the six steps will now be rearranged only for pedagogical reasons. If I were asked to apply these six steps in an actual study, I would use the ordered list as my guide, not the sequence to which we now turn.

Step 1: The Null Hypothesis

Research Navigator.com

Null hypothesis

When engaged in hypothesis testing, a researcher begins by stating a **null hypothesis.** If there is just one population involved in the study, the null hypothesis is a pinpoint statement as to the unknown quantitative value of the parameter in the population of interest. To illustrate what this kind of null hypothesis might look like, suppose that (1) we conduct a study in which our population contains all full-time students enrolled in a particular university, (2) our variable of interest is intelligence, and (3) our statistical focus is the mean IQ score. Given this situation, we could set up a null hypothesis to say that $\mu = 100$. This statement deals with a population *parameter,* it is *pinpoint* in nature, and *we* made it.

The symbol for null hypothesis is H_0, and this symbol is usually followed by (1) a colon, (2) the parameter symbol that indicates the researcher's statistical focus, (3) an equal sign, and (4) the pinpoint numerical value that the researcher has

selected. Accordingly, we could specify the null hypothesis for our imaginary study by stating H_0: $\mu = 100$.

If our study's statistical focus involved something other than the mean, we would have to change the parameter's symbol so as to make H_0 consistent with the study's focus. For example, if our imaginary study were to be concerned with the variance among students' heights, the null hypothesis would need to contain the symbol σ^2 rather than the symbol μ. Or, if we were concerned with the product-moment correlation between the students' heights and weights, the symbol ρ would have to appear in H_0.

With respect to the pinpoint numerical value that appears in the null hypothesis, researchers have the freedom to select any value that they wish to test. Thus in our example dealing with the mean IQ of university students, the null hypothesis could be set up to say that $\mu = 80$, $\mu = 118$, $\mu = 101$, or $\mu =$ any specific value of our choosing. Likewise, if our study focused on the variance, we could set up H_0, the null hypothesis, to say that $\sigma^2 = 10$ or that $\sigma^2 =$ any other positive number of our choosing. And in a study having Pearson's product-moment correlation coefficient as its statistical focus, the null hypothesis could be set up to say that $\rho = 0.00$ or that $\rho = -.50$ or that $\rho = +.92$ or that $\rho =$ any specific number between -1.00 and $+1.00$.

The only statistical restrictions on the numerical value that appears in H_0 are that it (1) must lie somewhere on the continuum of possible values that correspond to the parameter and (2) cannot be fixed at the upper or lower limit of that continuum, presuming that the parameter has a lowest and/or highest possible value. These restrictions rule out the following null hypotheses:

$$H_0: \sigma^2 = -15 \qquad H_0: \rho = +1.30$$
$$H_0: \sigma^2 = 0 \qquad H_0: \rho = -1.00$$

because the variance has a lower limit of 0 while Pearson's product-moment correlation coefficient has limits of ± 1.00.

Excerpts 7.1 and 7.2 show how researchers sometimes talk about their null hypotheses. In the first of these excerpts, it is clear that correlation is the statistical focus in each of the two null hypotheses. If either of these null hypotheses had been expressed in symbols rather than in words, it would have taken the form H_0: $\rho = 0.00$, where ρ would represent the correlation in the population of "lecturer practitioners" in nursing. In the first null hypothesis, the two variables involved in this correlation would be experience and stress; in the second null hypothesis, the two variables would be experience and burnout.

In Excerpt 7.2, the statistical focus of the null hypothesis is the mean. That is made clear by the inclusion of the symbol μ. As you can see, there are two μs in this null hypothesis. That is because there were two populations involved in this study, 10-year-old girls and 11-year-old girls. The symbol μ, of course, corresponds to the mean score on some variable of interest. As indicated in the excerpt, the

EXCERPTS 7.1–7.2 • *The Null Hypothesis*

The null hypotheses addressed were:

1. There is no correlation between LPs' experience index and their occupational stress (measured by OSI sub-scale indices).
2. There is no correlation between LPs' experience index and their burnout (measured by MBI sub-scale indices).

Source: Willamson, G. R., Webb, C., and Abelson-Mitchell, M. (2004). Developing lecturer practitioner roles using action research. *Journal of Advanced Nursing, 47*(2), p. 155.

Null hypothesis 2: There is no significant difference in mean throw distance for age groups, i.e., H_0: $\mu_{10} = \mu_{11}$, where

μ_{10} = mean distance of throw of 10-year-old girls

μ_{11} = mean distance of throw of 11-year-old girls.

Source: Salonia, M. A., Chu, D. A., Cheifetz, P. M., and Friedhoff, G. C. (2004). Upper-body power as measured by medicine-ball throw distance and its relationship to class level among 10- and 11-year-old female participants in club gymnastics. *Journal of Strength & Conditioning Research, 18*(4), p. 699.

---

researchers wanted to compare the two groups of girls in terms of how far they could throw a medicine ball.

Earlier, I indicated that every null hypothesis must contain a pinpoint numerical value. From what is stated in Excerpt 7.1 (or from what I have already said about that excerpt's null hypothesis), it is clear that the pinpoint number in this excerpt's H_0 is zero. But what about Excerpt 7.2? This excerpt's null hypothesis also has a pinpoint number, but it's hidden. If two things are equal, there is no difference between them, and the notion of no difference is equivalent to saying that a zero difference exists. Accordingly, the null hypothesis shown in Excerpt 7.2 could be rewritten H_0: $\mu_{10} - \mu_{11} = 0$, where the subscripts 10 and 11 designate the ages of the two populations of girls.

Although researchers have the freedom to select any pinpoint number they wish for H_0, a zero is often selected when two or more populations are being compared. When this is done, the null hypothesis becomes a statement that there is no difference between the populations. Because of the popularity of this kind of null hypothesis, people sometimes begin to think that a null hypothesis *must* be set up as a "no difference" statement. This is both unfortunate and wrong. When two populations are compared, the null hypothesis can be set up with any pinpoint value the researcher wishes to use. (For example, in comparing the mean height of men and women, we could set up a legitimate null hypothesis that stated

H_0: $\mu_{men} - \mu_{women} = 2$ inches.) When the hypothesis testing procedure is used with a single population, the notion of "no difference," applied to parameters, simply doesn't make sense. How could there be a difference, zero or otherwise, when there is only one μ (or only one ρ, or only one σ^2, etc.)?

In Excerpts 7.3 and 7.4, we see two additional null hypotheses. In the first of these excerpts, the null hypothesis stated that two population proportions were equal. If we let the capital letter P stand for population proportion, we can express the null hypothesis of Excerpt 7.3 like this: H_0: $P_{1995} = P_{2001}$. The subscripts here serve to distinguish the two different populations of interest to the researchers in this study.[1]

EXCERPTS 7.3–7.4 • *Two Additional Null Hypotheses*

The researchers tabulated frequencies and percentages and used a z-test for the equality of two population proportions to test the null hypothesis that the two population proportions (1995 respondents compared to 2001 respondents) were equal.

Source: Oster-Aaland, L. K., Sellnow, T. L., Nelson, P. E., and Pearson, J. C. (2004). The status of service learning in departments of communication: A follow-up study. *Communication Education, 53*(4), p. 351.

The null hypothesis was that there would be no intra-group (within-group) difference between the results obtained at baseline and after interventions 5, 9, and at 1 month after intervention 9 for Groups 1 and 2:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

Source: Dimou, E. S., Brantingham, J. W., and Wood, T. (2004). A randomized, controlled trial (with blinded observer) of chiropractic manipulation and achilles stretching vs. orthotics for the treatment of plantar fasciitis. *Journal of the American Chiropractic Association, 41*(9), p. 33.

Excerpt 7.4 shows a null hypothesis that involves four population means. The four μs in this null hypothesis do not correspond to four different groups; instead, these μs represent the performance of a single population measured at four points in time. Thus, the subscripts on the μs correspond to the first point in time (baseline), the second point in time (after intervention 5), and so on. This null hypothesis was actually used twice in the study, because there were two groups of people with a foot problem called plantar fasciitis. People in one of the groups were treated with chiropractic manipulation and Achilles stretching; people in the other group received shoe inserts. This null hypothesis was investigated separately for each of these two groups.

[1]In Chapter 17, we will consider in depth statistical tests that focus on percentages.

Before we leave our discussion of the null hypothesis, it should be noted that H_0 does *not* always represent the researcher's personal belief, or hunch, as to the true state of affairs in the population(s) of interest. In fact, the vast majority of null hypotheses are set up by researchers in such a way as to *disagree* with what they actually believe to be the case. We will return to this point later (when we formally consider the research hypothesis). For now, however, all I want to do is alert you to the fact that the H_0 associated with any given study probably is *not* an articulation of the researcher's honest belief concerning the population(s) being studied.

Step 6: The Decision Regarding H_0

At the end of the hypothesis testing procedure, the researcher will do one of two things with H_0. One option is for the researcher to take the position that the null hypothesis is probably false. In this case, the researcher **rejects** H_0. The other option available to the researcher is to refrain from asserting that H_0 is probably false. In this case, a **fail-to-reject** decision is made.

If, at the end of the hypothesis testing procedure, a conclusion is reached that H_0 is probably false, the researcher will communicate this decision by saying one of four things: that H_0 was rejected, that a statistically significant finding was obtained, that a **reliable difference** was observed, or that p is less than a small decimal value (e.g., $p < .05$). In Excerpts 7.5 through 7.7, we see examples of how researchers will sometimes communicate their decision to disbelieve H_0.

EXCERPTS 7.5–7.7 • *Rejecting the Null Hypothesis*

The null hypothesis was rejected.

Source: Nogueras, D. J. (2006). Occupational commitment, education, and experience as a predictor of intent to leave the nursing profession. *Nursing Economic$, 24*(2), p. 90.

The comparison group scored significantly higher on the scale "I Want to Be a Scientist" than the experimental group.

Source: Barnett, M., Lord, C., Strauss, E., Rosca, C., Langford, H., Chavez, D., and Deni, L. (2006). Using the urban environment to engage youths in urban ecology field studies. *Journal of Environmental Education, 37*(2), p. 8.

Among the respondents, White users [of the drug ecstasy] reported more days of use during the past 90 days ($p < 0.05$) than users from other racial/ethnic groups.

Source: Sterk, C. E., Theall, K. P., and Elifson, K. W. (2006). Young adult ecstasy use patterns: Quantities and combinations. *Journal of Drug Issues, 36*(1), p. 220.

Just as there are different ways for a researcher to tell us that H_0 is considered to be false, there are various mechanisms for expressing the other possible decision concerning the null hypothesis. Instead of saying that a fail-to-reject decision has been reached, the researcher may tell us that H_0 was tenable, that H_0 was **accepted,** that no reliable differences were observed, that no significant difference was found, that the result was not significant (often abbreviated as *ns* or *NS*), or that p is greater than a small decimal value (e.g., $p > .05$). Excerpts 7.8 through 7.11 illustrate these different ways of communicating a fail-to-reject decision.

EXCERPTS 7.8–7.11 • *Failing to Reject the Null Hypothesis*

Our analysis of walking failed to demonstrate a significant difference between healthy elderly fallers and non-fallers walking freely under single-task conditions.

Source: Toulotte, C., Thevenon, A., Watelain, E., and Fabre, C. (2006). Identification of healthy elderly fallers and non-fallers by gait analysis under dual-task conditions. *Clinical Rehabilitation, 20*(3), p. 274

Therefore, the null hypothesis was accepted.

Source: Rouse, W. J., and Hollomon, H. L. (2005). A comparison of student test results: Business and marketing education National Board Certified Teachers and non-national Board Certified Teachers. *Delta Pi Epsilon Journal, 47*(3), p. 138.

Participants were selected such that the three experimental groups did not differ in age, $F(2, 54) < 1$, *ns*.

Source: van der Sluis, S., van der Leij, A., and de Jong, P. F. (2005). Working memory in Dutch children with reading- and arithmetic-related LD. *Journal of Learning Disabilities, 38*(3), p. 210.

The main effect of gender was not significant, $F(1, 20) = 1.2$, $p > .05$.

Source: Andersen, G. J., and Enriquez, A. (2006). Aging and the detection of observer and moving object collisions. *Psychology and Aging, 21*(1), p. 79.

It is especially important to be able to decipher the language and notation used by researchers to indicate the decision made concerning H_0. This is because most researchers neither articulate their null hypotheses nor clearly state that they used the hypothesis testing procedure. Often, the only way to tell that a researcher has used this kind of inferential technique is by noting what happened to the null hypothesis.

Step 2: The Alternative Hypothesis

Near the beginning of the hypothesis testing procedure, the researcher must state an **alternative hypothesis.** Referred to as H_a (or as H_1), the alternative hypothesis takes the same form as the null hypothesis. For example, if the null hypothesis deals with the possible value of Pearson's product-moment correlation in a single population (e.g., H_0: $\rho = +.50$), then the alternative hypothesis must also deal with the possible value of Pearson's correlation in a single population. Or, if the null hypothesis deals with the difference between the means of two populations (perhaps indicating that $\mu_1 = \mu_2$), then the alternative hypothesis must also say something about the difference between those populations' means. In general, therefore, H_a and H_0 are identical in that they must (1) deal with the same number of populations, (2) have the same statistical focus, and (3) involve the same variable(s).

The only difference between the null and alternative hypothesis is that the possible value of the population parameter included within H_a will always differ from what is specified in H_0. If the null hypothesis is set up so as to say H_0: $\rho = +.50$, then the alternative hypothesis might be set up to say H_a: $\rho \neq +.50$; or, if a researcher specifies, in Step 1, that H_0: $\mu_1 = \mu_2$, we might find that the alternative hypothesis is set up to say H_a: $\mu_1 \neq \mu_2$.

Excerpt 7.12 contains an alternative hypothesis, labeled H_a, as well as the null hypothesis with which it was paired. Notice that both H_0 and H_a deal with the same population and have the same statistical focus (the mean). If expressed symbolically, these two hypotheses would have looked identical to the H_0 and H_a shown in the final sentence of the previous paragraph. Expressed in that manner, the μs appearing in both H_0 and H_a would be representing admission rates.

EXCERPT 7.12 • *The Alternative Hypothesis*

H_0 = no difference between the mean reductions in admission rates of the two populations

H_a = the population means are different

Source: Smith, R. B. (2001). Gatekeepers and sentinels: Their consolidated effects on inpatient medical care. *Evaluation Review, 25*(3), p. 293.

As was indicated in the previous section, the hypothesis testing procedure terminates (in Step 6) with a decision to either reject or fail to reject the null hypothesis. In the event that H_0 is rejected, H_a represents the state of affairs that the researcher will consider to be probable. In other words, H_0 and H_a always represent two opposing statements as to the possible value of the parameter in the population(s) of interest. If, in Step 6, H_0 is rejected, then belief shifts *from* H_0 *to* H_a.

Stated differently, if a reject decision is made at the end of the hypothesis testing procedure, the researcher will reject H_0 *in favor of* H_a.

Although researchers have flexibility in the way they set up alternative hypotheses, they normally will set up H_a either in a **directional** fashion or in a **nondirectional** fashion.[2] To clarify the distinction between these options for the alternative hypothesis, let's imagine that a researcher conducts a study to compare men and women in terms of intelligence. Further suppose that the statistical focus of this hypothetical study is on the mean, with the null hypothesis asserting that H_0: $\mu_{men} = \mu_{women}$. Now, if the alternative hypothesis is set up in a nondirectional fashion, the researcher will simply state H_a: $\mu_{men} \neq \mu_{women}$. If, on the other hand, the alternative hypothesis is stated in a directional fashion, the researcher will specify a direction in H_a. This could be done by asserting H_a: $\mu_{men} > \mu_{women}$ *or* by asserting H_a: $\mu_{men} < \mu_{women}$.

The directional/nondirectional nature of H_a is highly important within the hypothesis testing procedure. The researcher will need to know whether H_a was set up in a directional or nondirectional manner in order to decide whether to reject (or to fail to reject) the null hypothesis. No decision can be made about H_0 unless the directional/nondirectional character of H_a is clarified.

In most empirical studies, the alternative hypothesis is set up in a nondirectional fashion. Thus if I had to guess what H_a would say in studies containing the null hypotheses presented here on the left, I would bet that the researchers had set up their alternative hypotheses as indicated on the right.

Possible H_0	***Corresponding nondirectional H_a***
H_0: $\mu = 100$	H_a: $\mu \neq 100$
H_0: $\rho = +.20$	H_a: $\rho \neq +.20$
H_0: $\sigma^2 = 4$	H_a: $\sigma^2 \neq 4$
H_0: $\mu_1 - \mu_2 = 0$	H_a: $\mu_1 - \mu_2 \neq 0$

Researchers typically set up H_a in a nondirectional fashion because they do not know whether the pinpoint number in H_0 is too large or too small. By specifying a nondirectional H_a, the researcher permits the data to point one way or the other in the event that H_0 is rejected. Hence, in our hypothetical study comparing men and women in terms of intelligence, a nondirectional alternative hypothesis would allow us to argue that μ_{women} is probably higher than μ_{men} (in the event that we reject the H_0 because $\overline{X}_{women} > \overline{X}_{men}$); or such an alternative hypothesis would allow us to argue that μ_{men} is probably higher than μ_{women} (if we reject H_0 because $\overline{X}_{men} > \overline{X}_{women}$).

[2]A directional H_a is occasionally referred to as a *one-sided* H_a; likewise, a nondirectional H_a is sometimes referred to as a *two-sided* H_a.

Occasionally, a researcher will feel so strongly (based on theoretical consideration or previous research) that the true state of affairs falls on one side of H_0's pinpoint number that H_a is set up in a directional fashion. So long as the researcher makes this decision prior to looking at the data, such a decision is fully legitimate. It is, however, totally inappropriate for the researcher to look at the data first and then subsequently decide to set up H_a in a directional manner. Although a decision to reject or fail to reject H_0 could still be made after first examining the data and then articulating a directional H_a, such a sequence of events would sabotage the fundamental logic and practice of hypothesis testing. Simply stated, decisions concerning how to state H_a (and how to state H_0) must be made without peeking at any data.

Research Navigator.com

One-tailed test
Two-tailed test

When the alternative hypothesis is set up in a nondirectional fashion, researchers sometimes use the phrase **two-tailed test** to describe their specific application of the hypothesis testing procedure. In contrast, directional H_as lead to what researchers sometimes refer to as **one-tailed tests.** Inasmuch as researchers rarely specify the alternative hypothesis in their technical write-ups, the terms *one-tailed* and *two-tailed* help us to know exactly how H_a was set up. For example, consider Excerpts 7.13 and 7.14. Here, we see how researchers sometimes use the term *two-tailed* or *one-tailed* to communicate their decisions to set up H_a in a nondirectional or directional fashion.

EXCERPTS 7.13–7.14 • *Two-Tailed and One-Tailed Tests*

Two-tailed tests were used for all analyses.

Source: Baker, D. W., Cameron, K. A., Feinglass, J., Thompson, J. A., Georgas, P., Foster, S., Pierce, D., and Hasnain-Wynia, R. (2006). A system for rapidly and accurately collecting patients' race and ethnicity. *American Journal of Public Health, 96*(3), p. 533.

Directional (one-tailed) tests were used because improvement was hypothesized over sequential treatment sessions.

Source: Storchheim, L. F., and O'Mahony, J. F. (2006). Compulsive behaviours and levels of belief in obsessive–compulsive disorder: A case-series analysis of their interrelationships. *Clinical Psychology & Psychotherapy, 13*(1), p. 70.

If H_a is set up in a directional manner, the null hypothesis can be expressed as an **inexact H_0.** This type of null hypothesis functions exactly like the kind of H_0 we have been considering, so it really does not matter whether H_0 takes the form of an inexact statement or an exact statement. Nonetheless, I feel it necessary to illustrate what an inexact H_0 looks like so you are not thrown into a tizzy if you ever see one in a research report.

Suppose a researcher wants to compare a sample of high school students against a sample of college students in terms of their vocabulary. Further suppose that our hypothetical researcher can look ahead into the hypothesis testing procedure and knows that the alternative hypothesis in Step 2 will be set up in a directional manner to say H_a: $\mu_{\text{college}} > \mu_{\text{high school}}$. If the researcher knows from the beginning that H_a will be directional, then the null hypothesis (in Step 1) could be set up to say H_0: $\mu_{\text{college}} \leq \mu_{\text{high school}}$. This null hypothesis is inexact because it does not contain a pinpoint numerical value for the population parameter (as would be the case if the null hypothesis were to be set up to say H_0: $\mu_{\text{college}} - \mu_{\text{high school}} = 0$). Instead, this inexact H_0 says that the mean vocabulary among college students is equal to or lower than the mean vocabulary among high school students, with lower being anything from a tiny amount to an enormous difference.

Excerpt 7.15 provides an illustration of an inexact null hypothesis. If this H_0 had been expressed symbolically, it would have taken the form H_0: $\rho \leq 0.00$. It is worth noting that the null hypothesis, when set up to be inexact, does not overlap whatsoever with the alternative hypothesis. This is due to the general requirement that H_0 and H_a be mutually exclusive.

EXCERPT 7.15 • *An Inexact Null Hypothesis (and Its Alternative Hypothesis)*

H_0: The relationship between social performance and financial performance in the commercial banking industry is either zero or negative.

H_a: The relationship between social performance and financial performance in the commercial banking industry is positive.

Source: Simpson, W. G., and Kohers, T. (2002). The link between corporate social and financial performance: Evidence from the banking industry. *Journal of Business Ethics, 35*(2), p. 102.

In terms of the ultimate reject or fail-to-reject decision reached by the researcher, it makes absolutely no difference whether the null hypothesis is set up to be exact or inexact. I prefer to articulate any null hypothesis as an exact H_0, because this is consistent with the notion that the null hypothesis is a *point* on a numerical continuum, with the alternative hypothesis represented by either (1) the rest of that continuum, both above and below the null point, if H_a is nondirectional, or (2) the segment of the continuum that lies on just one side of the null point, if H_a is directional. Certain authors have a preference for conceptualizing H_0 in an inexact manner, thereby equating H_0 to one of those two segments. It really doesn't matter which definition of H_0 is used. (If H_a is nondirectional, however, there is no option. In that more common situation, H_0 must be exact.)

Step 4: Collection and Analysis of Sample Data

So far, we have covered Steps 1, 2, and 6 of the hypothesis testing procedure. In the first two steps, the researcher states the null and alternative hypotheses. In Step 6, the researcher will either (1) reject H_0 in favor of H_a or (2) fail to reject H_0. We now turn our attention to the principal stepping-stone that is used to move from the beginning points of the hypothesis testing procedure to the final decision.

Inasmuch as the hypothesis testing procedure is, by its very nature, an empirical strategy, it should come as no surprise that the researcher's ultimate decision to reject or to retain H_0 is based on the collection and analysis of sample data. No crystal ball is used, no Ouija board is relied on, and no eloquent argumentation is permitted. Once H_0 and H_a are fixed, only scientific evidence is allowed to affect the disposition of H_0.

The fundamental logic of the hypothesis testing procedure can now be laid bare because the connections between H_0, the data, and the final decision are as straightforward as what exists between the speed of a car, a traffic light at a busy intersection, and a lawful driver's decision as the car approaches the intersection. Just as the driver's decision to stop or to pass through the intersection is made after observing the color of the traffic light, the researcher's decision to reject or to retain H_0 is made after observing the sample data. To carry this analogy one step further, the researcher will look at the data and ask, "Is the empirical evidence inconsistent with what one would expect if H_0 were true?" If the answer to this question is yes, then the researcher has a green light and will reject H_0. On the other hand, if the data turn out to be consistent with H_0, then the data set serves as a red light telling the researcher not to discard H_0.

Because the logic of hypothesis testing is so important, let us briefly consider a hypothetical example. Suppose a valid intelligence test is given to a random sample of 100 males and a random sample of 100 females attending the same university. If the null hypothesis had first been set up to say H_0: $\mu_{male} = \mu_{female}$ and if the data reveal that the two sample means (of IQ scores) differ by only two points, the sample data would be consistent with what we expect to happen when two samples are selected from populations having identical means. Clearly, the notion of sampling error could fully explain why the two $\bar{X}$s might differ by two IQ points even if $\mu_{male} = \mu_{female}$. In this situation, no empirical grounds exist for making the data-based claim that males at our hypothetical university have a different IQ, on the average, than do their female classmates.

Now, let's consider what would happen if the difference between the two sample means turns out to be equal to 40 (rather than 2) IQ points. If the empirical evidence turns out like this, we would have a situation where the data are inconsistent with what one would expect if H_0 were to be true. Although the concept of sampling error strongly suggests that neither sample mean will turn out exactly equal to its population parameter, the difference of 40 IQ points between $\bar{X}_{males}$ and $\bar{X}_{females}$ is quite improbable if, in fact, μ_{males} and $\mu_{females}$ are equal. With results such as this, the researcher would reject the arbitrarily selected null hypothesis.

To drive home the point I'm trying to make about the way the sample data influence the researcher's decision concerning H_0, let's shift our attention to a real study that had Pearson's correlation as its statistical focus. In Excerpt 7.16, the hypothesis testing procedure was used to evaluate two bivariate correlations based on data that came from watching 40 pairs of children at play (with a sibling) and then talking with each child about conflicts that arose during the play. The two variables that were correlated were age of the child and the degree to which the child denied responsibility for an interpersonal conflict with his or her siblings. The correlation between these two variables was computed separately for each of two groups of children: younger siblings (who were between 3.5 and 5.3 years old) and their older siblings (who were between 5.5 and 8.9 years old).

EXCERPT 7.16 • *Rejecting H_0 When the Sample Data Are Inconsistent with H_0*

> For younger siblings, age was not related to denials, $r = -.02$, *ns*. However, for older siblings the relation was significant, $r = -.63$, [thus] indicating that as the age of older siblings increased (from 5.5 to 8.9), older siblings became considerably less likely to rely on denials.
>
> *Sources:* Wilson, A. E., Smith, M. D., Ross, H. S., and Ross, M. (2004). Young children's personal accounts of their sibling disputes. *Merrill-Palmer Quarterly, 50*(1), p. 53.

In the study associated with Excerpt 7.16, the hypothesis testing procedure was used separately to evaluate each of the two sample *r*s. In each case, the null hypothesis stated H_0: $\rho = 0.00$. The sample data, once analyzed, yielded correlations of $-.02$ and $-.63$. The first of these *r*s ended up being quite close to the pinpoint number in H_0, 0.00. The small difference between the null number and $-.02$ can easily be explained by sampling error. In other words, if the correlation in the population were truly equal to 0.00, it would not be surprising to have a sample *r* (with $n = 20$) that deviates from 0.00 but by only a small amount. Accordingly, the null hypothesis concerning the age-denial correlation for the young children was not rejected, as indicated by the notation *ns*.

The second correlation in Excerpt 7.16 turned out to be quite different from the null hypothesis number of 0.00. Statistically speaking, the *r* of $-.63$ was so inconsistent with H_0 that sampling error alone was considered to be an inadequate explanation for why the observed correlation was so far away from the pinpoint number in the null hypothesis. Although we would expect some discrepancy between 0.00 and the data-based value of *r* even if H_0 were true, we would *not* expect this big of a difference. Accordingly, the null hypothesis concerning the age-denial correlation for the older children was rejected, as indicated by the phrase "the relation was significant."

In Step 4 of the hypothesis testing procedure, the summary of the sample data will always lead to a single numerical value. Being based on the data, this number is technically referred to as the **calculated value.** (It is also called the **test statistic.**) Occasionally, the researcher's task in obtaining the calculated value involves nothing more than computing a value that corresponds to the study's statistical focus. This was the case in Excerpt 7.16, where the statistical focus was Pearson's correlation coefficient and where the researcher needed to do nothing more than compute a value for r.

In most applications of the hypothesis testing procedure, the sample data are summarized in such a way that the statistical focus becomes hidden from view. For example, consider Excerpts 7.17 and 7.18. In the first of these excerpts, the calculated value was labeled F and it turned out equal to 4.267. In Excerpt 7.18, the calculated value was t, and this time it turned out equal to -0.48. In each of these excerpts, the statistical focus was the mean.

EXCERPTS 7.17–7.18 • *The Calculated Value*

Men on average published 5.50 ($SD = 5.20$) articles over the course of 3 years, and the average for women was 4.66 ($SD = 4.18$) articles, $F(1, 534) = 4.267$, $p < .05$.

Source: Sellers, S. L., Smith, T., Mathiesen, S. G., and Perry, R. (2006). Perceptions of professional social work journals: Findings from a national survey. *Journal of Social Work Education, 42*(1), p. 146.

The children who listened to the musical story were less accurate ($M = 14.13$, $SD = 24.17$) than the children who listened to the spoken version of the story ($M = 17.50$, $SD = 21.33$). This difference, however, was not statistically significant ($t(40) = -0.48$, $p > .05$).

Source: Noguchi, L. K. (2006). The effect of music versus nonmusic on behavioral signs of distress and self-report of pain in pediatric injection patients. *Journal of Music Therapy, 43*(1), p. 27.

In each of these excerpts, two sample means were compared. In Excerpt 7.17, the mean of 5.50 was compared against the mean of 4.66. In Excerpt 7.18, the means were 14.13 and 17.50. Within each of these studies, the researchers put their sample data into a formula that produced the calculated value. The important thing to notice in these excerpts is that in neither case does the calculated value equal the difference between the two means being compared. In Chapter 10, we'll consider t-tests and F-tests in more detail, so you should not worry now if you do not currently comprehend everything that is presented in these excerpts. They are shown solely to illustrate the typical situation in which the statistical focus of a study is *not* reflected directly in the calculated value.

Before computers were invented, researchers would always have a single goal in mind when they turned to Step 4 of the hypothesis testing procedure. That goal was the computation of the data-based calculated value. Now that computers are widely available, researchers still are interested in the magnitude of the calculated value derived from the data analysis. Contemporary researchers, however, are also interested in a second piece of information generated by the computer. This second item is the data-based *p*-value.

Whenever researchers use a computer to perform the data analysis, they will either (1) tell the computer what the null hypothesis is going to be or (2) accept the computer's built-in default version of H_0. The researcher will also specify whether H_a is directional or nondirectional in nature. Once the computer knows what the researcher's H_0 and H_a are, it can easily analyze the sample data and compute the probability of having a data set that deviates as much or more from H_0 as does the data set being analyzed. The computer informs the researcher as to this probability by means of a statement that takes the form $p =$ _____, with the blank being filled by a single decimal value somewhere between 0 and 1.

Excerpts 7.19 and 7.20 illustrate nicely how a *p*-value is like a calculated value in that either one can be used as a single-number summary of the sample data. As you can see, two sample percentages were compared in Excerpt 7.19, while two correlation coefficients were computed in Excerpt 7.20. In the first of these excerpts, the researchers used a *p*-value to assess the likelihood, under the null hypothesis, of getting two percentages that differed as much or more than the two percentages actually observed (34 versus 20). The researchers associated with Excerpt 7.20 did the same

EXCERPTS 7.19–7.20 • *Using p as the Calculated Value*

Women receiving radiation therapy are more likely to have reduced shoulder mobility one year after breast cancer diagnosis (34% women had loss of range of motion compared with 20% who did not receive radiation therapy, $P = 0.03$).

Source: Internal, M. E., Freeman, J. L., Zhang, D. D., Jansen, C., Ostir, G., Hatch, S. S., and Goodwin, J. S. (2006). The relationship between depressive symptoms and shoulder mobility among older women: Assessment at one year after breast cancer diagnosis. *Clinical Rehabilitation, 20*(6), p. 518.

Age was not significantly correlated with pre-test scores, $r = -0.13, n = 111$, $P = 0.16$, nor improvement, $r = -0.02, n = 110, P = 0.85$.

Source: Law, A. V., and Shapiro, K. (2005). Impact of a community pharmacist-directed clinic in improving screening and awareness of osteoporosis. *Journal of Evaluation in Clinical Practice, 11*(3), p. 253.

sort of thing, because they used a *p*-value to determine how likely it would be, assuming the null hypothesis to be true, to end up with a sample correlation as large or larger than each of their computed *r*s. Within both of these excerpts, each *p* functioned as a measure of how inconsistent the sample data were compared with what we'd expect to happen if H_0 were true.

Be sure to note that there is an *inverse* relationship between the size of *p* and the degree to which the sample data deviate from the null hypothesis. In Excerpt 7.19, the *p*-value would have been larger than 0.03 if the two percentages had been closer together (or smaller than 0.03 if those percentages had been even further apart). In Excerpt 7.20, the *p*-value would have been smaller if the *r*s had been further away from zero (or smaller if the *r*s had turned out even lower than they did).

Step 5: The Criterion for Evaluating the Sample Evidence

After the researcher has summarized the study's data, the next task involves asking the question "Are the sample data inconsistent with what would likely occur if the null hypothesis were true?" If the answer to this question is "yes," then H_0 will be rejected; on the other hand, a negative response to this query will bring forth a fail-to-reject decision. Thus as soon as the sample data can be tagged as consistent or inconsistent (with H_0), the decision in Step 6 is easily made. "But how," you might ask, "does the researcher decide which of these labels should be attached to the sample data?"

If the data from the sample(s) are in perfect agreement with the pinpoint numerical value specified in H_0, then it is obvious that the sample data are consistent with H_0. (This would be the case if the sample mean turned out equal to 100 when testing H_0: $\mu = 100$, if the sample correlation coefficient turned out equal to 0.00 when testing H_0: $\rho = 0.00$, etc.) Such a situation, however, is unlikely. Almost always, there will be a discrepancy between H_0's parameter value and the corresponding sample statistic.

In light of the fact that the sample statistic (produced by Step 4) is almost certain to be different from H_0's pinpoint number (specified in Step 1), the concern over whether the sample data are inconsistent with H_0 actually boils down to the question "Should the observed difference between the sample evidence and the null hypothesis be considered to be a big difference or a small difference?" If this difference (between the data and H_0) is judged to be large, then the sample data will be looked on as being inconsistent with H_0 and, as a consequence, H_0 will be rejected. If, on the other hand, this difference is judged to be small, the data and H_0 will be looked on as consistent with each other and, therefore, H_0 will not be rejected.

To answer the question about the sample data's being either consistent or inconsistent with what one would expect if H_0 were true, a researcher can use either of two simple procedures. As you will see, both of these procedures involve comparing a single-number summary of the sample evidence against a criterion number. The single-number summary of the data can be either the calculated value or

the p-value. Our job now is to consider what each of these data-based indices is compared against and what kind of result forces researchers to consider their samples as representing a large or a small deviation from H_0.

One available procedure for evaluating the sample data involves comparing the calculated value against something called the **critical value.** The critical value is nothing more than a number extracted from one of many statistical tables developed by mathematical statisticians. Applied researchers, of course, do not close their eyes and point to just any entry in a randomly selected table of critical values. Instead, they must learn which table of critical values is appropriate for their studies and also how to locate the single number within the table that constitutes the correct critical value.

As a reader of research reports, you do not have to learn how to locate the proper table that contains the critical value for any given statistical test, nor do you have to locate, within the table, the single number that allows the sample data to be labeled as being consistent or inconsistent with H_0. The researcher will do these things. Occasionally, the critical value will be included in the research report, as exemplified in Excerpts 7.21 through 7.23.

EXCERPTS 7.21–7.23 • *The Critical Value and the Decision Rule*

Results indicated that subjects in the treatment group had significantly higher post-test self-concept scores than the control group (obtained t of 6.58 is greater than the critical t-value of 1.96 with $df = 66$ at alpha level of .05).

Source: Egbochuku, E. O., and Obiunu, J. J. (2006). The effect of reciprocal peer counseling in the enhancement of self-concept among adolescents. *Education, 126*(3), p. 504.

Marital status has a noticeable but not significant influence on students' use of their time for economic activities. This is because the chi-square value of 4.19 is less than the critical value of 5.99 at .05 level of probability.

Source: Ogonor, B. O., and Nwadiani, M. (2006). An analysis of non-instructional time management of undergraduates in southern Nigeria. *College Student Journal, 40*(1), pp. 209–210.

The hypothesis testing for hierarchical regression analysis (Cohen et al., 2003) found that the computed $F(5, 288) = 3.10$ was greater than the critical $F(5, 288) = 2.25$. Therefore, the hypothesis of this study was supported. The conclusion was to reject the null hypothesis at an alpha of .05.

Source: Pluta, D. J., and Accordino, M. P. (2006). Predictors of return to work for people with psychiatric disabilities: A private sector perspective. *Rehabilitation Counseling Bulletin, 49*(2), p. 105.

Once the critical value is located, the researcher will compare the data-based summary of the sample data against the scientific dividing line that has been extracted from a statistical table. The simple question being asked at this point is whether the calculated value is larger or smaller than the critical value. With most tests (such as *t*, *F*, chi-square, and tests of correlation coefficients), the researcher will follow a decision rule that says to reject H_0 if the calculated value is at least as large as the critical value. With a few tests (such as *U* or *W*), the decision rule tells the researcher to reject H_0 if the calculated value is smaller than the critical value. You do not need to worry about which way the decision rule works for any given test because this is the responsibility of the individual who performs the data analysis. The only things you need to know about the comparison of calculated and critical values are (1) that this comparison allows the researcher to decide easily whether to reject or fail to reject H_0 and (2) that some tests use a decision rule that says to reject H_0 if the calculated value is larger than the critical value, whereas other tests involve a decision rule that says to reject H_0 if the calculated value is smaller than the critical value.

The researchers associated with Excerpts 7.21, 7.22, and 7.23 helped the readers of their research reports by specifying not only the critical value but also the nature of the decision rule that was used when the calculated value was compared against the critical value. In most research reports, you will not see either of these things; instead, you will only be given the calculated value. (On rare occasions, you won't even see the calculated value.) As indicated previously, however, you should not be concerned about this because it is the researcher's responsibility to obtain the critical value and to know which way the decision rule operates. When reading most research reports, all you can do is trust that the researcher did these two things properly.

Research Navigator.com

Level of significance

The second way a researcher can evaluate the sample evidence is to compare the data-based *p*-value against a preset point on the 0-to-1 scale on which the *p* must fall. This criterion is called the **level of significance,** and it functions much as does the critical value in the first procedure for evaluating sample evidence. Simply stated, the researcher compares his or her data-based *p*-value against the criterion point along the 0-to-1 continuum so as to decide whether the sample evidence ought to be considered consistent or inconsistent with H_0. The decision rule used in this second procedure is always the same: If the data-based *p*-value is equal to or smaller than the criterion, the sample is viewed as being *in*consistent with H_0; if, on the other hand, *p* is larger than the criterion, the data are looked on as being consistent with H_0.

I will discuss the level of significance in more depth in the next section, since it is a concept that must be dealt with by the researcher no matter which of the two procedures is used to evaluate the sample data. (With the second procedure, the level of significance *is* the criterion against which the data-based *p*-value is compared; with the first procedure, the level of significance influences the size of the critical value against which the calculated value is compared.) Before we leave this

section, however, I need to point out that the same decision will be reached regarding H_0 no matter which of the two procedures is used in Step 5 of the hypothesis testing procedure. For example, suppose a researcher conducts an *F*-test and rejects H_0 because the calculated value is larger than the critical value. If that researcher were to compare the data-based *p* against the level of significance, it would be found that the former is smaller than the latter, and the same decision about H_0 would be made. Or, suppose a researcher conducts a *t*-test and fails to reject H_0 because the calculated value is smaller than the critical value. If that researcher were to compare the data-based *p* against the level of significance, it would be found that the former is larger than the latter, and the same fail-to-reject decision would be made.

Step 3: Selecting a Level of Significance

After the data of a study are collected and summarized, the six-step hypothesis testing procedure allows absolutely no subjectivity to influence, or bias, the ultimate decision that is made concerning the null hypothesis. This goal is accomplished by reliance on a scientific cutoff point to determine whether the sample data are consistent or inconsistent with H_0. By referring (in Step 5) to a numerical criterion, it becomes clear whether or not sampling error provides, by itself, a sufficient explanation for the observed difference between the single-number summary of the researcher's data (computed in Step 4) and H_0's pinpoint numerical value (articulated in Step 1). If the single-number summary of the data is found to lie on H_a's side of the criterion number (or if the data-based *p* lands on H_a's side of the level of significance), a decision (in Step 6) is made to reject H_0 in favor of H_a (set forth in Step 2); on the other hand, if the calculated value lands on H_0's side of the critical value (or if the data-based *p* lands on H_0's side of the level of significance), a fail-to-reject decision is made.

Either the critical value or the level of significance serves as a scientific cutoff point that determines what decision will be made concerning the null hypothesis. The six-step hypothesis testing procedure not only allows the researcher to do something that affects the magnitude of this criterion—*it actually forces the researcher to become involved in determining how rigorous the criterion will be.* The researcher should not, as I have pointed out, do anything like this after the data have been collected and summarized. However, the researcher *must* do something prior to collecting data that has an impact on how large or small the criterion number will be.

After the null and alternative hypotheses have been set up, but before any data are collected, the researcher must select a level of significance. This third step of the hypothesis testing procedure simply asks the researcher to select a positive decimal value of the researcher's choosing. Although the researcher has the freedom to select any value between 0 and 1 for the level of significance, most researchers select a small number such as .10, .05, or .01. The most frequently selected number is .05.

Before explaining how the researcher-selected level of significance influences the size of the critical value, I need to alert you to the fact that not all researchers

use the phrase *level of significance* to designate the decimal number that must be specified in Step 3. Instead of indicating, for example, that the level of significance is set equal to .05, some researchers will state that "the **alpha level** (α) is set equal to .05," others will assert that "$p = .05$," and still others will indicate that "H_0 will be rejected if $p < .05$." Likewise, a decision to use the .01 level of significance might be expressed using statements such as "alpha = .01," "$\alpha = .01$," or "results will be considered significant if $p < .01$."

In Excerpts 7.24 through 7.28, we see different ways in which researchers report what level of significance was selected within their studies.

If the single-number summary of the sample data is a p-value, the pragmatic value of the level of significance is clear. In this situation, p is compared directly

EXCERPTS 7.24–7.28 • *The Level of Significance*

A significance level of .05 was used for all statistical analyses.

Source: Mangione, K. K., Craik, R. L., Tomlinson, S. S., and Palombaro, K. M. (2005). Can elderly patients who have had a hip fracture perform moderate- to high-intensity exercise at home? *Physical Therapy, 85*(8), p. 734.

We used an alpha level of .05 for all analyses.

Source: Elias, S. M., and Cropanzano, R. (2006). Gender discrimination may be worse than you think: Testing ordinal interactions in power research. *Journal of General Psychology, 133*(2), p. 124.

The level of significance was set at $p < 0.05$.

Source: Davids, J. R., Peace, L. C., Wagner, L. V., Gidewall, M. A., Roberson, W. M., and Blackhurst, D. W. (2006). Validation of the Shriner's Hospital for Children Upper Extremity Evaluation (SHUEE) for children with hemiplegic cerebral palsy. *Journal of Bone & Joint Surgery, 88*(2), p. 328.

An α of 0.05 was selected for tests of significance.

Source: Sloka, J. S., Pryse-Phillips, W., and Stefanelli, M. (2006). The relation between menarche and the age of first symptoms in a multiple sclerosis cohort. *Multiple Sclerosis, 12*(3), p. 334.

For all the analyses reported below, the level of confidence for rejecting a null hypothesis was 0.05.

Source: Pichette, F. (2005). Time spent on reading and reading comprehension in second language learning. *Canadian Modern Language Review, 62*(2), p. 252.

against α to determine whether or not H_0 should be rejected. But even if the single-number summary of the sample data is a calculated value, the level of significance still performs a valuable, pragmatic function. This is because a critical value cannot be located (in Step 5) unless the level of significance has first been set. As indicated in our earlier discussion of Step 4, there are many tables of critical values. Once the proper table is located, the researcher still has the task of locating the single number within the table that will serve as the critical value. The task of locating the critical value is easy, so long as the level of significance has been specified.[3]

Although the level of significance plays an important pragmatic role within the six-step hypothesis testing procedure, the decimal number selected in Step 3 is even more important from a different perspective. When I introduced the concept of the null hypothesis and when I talked about the reject or fail-to-reject decision that researchers will make regarding the null hypothesis, I was careful to use language that did *not* suggest that H_0 is ever *proven* to be true or false by means of hypothesis testing. Regardless of the decision made about H_0 after the calculated and critical values (or p and α) are compared, it is possible that the wrong decision will be reached. If H_0 is rejected in Step 6, it is conceivable that this action represents a mistake, since H_0 may actually be true. Or, if H_0 is not rejected, it is conceivable that *this* action represents a mistake, since H_0 may actually be an inaccurate statement about the value of the parameter in the population(s).

In light of the fact that a mistake can conceivably occur regardless of what decision is made at the end of the hypothesis testing procedure, two technical terms have been coined to distinguish between these potentially wrong decisions. A **Type I error** designates the mistake of rejecting H_0 when the null hypothesis is actually true. A **Type II error,** on the other hand, designates the kind of mistake that is made if H_0 is not rejected when the null hypothesis is actually false. The following chart may help to clarify the meaning of these possible errors.

		Is H_0 Really True?	
		Yes	**No**
Researcher's Decision	Reject H_0	Type I Error	Correct Decision
	Fail-to-Reject H_0	Correct Decision	Type II Error

[3]With certain tests, researchers cannot locate the critical value unless they also know (1) whether their test is one- or two-tailed in nature and (2) how many degrees of freedom are connected with the sample data. I will discuss the concept of degrees of freedom in later chapters.

Beyond its pragmatic utility in helping the researcher locate the critical value (or in serving as the criterion against which the data-based p is compared), the level of significance is important because it establishes the probability of a Type I error. In other words, the selected alpha level determines the likelihood that a true null hypothesis will be rejected. If the researcher specifies, in Step 3, that $\alpha = .05$, then the chances of rejecting a true null hypothesis become equal to 5 out of 100. If, on the other hand, the alpha level is set equal to .01 (rather than .05), then the chances of rejecting a true null hypothesis would become equal to 1 out of 100. The alpha level, therefore, directly determines the probability that a Type I error will be committed.[4]

After realizing that the researcher can fully control the likelihood of a Type I error, you may be wondering why the researcher does not select an alpha level that would dramatically reduce the possibility that a true H_0 will be rejected. To be more specific, you may be inclined to ask why the alpha level is not set equal to .001 (where the chance of a Type I error becomes equal to 1 out of 1,000), equal to .00001 (where the chance of Type I error becomes equal to 1 out of 100,000), or even equal to some smaller decimal value. To answer this legitimate question, we must consider the way in which a change in the alpha level has an effect on both Type I error risk *and* Type II error risk.

If the alpha level is changed, it's as if there is an apothecary scale in which the two pans hanging from opposite ends of the balance beam contain, respectively, Type I error risk and Type II error risk. The alpha level of a study could be changed so as to decrease the likelihood of a Type I error, but this change in alpha will simultaneously have an opposite effect on the likelihood of a Type II error. Hence, researchers rarely move alpha from the more traditional levels of .05 or .01 to levels that would greatly protect against Type I errors (such as .0001) because such a change in the alpha level would serve to make the chances of a Type II error unacceptably high.

In Excerpts 7.29, 7.30, and 7.31, we see three cases where a connection is drawn between the selected level of significance and the likelihood of a Type I error and/or a Type II error. The third group of researchers deserve your respect for having explained why they chose the levels of significance they did. Far too many researchers, without thinking, set alpha equal to .05 simply because this is the most popular level of significance. If they weighed the risks of Type I and Type II errors in their own studies, they might choose some level of significance other than .05.

Near the beginning of this chapter, I pointed out that H_0 is normally set up so as to disagree with the researcher's personal hunch regarding the population parameter(s) focused on in the study. For example, if a researcher thinks that a new pill will reduce the mean stress level among students preparing to take their final examinations, a study might be set up involving an experimental group and a placebo group. Within this study, the researcher's null hypothesis would probably be set up to say that the pill has no effect on stress (i.e., H_0: $\mu_{experimental} = \mu_{placebo}$).

[4]As you will see later, the alpha level defines the probability of a Type I error only if (1) important assumptions underlying the statistical test are valid and (2) the hypothesis testing procedure is used to evaluate only *one* null hypothesis.

EXCERPTS 7.29–7.31 • *Alpha and the Risk of Type I and Type II Errors*

A type I error level of 5% was chosen for statistical significance.

Source: Sadri, H., MacKeigan, L. D., Leiter, L. A., and Einarson, T. R. (2005). Willingness to pay for inhaled insulin: A contingent valuation approach. *Pharmaco Economics, 23*(1), p. 1220.

For all tests of significance, the alpha error level was set at $p = .05$.

Source: Nagata, H., Dalton, P., Doolittle, N., and Breslin, P. A. S. (2005). Psychophysical isolation of the modality responsible for detecting multimodal stimuli: A chemosensory example. *Journal of Experimental Psychology: Human Perception and Performance, 31*(1), p. 102.

In order to balance a legitimate concern about the potential for Type I error with concern about Type II error, particularly given the small sample size, we set alpha at .05 for the hypotheses tested.

Source: Woodhouse, S. S., Schlosser, L. Z., Crook, R. E., Ligiéro, D. P., and Gelso, C. J. (2003). Client attachment to therapist: Relations to transference and client recollections of parental caregiving. *Journal of Counseling Psychology, 50*(4), p. 405.

In light of the fact that researchers typically like to reject H_0 to gain empirical support for their honest hunches, and in light of the fact that a change in the level of significance has an impact on the likelihood of Type II errors, you now may be wondering why the researcher does not move alpha in the opposite direction. It is true that a researcher would decrease the chance of a Type II error by changing alpha—for example, from .05 to .40—since such a change would make it more likely that H_0 would be rejected. Researchers do not use such high levels of significance simply because the scientific community generally considers Type I errors to be more dangerous than Type II errors. In most disciplines, few people would pay attention to researchers who reject null hypotheses at alpha levels higher than .20, because such levels of significance are considered to be too lenient (i.e., too likely to yield reject decisions that are Type I errors).

The most frequently seen level of significance, as illustrated earlier in Excerpts 7.24 through 7.31, is .05. This alpha level is considered to represent a happy medium between the two error possibilities associated with any application of the six-step hypothesis testing procedure. If, however, a researcher feels that it is more important to guard against the possibility of a Type I error, a lower alpha level (such as .01 or .001) will be selected. On the other hand, if it is felt that a Type II error would be more dangerous than a Type I error, then a higher alpha level (such as .10 or .15) will be selected. Excerpts 7.32 and 7.33 illustrate how (and why) researchers sometimes set alpha equal to something other than .05. In Excerpt 7.32, the

EXCERPTS 7.32–7.33 • *Reasons for Using Alpha Levels Other Than .05*

> Although the level of significance of a statistical test is generally set at 0.05 by convention, [our study's objectives] justified the acceptability of a higher Type I error rate, resulting in greater statistical power. The alpha level for statistical tests in this study was set a priori at 0.10.
>
> *Source:* Gall, J. (2006). Orienting tasks and their impact on learning and attitudes in the use of hypertext. *Journal of Educational Multimedia and Hypermedia, 15*(1), p. 15.

> A Bonferroni approach to control for Type I error across the 10 correlations indicated that a p value of less than .005 (.05/10 = .005) was required for significance.
>
> *Source:* Wrobleski, K. K., and Snyder, C. R. (2005). Hopeful thinking in older adults: Back to the future. *Aging Research, 31*(2), p. 223.

researcher wanted to guard against making Type II errors (by making sure his statistical tests had high "power"), so he set the level of significance equal to .10 rather than .05. In contrast, the researchers in Excerpt 7.33 wanted to guard against making Type I errors when statistically evaluating their study's 10 different correlation coefficients, so they changed the level of significance from .05 to .005 by making a "Bonferroni" adjustment. In the next chapter, we will consider both power and the Bonferroni adjustment. For now, the only thing you need to know is that not all alpha levels are set equal to .05.

Before concluding our discussion of the level of significance, I need to clarify two points of potential confusion. To accomplish this goal, I want to raise and then answer two questions: "Does the alpha level somehow determine the likelihood of a Type II error?" and "If H_0 is rejected, does the alpha level indicate the probability that H_0 is true?"

The first point of potential confusion concerns the relationship between alpha and Type II error risk. Since alpha does, in fact, determine the likelihood that the researcher will end up rejecting a true H_0, and since it is true that a change in alpha affects the chance of a Type I error *and* the chance of a Type II error (with one increasing, the other decreasing), you may be tempted to expect the level of significance to dictate Type II error risk. Unfortunately, this is not the case. The alpha level specified in Step 3 does influence Type II error risk, but so do other features of a study such as sample size, population variability, and the reliability of the measuring instrument used to collect data.

The second point of potential confusion about the alpha level again concerns the decision reached at the end of the hypothesis testing procedure. If a study's H_0 is rejected in Step 6, it is *not* proper to look back to see what alpha level was specified in Step 3 and then interpret that alpha level as indicating the probability that

H_0 is true. For example, if a researcher ends up rejecting H_0 after having set the level of significance equal to .05, you cannot legitimately conclude that the chances of H_0 being true are less than 5 out of 100. The alpha level in any study indicates only what the chances are that the forthcoming decision will be a Type I error. If alpha is set equal to .05, then the chances are 5 out of 100 that H_0 will be rejected *if H_0 is actually true.* Statisticians sometimes try to clarify this distinction by pointing out that the level of significance specifies "the probability of a reject decision, given a true H_0" and *not* "the probability of H_0 being true, given a reject decision."

Results That Are Highly Significant and Near Misses

As indicated earlier, the level of significance plays a highly important role in hypothesis testing. In a very real sense, it functions as a dividing line. Statistical significance is positioned on one side of that line, the lack of statistical significance on the other. That dividing line is clearly visible if the researcher decides to reject or fail-to-reject H_0 by comparing the data-based p against the level of significance. But even when the procedure for deciding H_0's fate involves comparing the data-based calculated value against a tabled critical value, the level of significance is still involved. That's because α influences the size of the critical value.

Because the level of significance plays such an important role—both pragmatically and conceptually—in hypothesis testing, it often is included when the decision about H_0 is declared. With the level of significance set at .05 (the most popular α-level), a decision to reject H_0 is often summarized by the notation $p < .05$, while a decision not to reject H_0 is summarized by the notation $p > .05$. Earlier, you saw such notational summaries in Excerpts 7.7, 7.11, 7.17, and 7.18.

Many researchers do not like to summarize their results by reporting simply that the null hypothesis either was or was not rejected. Instead, they want their readers to know how much of a discrepancy existed between the data-based p and the level of significance (or between the data-based calculated value and the critical value). In doing this, the researcher's goal is to provide evidence as to how strongly the data challenge H_0. In other words, these researchers want you to know if they beat the level of significance by a wide margin (presuming that H_0 was rejected) or if they just missed beating α (presuming that H_0 was retained).

Consider Excerpts 7.34 and 7.35. In the first of these excerpts, the researchers presented three ps, each of which turned out to be smaller than .000001. In Excerpt 7.35, the researcher reports a single p that turned out equal to .0000003. Instead of simply summarizing their results by saying "$p < .05$," these researchers wanted to show us that they beat the .05 level of significance "by a mile." Further evidence of that presumed motivation for reporting these unusually low p values is the phrase **highly significant** in Excerpt 7.34.

Highly significant

Although p-values like those shown in Excerpts 7.34 and 7.35 are not seen very often in research reports, I can assure you that you will frequently encounter