Chapter 8

Hypothesis-Testing Logic

As you read the chapter, consider the following questions:

8.1 What is the standard error?
8.2 How can we use a population mean and the standard error to test a hypothesis?
8.3 What are the steps in hypothesis testing?
8.4 What is statistical significance and what does it tell us about our hypothesis?
8.5 What types of errors exist in our hypothesis-testing procedure?
8.6 How can I reduce the chance of an error in testing hypotheses?

In the last chapter, we began to discuss how we use z scores to test hypotheses about data using the Unit Normal Table. In this chapter, we will continue this discussion and further define the steps and logic of hypothesis testing using inferential statistics. To better understand our goals in hypothesis testing, let's review the example we used in Chapter 7: You are taking a course in which the instructor has introduced online quizzes to help you study for exams. She has taught this course many times, but this is the first time she has used the online quizzes in the course. You want to know whether the online quizzes are helping the students perform better on the exams (see Photo 8.1). To answer this question, you decide to compare the mean on the final exam for your class to the final exam mean for all previous classes. Your instructor tells you that the mean for the final exams from all previous students is 75 (i.e., $\mu = 75$) with a standard deviation of 3 (i.e., $\sigma = 3$). She also tells you that your class mean on the final exam was 82 (i.e., $X = 82$). In the last chapter, we used this information to calculate the location of your class mean in the population of final exam scores with $\mu = 75$ and $\sigma = 3$. We also used the Unit Normal Table to determine that a score of 82 is quite unlikely in this distribution (i.e., there is only a 0.99% chance of getting a score of 82 or higher in this distribution). However, this only tells us the location of the mean score in the population of all final exam scores. In order to figure out if the mean exam score for your class is significantly higher than the mean scores

Photo 8.1 Research question: Does taking online quizzes help students prepare for exams?

©iStock/LuckyBusiness

for the previous classes, we will need to look at the location of this mean in the distribution of sample means based on the class (i.e., sample) size. That is the general process for testing hypotheses using any of our inferential statistics tests: to find the location of the sample mean in the distribution of sample means for that population and decide if it is an extreme (i.e., unlikely) score in that distribution. We will start with the hypothesis test using *z* scores and this research question (see Photo 8.1) because you are already familiar with *z* scores from the previous chapter.

USING THE NORMAL DISTRIBUTION TO TEST HYPOTHESES

Recall from Chapter 7 that in the online quiz example, we are looking at a normal population of final exam scores to compare with your class mean of 82. Although we were able to find the location of this score in the population of final exam scores, this did not quite tell us what we want to know—whether this is a likely mean from the distribution of sample means for the class size of 30. What we really want to know is whether the mean of 82 is an unlikely mean score for classes that did not have online quizzes available. This will tell us (with a certain probability) that the online quiz class mean is different from the mean scores without the online quizzes. Thus, we need to figure out how likely a mean this is in the distribution of sample means. In order to find out how likely this mean is in the distribution of sample means, we will need to look more closely at that distribution.

The Distribution of Sample Means Revisited

As I have described in previous chapters, the distribution of sample means is a special distribution that contains all the sample means we would get if we were to draw all the possible samples of a specific size from the population and determine each sample's mean score. In other words, the distribution of sample means' scores are the sample means from all possible samples of specific size drawn at random from the population. Recall from Chapter 7 that a z score is calculated based on the distance of the score from the mean divided by the standard deviation of the distribution (i.e., $z = \frac{(X - \mu)}{\sigma}$). Thus, if we want to calculate a z score for a sample mean to determine how likely it is that it came from the distribution of sample means for that population, we will need to know the mean and standard deviation for this distribution. If we were to calculate the mean of all the sample means from samples drawn from the population, we would end up with the population mean μ; thus, the mean of the distribution of samples is equal to the population mean μ. So, if we know the population mean, we will also know the mean of the distribution of sample means.

Things get a bit trickier in determining the standard deviation for the distribution of sample means. Recall from the discussion of sampling that the larger our sample is, the closer we will get to the actual population values in our sample. Thus, sample size will influence the spread of the scores in our distribution of sample means. The larger the sample size (n), the lower the variability in the distribution of sample means because we're getting a better estimate of the population mean with each sample. Thus, the standard deviation for the distribution of sample means is based on σ and n. If we know these values, we can calculate the standard deviation of the distribution of sample means, known as the **standard error**. The standard error represents the sampling error present in our samples (i.e., how much we expect the sample to differ from the population). We can calculate the standard error using the following formula:

Standard error: the estimate of sampling error that is determined from the standard deviation of the distribution of sample means

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

where $\sigma_{\bar{X}}$ is the standard error, σ is the population standard deviation, and n is the sample size for the sample we are looking at. In other words, $\sigma_{\bar{X}}$ is the standard deviation of the distribution of sample means we want to locate our sample mean in.

Finally, we need to consider the shape of the distribution of sample means. If the population is normal, then the distribution of sample means is also normal. This means we can use the Unit Normal Table to find the proportion of scores in different parts of the distribution of sample means, as we did for the distributions in Chapter 7. However, even if the population distribution's shape is unknown (or known to be something other than normal), we can still determine the shape of the distribution of sample means using the **central limit theorem**. The central

Central limit theorem: a mathematical description of the shape of the distribution of sample means that states that for a population with mean μ and standard deviation σ, the distribution of sample means for sample size n will have a mean equal to μ, standard deviation equal to the standard error, and a shape approaching a normal distribution as n becomes very large

limit theorem is a mathematical description of the shape of the distribution of sample means that will allow us to determine if the distribution of sample means is normal in shape. This turns out to be very important in inferential statistics because, in many cases, we do not know the shape of the population. The central limit theorem states that for a population with mean μ and standard deviation σ, the distribution of sample means for sample size n will have a mean equal to μ, standard deviation equal to the standard error, and a shape approaching a normal distribution as n becomes very large (i.e., approaches infinity). In practical terms, the shape of the distribution of sample means is almost exactly normal any time n is greater than 30. Thus, we can use the Unit Normal Table to determine the proportion of scores in different sections of the distribution of sample means whenever our sample size is greater than 30.

Conducting a One-Sample *z* Test

Based on the description of the distribution of sample means in the previous section, you should be able to see how we can use a known population μ and σ to calculate a z score for a sample mean to determine its location in the distribution of sample means. We will need to adjust our z score formula a bit to fit the distribution of sample means:

$$z_{\bar{X}} = \frac{\left(\bar{X} - \mu\right)}{\frac{\sigma}{\sqrt{n}}}$$

The new z score ($z_{\bar{X}}$) for the distribution of sample means will tell us the location of our sample mean within this distribution. If our population is a normal distribution or our sample size is greater than 30, we can then use the Unit Normal Table to determine how extreme a score this is in the distribution of sample means. For our example with μ = 75, σ = 3, and $\bar{X} = 82$, we can calculate $z_{\bar{X}}$:

$$z_{\bar{X}} = \frac{(82 - 75)}{\frac{3}{\sqrt{30}}} = \frac{7}{3/5.48} = \frac{7}{.55} = +12.73$$

This is a different value than the z score of +2.33 that we calculated in Chapter 7 for the location of the score in the population of final exam scores. That is because we have calculated the location of the sample mean in the distribution of sample means here that has a smaller standard deviation than the population of scores.

If you look in a Unit Normal Table, you will see that a z score of 12.73 is very large—in fact, the table given in Appendix B only goes up to a z score +4.00, where only 0.003% of the scores in the distribution are at this z score or higher. Thus, there will be less than 0.003% of the scores in the distribution of sample means at +12.73 or higher. This tells us that our class mean of 82 is a very rare sample mean in the distribution of sample means for final exam means in classes without online quizzes. But how rare does it have to be before we can decide that it probably doesn't belong in this distribution of means? The standard we use in the behavioral sciences for determining this is 5% (i.e., a proportion of 0.05 in the distribution), meaning there's only a 5% chance (or less) that our sample mean came from this distribution. If the percentage of scores for a sample mean z score is at 5% or less for that score or higher (or lower for negative z scores), then we can conclude that it is rare enough for us to decide that it is different

from the means in this distribution. In other words, our class mean of 82 is higher (with less than 0.003% probability) than what is expected for the final exam mean from a class that did not have online quizzes. Therefore, we can conclude that the online quizzes did help students score higher on the final exam. This is the general process we use in inferential statistics. In the next section, we will consider this process as a series of steps to follow to test our hypothesis.

Stop and Think

8.1 Using the standard cutoff probability of 0.05 or lower, find the *z* score in the Unit Normal Table that corresponds to this probability value (Hint: Look for the closest value to 0.05 without going over in the proportion in *Tail* column). Using the *z* score you found as a comparison, if your sample mean had a *z* score of +4.50 for the distribution of sample means, what would this tell you about your hypothesis?

8.2 The Graduate Record Exam (GRE) Verbal test is reported to have a mean score of about 150 and a standard deviation of about 8 (Educational Testing Service, 2016). The GRE prep course you are thinking of taking advertises that it can improve scores on this test. They report that the mean score for this year's class of 100 students scored a mean of 160 on the test. Based on these values, is it worth it to take the class?

LOGIC OF HYPOTHESIS TESTING

As you saw in the example in the previous section, the starting place for our hypothesis-testing procedure is a research question we want to answer. For the example I have been using in this chapter, the question was whether or not online quizzes helped students achieve a higher score on the course's final exam. The one-sample *z* test we conducted helped us determine an answer to this question. Now let's consider another research question: Does memory ability change as one ages? A reasonable hypothesis is that memory ability does change with age. How can we test this hypothesis? We can conduct a study comparing memory for older and young adults and compare their memory scores. A hypothesis-testing procedure using inferential statistics can help us.

The hypothesis-testing procedure can be summarized in five steps:

Step 1: State your research question and make hypotheses about the answer.

Step 2: Set a decision criterion for making a decision about the hypotheses.

Step 3: Collect your sample data.

Step 4: Calculate statistics.

Step 5: Make a decision about the hypotheses.

Table 8.1 provides an overview of these steps that you can refer to as I discuss them further in this chapter. In the next few sections, we will go through each step for our memory and aging research question.

Table 8.1 Overview of the Hypothesis-Testing Steps

Step 1: State Hypotheses	State research question and develop null and alternative hypotheses using literature in the research area.
Step 2: Set Decision Criterion	Set the decision criterion alpha (α) as a probability that the sample mean is a score in the distribution of sample means; consider how your alpha level will influence the chance of Type I and Type II errors in your test.
Step 3: Collect Sample Data	Design your study to test your hypotheses, recruit sample participants/subjects, and collect data on the dependent variables of interest.
Step 4: Calculate Statistics	Summarize data with descriptive statistics; choose an appropriate inferential statistics test and calculate the inferential statistic and corresponding probability (p) value for that statistic.
Step 5: Make a Decision	Compare the statistic p value with α; make a decision to either reject or retain the null hypothesis based on this comparison and then decide if you can accept the alternative hypothesis.

Step 1: State Hypotheses

For this example, we have already stated our research question: Does memory ability change with age? We have also stated our hypothesis about the answer to this question: Memory does change with age. Thus, part of this step is already complete. One thing to note is that the hypothesis we are making is about the population of people, not about our sample. It is the population we want to learn about when we conduct our study. We are only using the sample to represent this population because we cannot test the entire population. Thus, the hypotheses we make are always about a population we want to learn about. We could state our hypothesis as "In the population of people, memory changes with age."

Scientific/Alternative hypothesis: the hypothesis that an effect or relationship exists (or exists in a specific direction) in the population

Null hypothesis: the hypothesis that an effect or relationship does not exist (or exists in the opposite direction of the alternative hypothesis) in the population

In the hypothesis-testing procedure, the hypothesis made by the researcher is usually the **scientific/alternative hypothesis** (it is the alternative hypothesis to an important hypothesis that you will read about next). The scientific or alternative hypothesis is the hypothesis either that an effect of the independent or subject variable exists (for an experiment or quasi-experiment) or a relationship between the variables exists (for a correlational study). For our example, the hypothesis that memory changes with age in the population of people is the scientific/alternative hypothesis that is consistent with predicting that aging causes a change in memory ability for individuals in the population. However, we also must consider a second hypothesis in our test: the **null hypothesis**. The null hypothesis is the opposite hypothesis to the scientific/alternative hypothesis: that an effect or relationship does not exist in the population. The null hypothesis is also

important to state in Step 1 of our procedure because, as you will see later in this chapter, it is the null hypothesis we are directly learning about when we calculate our inferential statistics in Step 4 and make a decision about in Step 5.

For our example, then, we will have two hypotheses to state to complete Step 1: the scientific/alternative hypothesis (denoted by H_1 or sometimes as H_a for *alternative*) and the null hypothesis (denoted by H_0). We can state these hypotheses as

H_0: *In the general population, memory abilities do not change with age* or *In the general population, different age groups have the same mean memory scores*.

The null hypothesis makes the opposite prediction from the alternative hypothesis: *Memory abilities do not change with age.*

H_1: *In the general population, memory abilities change with age* or *In the general population, different age groups have different mean memory scores*.

What we have considered above is called a **two-tailed hypothesis** because we are considering both possible directions of the difference between means in the hypothesis. In other words, our alternative hypothesis does not predict whether younger or older individuals will have higher memory scores; it simply states that the mean scores for younger and older individuals in the population will be *different*. It does not include a prediction about which population will have higher scores. However, for this study, you might find previous studies that indicate that as people age their memory abilities decline. Thus, you could make a directional or **one-tailed hypothesis**. As a one-tailed hypothesis, our alternative hypothesis could be stated as

Two-tailed hypothesis: both directions of an effect or relationship are considered in the alternative hypothesis of the tes

One-tailed hypothesis: only one direction of an effect or relationship is predicted in the alternative hypothesis of the test

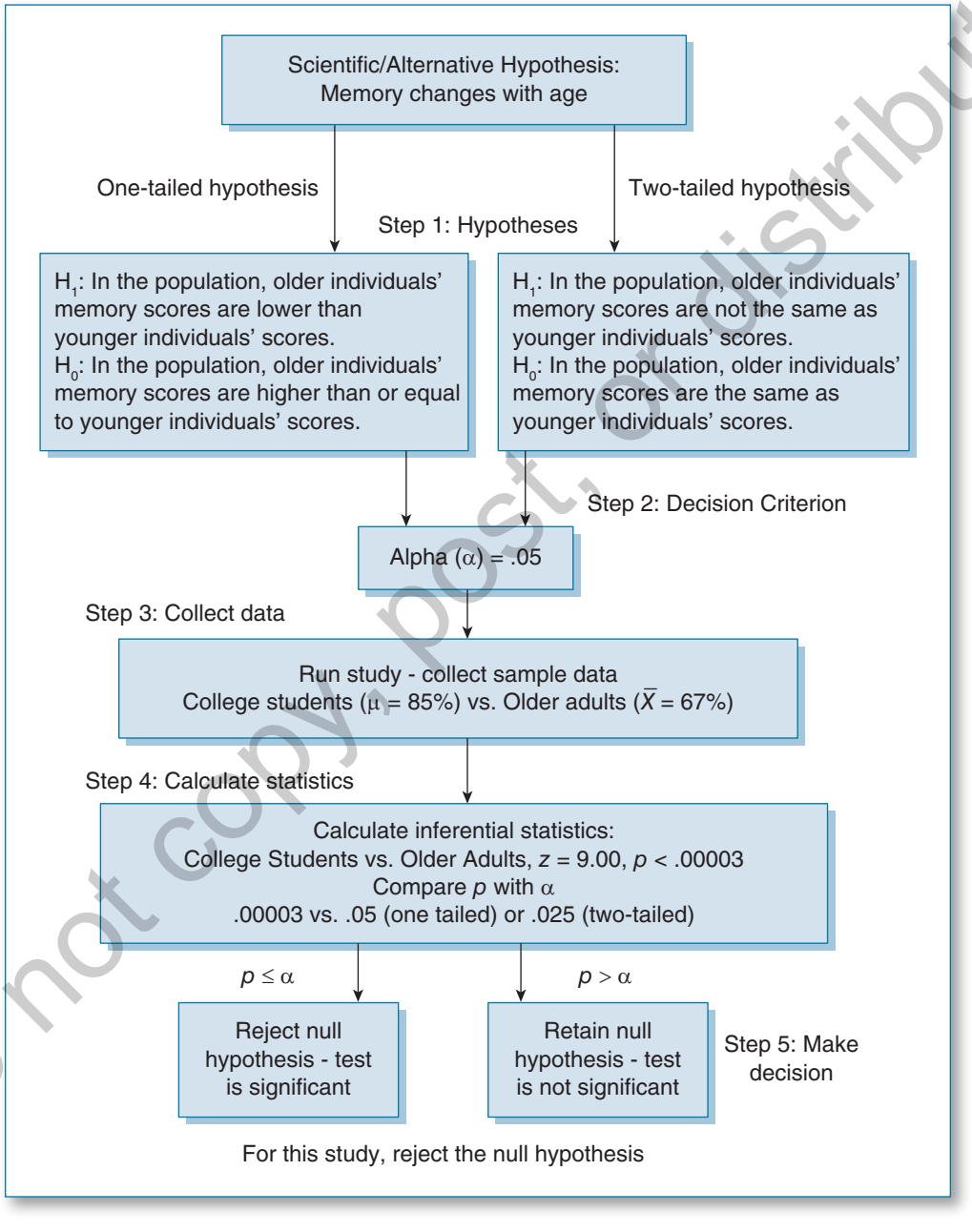
H_1: *In the general population, older individuals have lower memory scores than younger individuals*.

We could also make the opposite prediction (e.g., H_1: *In the general population, older individuals have higher memory scores than younger individuals*), but the first hypothesis stated above is more likely to be consistent with the results of previous studies. For this alternative hypothesis, our null hypothesis must include any other possible outcomes, so our null hypothesis is

H_0: *In the general population, older individuals have higher memory scores than younger individuals or the memory scores of the two age groups are the same*.

For a one-tailed hypothesis, the null hypothesis contains the predictions of no effect or relationship *and* the effect or relationship in the direction opposite to that predicted in the alternative hypothesis. See the top portion of the flowchart in Figure 8.1 for a comparison of one-tailed and two-tailed hypotheses for this study.

Figure 8.1 Flowchart of the Steps in Hypothesis Testing for Memory and Aging Example



One-tailed hypotheses are typically made only when a researcher has a logical reason to believe that one particular direction of the effect will occur. Thus, one-tailed hypotheses are often made when the other direction of the effect logically should not occur or does not answer the research question. They may also be made when the literature review of an area indicates that one direction of the effect has been shown consistently over a number of research studies.

Stop and Think

8.3 For each of the following statements, indicate if a one- or two-tailed test is most appropriate:

- Taking aspirin reduces the chance of a heart attack.
- Quizzing yourself before a test will increase your test score compared with simply rereading your notes.
- Completing a puzzle under a time constraint will affect your accuracy.
- Sleep affects depression.

8.4 For each statement above, state the alternative and null hypotheses.

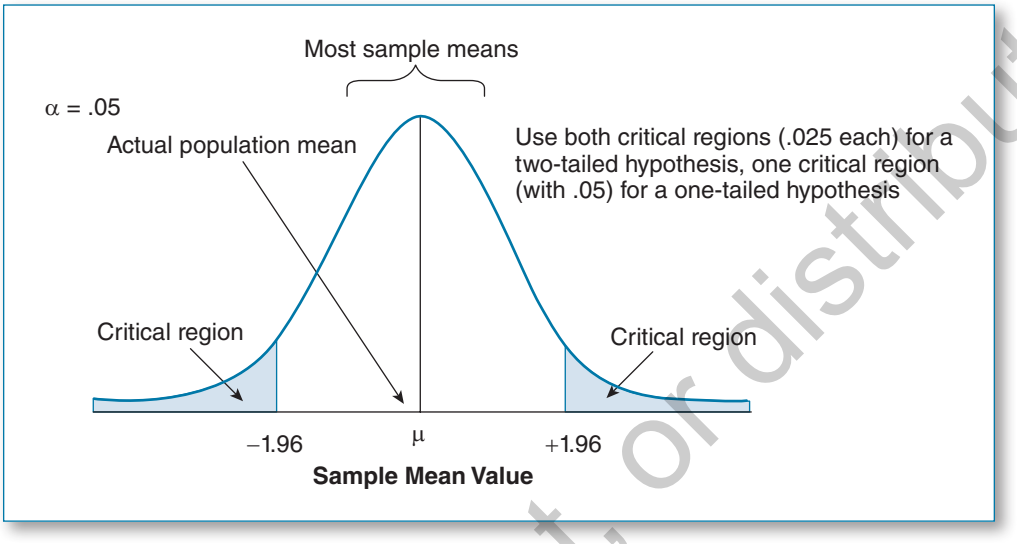
Step 2: Set Decision Criterion

Now that we have completed Step 1 and have stated our alternative and null hypotheses, we can move on to Step 2 and set our decision criterion. Let's consider what we are doing when we set this value.

In inferential statistics tests, we are calculating a value that tells us the location of the sample mean in a distribution, just as we did with z scores in the last chapter. But the values we have from the sample are descriptive statistics in the form of a mean score (or, in some cases, a value indicating the strength of a relationship—more will be described about this type of test in Chapter 15). Thus, we are looking at the location of a sample mean in the distribution of sample means as we did earlier in this chapter. The decision criterion value marks off a portion of this distribution that represents the most extreme scores. It is also called the **alpha level**, because the criterion probability is denoted by the Greek letter α. As described earlier, the criterion is typically set at 0.05 (i.e., 5% of the most extreme scores) in research in the behavioral sciences. This means that we want to look at the proportion in tail from the Unit Normal Table that sets off this part of the distribution (and its corresponding z score—see Stop and Think 8.1 earlier in this chapter). Figure 8.2 illustrates this portion of the distribution of sample means. The shaded areas are the portion of the distribution that are considered the most extreme scores equal to the decision criterion. If we have a two-tailed hypothesis, we

Alpha level: the probability level used by researchers to indicate the cutoff probability level (highest value) that allows them to reject the null hypothesis

Figure 8.2 Distribution of Sample Means When the Null Hypothesis Is True



must consider both shaded tails of the distribution so the criterion proportion is split in two (i.e., 0.025 in the upper tail and 0.025 in the lower tail). If we have made a one-tailed hypothesis, then we only need to consider one shaded tail that contains the entire proportion—which tail depends on the direction we have predicted: The upper tail if we predict the mean will be higher and the lower tail if we predict that the mean will be lower. The shaded portion is known as the **critical region** of the distribution because it is the part we are looking at to see if we can reject the null hypothesis (one of our possible decisions in Step 5).

Critical region: the most extreme portion of a distribution of statistical values for the null hypothesis determined by the decision criterion (i.e., alpha level—typically 5%)

Notice in Figure 8.2 that the distribution of sample means corresponds to the sample means when the null hypothesis is true. This is the distribution we will consider in our hypothesis test. We will locate our test statistic in this distribution (is it in the shaded portion(s) or not?) and make a decision about the null hypothesis depending on whether our sample mean is extreme for this distribution or not. This is because the evidence provided by the inferential test is the likelihood of obtaining the data in the study if we assume the null hypothesis is, in fact, true. That is what the inferential test focuses on: What is the chance of obtaining the data in this study when the null hypothesis is true? If the chance is fairly high, then there is no evidence to reject the null hypothesis. If the chance is very low, then the researcher takes that as evidence against the null hypothesis, rejects it, and supports the alternative hypothesis that there is an effect or relationship. It is important to set your decision criterion before you begin the study so that you have a clear basis for making a decision when you get to Step 5. If

we wait to choose our alpha level, we might be tempted to make the wrong decision because our probability value seems low enough. This could result in an error in our hypothesis-testing procedure. I will discuss these errors and how we use our decision criterion to make a decision further as we consider Step 5: Making a decision.

Step 3: Collect Sample Data

In Step 3, we are ready to design our study to test our hypothesis, recruit a sample, and collect our data. This process was discussed in more detail in Chapters 1 and 2, where we considered where data come from. This might be a good time to review the summaries of those chapters.

For our example (looking at whether memory changes with age), we might design a study looking at memory abilities for college students and older adults. For example, suppose we know the population mean and standard deviation of a standardized memory test for college students because many of them have taken it as they participated in research studies. We might then design a study where we recruit a sample of older adults to complete the memory test to see how their mean score on the test compares. Our sample of older adults represents the population of all older adults (e.g., over the age of 60). We can then consider where our sample mean falls in the distribution of sample means for college students to see if the older adults' mean score is an extreme score in this distribution based on our decision criterion. If it is, we can decide to reject the null hypothesis (H_0: Memory does not change with age) and conclude that the older adults' memory scores appear to be part of a different distribution with a lower (or higher) mean. See Figure 8.1 for an overview of our study.

Step 4: Calculate Statistics

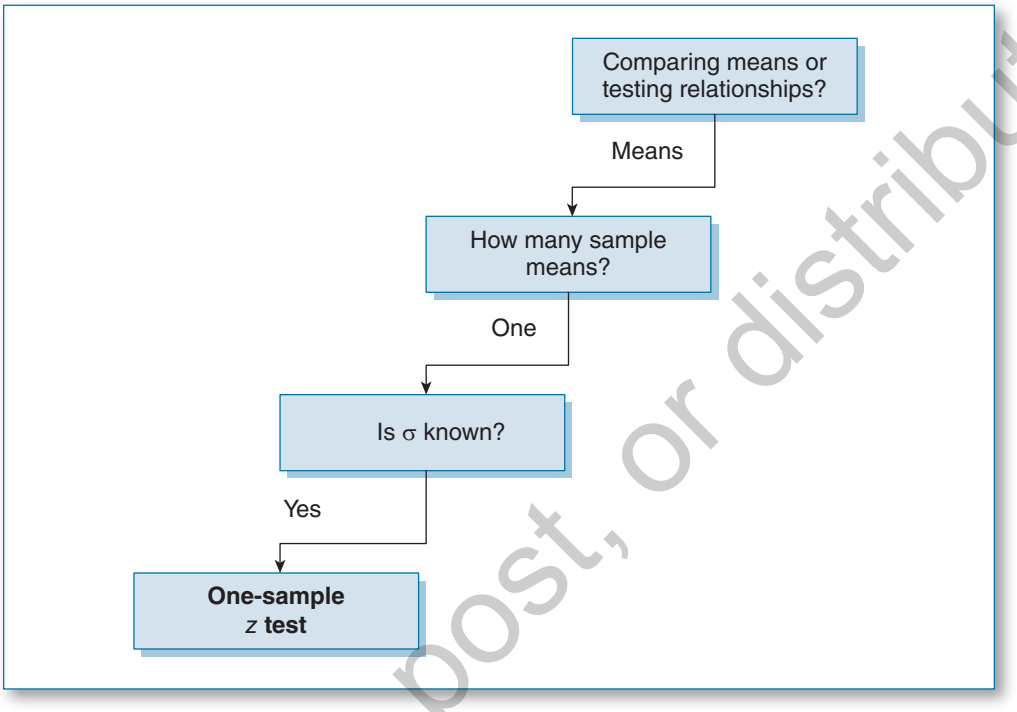
As Figure 8.1 shows, the known population mean μ for the standardized memory test in our study is 85% and our sample mean $\bar{X}$ is 67%. What we want to know from our hypothesis test is whether 67% is different enough from 85% to conclude that older adults show different memory abilities from the young adults. To determine this, we will need to calculate an inferential statistic. If we know the standard deviation for the memory test for the population of college students, we can use our one-sample *z* test. Figure 8.3 shows the relevant portion of the inferential statistics flowchart for a one-sample *z* test. Suppose we know that the standard deviation is $\sigma = 20$ and that the population of memory scores is a normal distribution. With this information, we're ready to calculate the $z_{\bar{X}}$. For this example, the calculation is

$$z_{\bar{X}} = \frac{(85-67)}{\frac{20}{\sqrt{100}}} = \frac{18}{20/10} = \frac{18}{2} = +9.00$$

We can use the Unit Normal Table in Appendix B to find the probability value (also known as a ***p* value**) associated with this *z* score. We have already seen earlier in this chapter that the highest value in the table is 4.00 with a *p* value of 0.00003. So we know that the *p* value for 9.00 will be

***p* value:** probability value associated with an inferential test that indicates the likelihood of obtaining the data in a study when the null hypothesis is true

Figure 8.3 Portion of the Statistical Test Decision Flowchart for a One-Sample *z* Test



lower than 0.00003. This *p* value is what we need before we move on to Step 5 and make a decision.

However, before we move on to our last step, let's consider what would happen if we did not know the population standard deviation value, as is often the case in a research study. Without the σ value, we cannot calculate the standard error as we have done for our one-sample *z* test. Instead, our standard error will need to be calculated from an estimate of the population σ. The best guess we can make for this value is the standard deviation in our sample, because the sample values are meant to represent the ones we would find in the population. Thus, our standard error formula would be

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

This calculation of the standard error changes the test statistic calculation and becomes a new statistic known as *t* (instead of *z*). We could then conduct a one-sample *t* test instead of a *z* test to test our hypothesis. We will discuss this test in Chapter 9.

Things get even more complicated if we don't know the population μ either. In this case, we would have to represent both age groups' populations in our study with separate samples: a sample of college students and a sample of older adults. We can still calculate a *t* statistic, but it will be a new inferential test called an independent samples *t* test. This test is discussed in Chapter 11. Later chapters in this text will also consider cases where there are more than two

samples in a study with an inferential test known as analysis of variance (ANOVA). However, in each inferential test, we are using the same hypothesis-testing procedure with the five steps described in this chapter.

All the inferential statistics described in this text use a calculation that relies on differences or relationships seen in the data with an estimate of sampling error divided out of these differences or relationships. Thus, the test statistic is a value representing the differences or relationships seen in the sample data corrected for chance differences or relationships that could be seen due to relying on samples (instead of whole populations). If the test statistic is a low value, then the differences/relationships seen in the sample are likely due to chance sampling factors. If the test statistic is a high value, then the differences/relationships seen in the sample are likely due to actual differences/relationships that exist in the population. These outcomes correspond to the decisions we can make in the test.

Step 5: Make a Decision

In our last step, we need to make a decision about the null hypothesis based on how unlikely our sample mean is in the distribution of sample means that would exist when the null hypothesis is true. We will either find evidence against the null hypothesis and reject it or fail to find this evidence and retain it. How unlikely does a sample mean have to be before we decide it did not come from this distribution of sample means and reject the null hypothesis? This is where our alpha level comes into play. It is the proportion of our distribution of sample means that falls into the critical regions. Figure 8.2 shows these regions for $\alpha = 0.05$, our standard alpha level (bounded by the z scores of ± 1.96). It is the highest probability that a sample mean came from this distribution of sample means that we will accept as evidence against the null hypothesis. It is set by the researcher at a low value (such as 0.05) to allow rejection of the null hypothesis only when it is very unlikely that the sample mean came from the distribution of sample means for the null hypothesis. *In other words, the decision to reject or not reject the null hypothesis is based on the probability of obtaining the data in the sample when the null hypothesis is true.* A low alpha level helps us avoid an error in our hypothesis-testing procedure.

The probability of a sample mean appearing in the distribution of sample means is compared with the alpha level in an inferential test. Remember that most sample means occur near the actual population mean, so if the probability is high that the sample mean came from this distribution, then it is more likely that the null hypothesis is true, given the data we collected. If the p value is equal to or lower than the alpha level, then we can reject the null hypothesis as unlikely to be true. If the p value is higher than the alpha level, we cannot reject the null hypothesis, as it might be true (however, the test does *not* provide evidence *for* the null hypothesis, only against it).

Now, consider our example once again. Figure 8.1 shows the z and p values we determined in the previous section for our sample data. In Step 5, we compare the p value with our α level. For this example, we compare 0.0003 versus half of 0.05, which is 0.025 (we have two critical regions with a two-tailed test, so we divide α in half here). If our p value is lower than α, as it is here, we can reject the null hypothesis that memory does not change with age. If we reject this hypothesis, then we will have supported the alternative hypothesis that it does change with age. Looking back at the means, the older adults showed a lower mean score than the population of college students. Thus, we can conclude for this study that memory declines with age.

Stop and Think

8.5 Suppose the *z* score we had calculated for our example in this chapter was +2.30. In this case, what would our *p* value be? With an $\alpha = 0.05$, what decision would you make for this example?

8.6 Consider another example: An anxiety questionnaire is known to have $\mu = 50$ and $\sigma = 5$ in the general population. A sample of 50 college students is given the questionnaire after being asked to prepare a 5-minute speech on a topic of their choosing to see if this task elevates their anxiety level from what is expected based on the population mean score on the questionnaire. The sample mean is $\bar{X} = 58$.

- State the alternative and null hypotheses for this study. Is the alternative hypothesis one-tailed or two-tailed?
- Calculate the one-sample *z* score for this sample. What is the probability of obtaining this *z* score when the null hypothesis is true?
- For an alpha level of 0.05, what decision would you make for this study? What can you conclude about the speech preparation task and its effect on anxiety level?

TYPES OF HYPOTHESIS-TESTING ERRORS

One thing that is important to keep in mind is that we are using probability to make our decision in the hypothesis-testing procedure. The test statistic's *p* value tells us the chance of obtaining that statistical value (using the sample data to calculate it) when the null hypothesis is true. Even if we reject the null hypothesis as our decision in the test, there is still a small chance that it is, in fact, true. Likewise, if we retain the null hypothesis as our decision, it is still possible that it is false. These are the kinds of errors that can be made in our hypothesis test, and they are always possible because we are relying on a probability (not a certainty) to make a decision.

Table 8.2 illustrates the different possible outcomes of a hypothesis test. The columns represent the reality for the population being studied: Either the null hypothesis is true and there is no effect/relationship (e.g., older adults do not have different memory scores and their mean memory score is the same as the population mean for the younger adults), or the null hypothesis is false and there is an effect/relationship (e.g., older adults do have different memory scores and their mean memory score is not the same as the population mean for the younger adults). When a hypothesis test is conducted, the researcher does not know whether the null hypothesis is true or false. However, as described in the previous section, the inferential test is conducted to look for evidence that the null hypothesis is not true. If that evidence is found, the researcher decides that the null hypothesis is false and rejects it. This is the outcome represented in the first row of Table 8.2. If, in fact, the null hypothesis is false, the researcher has made a correct decision in the test, because the decision matches the reality about the null hypothesis. However, it is possible to make the wrong decision. Thus, the outcome to reject the null hypothesis in the first row under the column where the null hypothesis is actually true is

an error. The researcher's decision does not match the reality for the null hypothesis. This is called a **Type I error** and indicates that the researcher has rejected the null hypothesis when it is really true (e.g., we find in our study that older adults have a different memory score mean from the young adults, but in the population of older adults, they do not have a different mean score). The chance of making a Type I error is determined ahead of time by the researcher when an alpha level is chosen. Thus, in tests with $\alpha = 0.05$, there is a 5% chance of making a Type I error.

Type I error: an error made in a hypothesis test when the researcher rejects the null hypothesis when it is actually true

Table 8.2 Possible Outcomes of a Statistical Test

Decisions	*Null Hypothesis Is Actually False*	*Null Hypothesis Is Actually True*
Reject the null hypothesis	Correct decision!	Type I error
Fail to reject the null hypothesis	Type II error	Correct decision!

The second row in Table 8.2 illustrates test outcomes for the other decision that can be made in the significance test: retaining or failing to reject the null hypothesis, which occurs when evidence against it is not found in the test. A correct decision is made in the test when the decision is to fail to reject the null hypothesis and this hypothesis is really false (bottom right box). However, another type of error, called a **Type II error**, can be made when the null hypothesis is not rejected but is actually false (e.g., we find in our study that older adults do not have a different mean memory scores than young adults, but in the population, older adults do have a different mean score from the younger adults). This means that an effect or relationship exists in the population but was not detected in the data for the sample. The chance of a Type II error is more difficult to determine. There are several factors that can influence the probability of a Type II error, including the alpha level chosen, the size of the effect or relationship, and the sample size in the study. The researcher can lower the chance of a Type II error by using an optimal sample size and making sure that the study is designed to maximize the effect or relationship being studied. By keeping the Type II error rate low, you are increasing the **power** of your hypothesis test to detect an effect or relationship that actually exists. Thus, it is important to keep Type II errors in mind as you design your study to conduct a powerful test of the hypothesis.

Type II error: an error made in a hypothesis test when the researcher fails to reject the null hypothesis when it is actually false

Power: the ability of a hypothesis test to detect an effect or relationship when one exists (equal to 1 minus the probability of a Type II error)

Predicting the Null Hypothesis

As mentioned above, in many cases, the alternative hypothesis is also the researcher's hypothesis. The researcher predicts that an effect or relationship exists in the population. However, in some cases, the researcher may wish to predict that an effect or relationship does not exist in the population. Is this an appropriate thing for a researcher to do when using inferential statistics? Many would argue that it is not appropriate for a researcher to predict the null hypothesis because significance tests do not provide evidence for the null hypothesis. In fact, most papers that are published in psychological journals describe studies that showed significant results (Francis, 2013), because it can be difficult to draw strong conclusions from studies that do not show significant results. However, power analyses can be used to estimate the chance of a Type II error occurring and the null hypothesis being falsely retained. While any single study with nonsignificant results is not sufficient to provide support for the null hypothesis, a series of studies that have a reasonable level of power to detect effects (80% or higher is the generally accepted level; Cohen, 1988) that all show the same nonsignificant results may provide some support for the null hypothesis. Thus, if researchers want to predict the null hypothesis, they must be prepared to conduct several studies in order to obtain some support for their hypothesis. Many researchers (e.g., Francis, 2013; Greenwald, 1975) also argued that a bias against the null hypothesis can result in researchers ignoring studies that do not find significant effects (which can be caused by the bias against publishing them). In addition, because it is important that theories of behavior can be falsified, it is sometimes necessary to predict the null hypothesis in order to truly test a theory. Finally, in order to get around this issue, several researchers (e.g., Cohen, 1990, 1994; Loftus, 1993) have suggested alternatives to the hypothesis-testing procedure described in this chapter as a means of interpreting data.

STATISTICAL SIGNIFICANCE

One concept not yet discussed in this chapter is what it means for a hypothesis test to be a **significant test**. If the *p* value for the test statistic is less than or equal to the alpha level, the test is said to be significant. In other words, a significant inferential test means that the null hypothesis can be rejected, the alternative hypothesis has been supported, and the researcher can conclude that there is an effect or relationship for the data in the current study. This means that hypothesis tests where the decision is to reject the null hypothesis are reported as *significant* tests.

Significant test: the *p* value is less than or equal to the alpha level in an inferential test, and the null hypothesis can be rejected

Note that this term does not mean *important* in the way this term is typically used outside of statistics. A hypothesis test can be significant in the statistical sense without being very important at all. In fact, with a large enough sample size, it is often easy to obtain a significant difference between groups that is based on subject differences unrelated to the study or a significant statistical relationship between factors that are not related in any meaningful or causal way (e.g., amount of rainfall in a month and number of people buying soda in that month). So be aware that statistical significance may not mean that a result tells us something important about behavior.

Stop and Think

8.7 For each description below, indicate the situation: correct decision, Type I error, or Type II error.

- An effect of amount of sleep on mood exists, but the results of the study were not significant.
- A relationship between early reading and later academic achievement exists, and the results of the study were significant.
- An effect of caffeine on work productivity does not exist, but the results of the study were significant.

Calculation Summary

Standard error: Population standard deviation divided by the square root of the sample size

One-sample *z* test: Sample mean minus the population mean, divided by the standard error

CHAPTER SUMMARY

8.1 What is the standard error?

Standard error is a measure of variability in the distribution of sample means that takes sample size into account. It provides an estimate of sampling error for our hypothesis test.

8.2 How can we use a population mean and the standard error to test a hypothesis?

The one-sample *z* test and *t* test both consider the difference between a measured sample mean and a known population mean with our estimate of sampling error in the form of the standard error removed in the calculation of the test statistic. This statistical value is then used to determine the probability (*p*) value of getting our sample mean in the distribution of sample means for the population. A low *p* value indicates an extreme score for the distribution, making it possible to reject the null hypothesis that the sample mean is from this distribution.

8.3 What are the steps in hypothesis testing?

The five steps of hypothesis testing take us through the procedure described (see Table 8.1 for an overview of the procedures by step):

Step 1: State hypotheses.

Step 2: Set decision criterion.

Step 3: Collect sample data.

Step 4: Calculate statistics.

Step 5: Make a decision.

8.4 What is statistical significance and what does it tell us about our hypothesis?

A statistically significant hypothesis test is one where we have decided to reject the null hypothesis based on the evidence found in the sample data against it. If we reject the null hypothesis, we can accept the alternative hypothesis, which is typically the hypothesis we have made as researchers.

8.5 What types of errors exist in our hypothesis-testing procedure?

Hypothesis-testing procedures can result in either Type I or Type II errors, depending on the decision we make in Step 5. If we reject the null hypothesis in error (i.e., the null hypothesis is actually true), then we are making a Type I error. If we retain the null hypothesis in error (i.e., the null hypothesis is actually false), then we are making a Type II error.

8.6 How can I reduce the chance of an error in testing hypotheses?

Setting our decision criterion alpha to a low value will reduce the chance of a Type I error. Increasing our sample size and/or effect size will increase power, which is the same as reducing the chance of a Type II error.

THINKING ABOUT RESEARCH

A summary of a research study in psychology is given below. As you read the summary, think about the following questions:

1. Identify the five steps of hypothesis testing in this article description. Indicate what was determined at each step. State what you think are the null and alternative hypotheses for this study.
2. Why do you think the participants were blindfolded in this this study? What source of possible bias does this control for?
3. What is the likely reason the authors used a one-sample *t* test instead of a *z* test to analyze their results?
4. Based on the description of this study, what population mean μ do you think the authors compared their sample mean with?
5. Based on the information in this chapter, what formula do you think they used to calculate standard error in their inferential test?

Wagman, J. B., Zimmerman, C., & Sorric, C. (2007). "Which feels heavier—a pound of lead or a pound of feathers?" A potential perceptual basis of a cognitive riddle. *Perception, 36,* 1709–1711.

Purpose of the Study. Wagman et al. investigated the perceptual causes of why people answer the riddle about the respective heaviness of equal masses of lead and feathers as if one feels heavier. In their study, participants were asked to hold a box of lead bearings and a box of feathers of equal weight and indicate which box felt heavier. Based on the size/weight illusion (where larger objects are expected to be heavier regardless of actual mass) as applied to mass distribution of objects, the researchers predicted that participants would select the box of lead at a different rate than chance due to the different mass distributions of lead and feathers within the boxes.

Method of the Study. Participants included 23 blindfolded students. Each participant completed 20 trials, in which they held one box in their palm and then a second box in the same palm. One box held lead pellets and one box held feathers. The objects were secured within the box to keep them from creating any sound stimuli that could be used to make judgments. They were then asked to indicate which box felt heavier, the first box or the second box. Lead and feather boxes were presented in a random order on each trial. If participants could not determine a difference in heaviness between the boxes, chance performance (10 responses for the 1st box and 10 responses for the 2nd box) was expected.

Results of the Study. To test their hypothesis that the boxes did not feel equally heavy to the participants, the researchers conducted a one-sample *t* test on the number of trials (out of 20) that each participant reported the box of lead felt heavier. The mean number of times participants reported the box of lead felt heavier was 11.12 times out of the 20 trials. They found that this sample differed significantly from chance with a calculated *t* value of 2.64 with a *p* value of 0.015. APA style for reporting the statistic is $t(22) = 2.64$, $p = 0.015$.

Conclusions of the Study. The results of the study suggest that objects with equal mass can be perceived at different heaviness due to the difference in mass distribution within a held box of lead and feathers.

TEST YOURSELF

1. The standard error is _____.
   a. determined from the population standard deviation and the sample size
   b. is an estimate of the sampling error
   c. the variability of the distribution of sample means
   d. all of the above
2. The alpha level is the _____.
   a. chance that the null hypothesis is true
   b. the chance that the null hypothesis is false
   c. the decision criterion for rejecting the null hypothesis set by the researcher
3. The researcher's hypothesis is typically the opposite of the _____ hypothesis.
   a. alternative
   b. null
   c. population
4. The hypothesis-testing procedure can provide evidence against the _____.
   a. null hypothesis
   b. alternative hypothesis
   c. distribution of sample means standard error

5. The possible decision(s) in Step 5 of the hypothesis-testing procedure are to _____.
   a. reject the null hypothesis
   b. accept the null hypothesis
   c. retain the null hypothesis
   d. only (a) and (b)
   e. only (a) and (c)
6. The hypothesis-testing procedure will tell us the probability that the null hypothesis is true.
   a. True
   b. False
7. The best estimates of the population mean and standard deviation when these values are not known are the mean and standard deviation values in the sample.
   a. True
   b. False
8. The inferential test statistic represents the difference between means with sampling error removed.
   a. True
   b. False
9. Explain why errors are always possible during hypothesis testing.
10. You pulled several all-nighters last semester to study for your final exams. You want to know if staying up all night hurt your exam performance so you will know if it is worth it to stay up all night to study. You calculate the mean score for all of the finals you have ever taken in college (your exam population μ) and find that $\mu = 87\%$ with $\sigma = 5\%$. Assume you know that this population of scores has a normal distribution. You use as your sample the mean score on all five of the final exams you took last semester, $X = 83\%$.
    a. What are the null and alternative hypotheses for this example? Is this a one- or two-tailed test?
    b. Use a one-sample *z* test to determine if your all-nighters hurt your performance.
    c. Suppose that in reality, all-nighters do hurt your performance on exams. In this case, what type of decision has occurred in your test: correct decision, Type I error, or Type II error?
11. What is the easiest way to reduce Type II errors? What problem does this method of reducing Type II errors create? (Hint: Consider statistical significance vs. practical significance.)