6

Confidence Intervals

In this chapter, you'll enter the world of inferential statistics. As you get started, think back over the material you've covered so far. For example, you've already learned about the mean, standard deviation, samples, populations, statistics, parameters, and sampling error. You've also been introduced to *Z* scores, the Table of Areas Under the Normal Curve and the Central Limit Theorem. Now it's time to bring all those elements together.

As you begin to bring the different elements together, there's a chance you'll begin taking advantage of other resources—Web sites, other texts, or additional learning aids. As before, let me encourage you to do that. Should you take that path, however, let me also remind you again about the noticeable differences that often emerge when it comes to the matter of symbolic notation. Different statisticians may use different symbols for the same concept—that's just the way it is, and there's no reason to let those little bumps in the road throw you.

Having said that, here's what lies ahead. The general application we'll cover in this chapter is known as the *construction of a confidence interval*. More specifically, we're going to deal with the construction of a confidence interval for the mean and the construction of a confidence interval for a proportion. We'll begin with the confidence interval for the mean.

Before We Begin

By now you should be adequately armed to jump into the world of statistical inference. You have the important concepts under your belt, but your patience is probably wearing thin. Therefore, there's no reason to waste too much time, except to offer up one of my favorite statistical sayings: *We really don't give a hoot about a sample, except to the extent that it tells us something about the population.* In fact, that's what the field of inferential statistics is all about—samples really aren't of interest to us, except that they provide us information that we can use to make inferences about populations. That is an extremely simple but important notion, so allow me to repeat it: *We really don't give a hoot about a sample, except to the extent that it tells us something about the population.* Simply put, you're getting ready to apply that adage. You're going to use some information gained from a sample so you can make some statements about a population.

Confidence Interval for the Mean

Let's suppose we want to estimate the mean of a population (μ) on the basis of a sample mean ($\bar{X}$). By now you should know we can't simply calculate a sample mean ($\bar{X}$) and assume that it equals the mean of the population (μ). A sample mean *might* equal the mean of the population, but we can't assume that it will. We can't do that because there's always the possibility of sampling error. Because our ultimate aim is to estimate the true value of the population mean (μ), we'll have to use a method that takes into account this possibility of sampling error.

We'll use our sample mean as the starting point for our estimate. Then we'll build a band of values, or an *interval,* around the sample mean. To do this, we'll add a certain value to our sample mean, and we'll subtract a certain value from our sample mean. When we're finished, we'll be able to assert that we believe the true mean of the population (μ) is between this value and that value. For example, we'll eventually be in a position to make a statement such as "I believe the mean age of all students at the university (μ) is somewhere between 23.4 years and 26.1 years." This statement expresses a **confidence interval for the mean** of a population based on a sample mean.

Let's think about that for a moment. This method allows us to express an interval in terms of two values. The two values are the upper and lower limits of the interval—an interval within which we *believe* the mean of the population is found. We may be right (our interval may contain the true mean of the population), or we may be wrong (our interval may not contain the true mean of the population). Even though there's some uncertainty in our estimate, we'll know the probability, or likelihood, that we've made a mistake. That's where the term *confidence* comes into play—we'll have a certain level of confidence in our estimate. What's more, we'll know, in advance, *how much* confidence we can place in our estimate.

As it turns out, there are two different approaches to the construction of confidence intervals for the mean. One approach is used when we know the value of the population standard deviation (σ), and another approach is used when we don't know the value of the population standard deviation (σ). The second approach is used more frequently, but it's the first approach that really sets the stage with the fundamental logic. For that reason, we'll begin with confidence intervals for the mean with σ known; after that, we'll turn to confidence intervals for the mean with σ unknown. Once you've mastered the logic of the first approach, the move to the second application will be easier.

LEARNING CHECK

Question: What is a confidence interval for the mean?

Answer: It's an interval or range of values within which the true mean of the population is believed to be located.

Confidence Interval for the Mean With σ Known

We'll begin our discussion of confidence intervals with a somewhat unusual situation—one in which we're trying to estimate the mean of a population when we already know the value of σ (the standard deviation of the population). Why, you might ask yourself, would we have to estimate the mean of a population if we already know the value of the standard deviation of the population? Wouldn't we have to know the mean of the population to calculate the standard deviation? Those are certainly reasonable questions. Although situations in which you'd know the value of the standard deviation of the population are rare, they do exist.

Some researchers, for example, routinely use standardized tests to measure attitudes, aptitudes, and abilities. Personality tests, IQ tests, and college entrance exams are often treated as having a known mean and a known standard deviation for the general population (σ). The Scholastic Aptitude Test (SAT), for example, has two parts—math and verbal. Each part has been constructed or standardized in a way that yields a mean of 500 and a standard deviation of 100 for the general population of would-be college students. An example like that—one involving some sort of standardized test—is a typical one, so that's a good place to start.

An Application

Let's assume that we're working for the XYZ College Testing Prep Company—a company that provides training throughout the nation for students preparing to take the SAT college entrance examination. Part of our job is to monitor the success of the training. Let's assume we have collected information from a sample of 225 customers—225 students from throughout the nation who took our prep course—telling us how well they did on each section of the SAT. Let's say that we're only interested in the math scores right now, so that section will be our focus.

Now, let's say that the results indicate a sample mean ($\bar{X}$) of 606. In other words, the mean score on the math section for our 225 respondents was 606. The question is how to use that sample mean to estimate the mean score for all of our customers (the population). We know that we can't simply assume that the sample mean of 606 applies to our total customer base. After all, it's just one sample mean. A different sample of 225 customers might yield a different sample mean.

We can, however, use the sample mean of 606 as a starting point, and we can build a confidence interval around it. In other words, we'll start by treating the value or 606 as our best guess, so to speak. The true mean of the population (the population of our entire customer base—let's say 10,000 customers) may be above or below that value, but we'll start with the value of 606 nonetheless. After all, with random sampling on our side, our sample mean is likely to be fairly close to the value of the population mean. At the same time, though, we know that our value of 606 may not equal the true mean of the population, so we're going to build in a little cushion for our estimate. The question is, How do we establish the upper and lower values—how do we build in the cushion?

We build the cushion by adding a certain value to the sample mean and subtracting a certain value from the sample mean (don't worry right now about how much we add and subtract—we'll get to that eventually). When we add a value to the sample mean, we establish the upper limit of our confidence interval; when we subtract a value from the mean, we establish the lower limit of the confidence interval.

LEARNING CHECK

Question: In general, how is a confidence interval for the mean constructed?

Answer: A sample mean is used as the starting point. A value is added to the mean and subtracted from the mean. The results are the upper and lower limits of the interval.

Given what our purpose is, along with the notion that we're going to use our sample mean as a starting point, you shouldn't be terribly confused when you look at the formula for the construction of a confidence interval. After all, it's simply a statement that you add something to your sample mean and you subtract something from your sample mean. The formula that follows isn't the complete formula, but take a look at it with an eye toward grasping the fundamental logic.

$$\text{Confidence Interval, or CI} = \text{Sample Mean} \pm Z\,(\,?\,)$$

The sample mean will be the starting point.
A value will be added to the mean and subtracted from the mean.

It's clear from the formula that we're going to be working with a sample mean ($\overline{X}$), and we'll be using a Z value, but two questions still remain: Why the Z value, and what does the question mark represent?

Reviewing Z Values

To answer those questions, let's start by reviewing something you learned earlier about Z values (see Chapter 4 if you're in any way unclear about Z values). Think back to what you learned about a Z value in relationship to the normal curve—namely, that a Z value is a point along the baseline of the normal curve. Think about the fact that Z values are expressions of standard deviation units.

To understand why this is important in the present application, let me ask you to shift gears for just a moment. We'll eventually get back to our example, but for the moment, put that aside. Instead of thinking in terms of a sample of SAT scores, assume that you're working with a large *population* of scores on some other type of test. For example, think in terms of a large number of students who took a final exam in a chemistry course. Assume the scores are normally distributed, with a mean of 75 and a standard deviation of 8.

Since the distribution is normal, 95% of the scores would fall between 1.96 standard deviations above and below the mean. That's something you learned when you learned about the normal curve and the Table of Areas Under the Normal Curve. If 95% of the cases fall between 1.96 standard deviations above and below the mean, it's easy to figure out the actual value of the scores that would encompass 95% of the cases. All you'd have to do is multiply the standard deviation of your distribution (8) times 1.96. You'd add that value (1.96×8) to the mean, and then you'd subtract that value from the mean. That would be the answer to the problem. Here's how the process would play out.

- Assuming that a large number of scores on a final exam are normally distributed, you'd expect 95% of the scores in your distribution to fall between ±1.96 standard deviations from the mean (that is, 1.96 standard deviations above and below the mean).
- The mean = 75
- The standard deviation = 8
- $1.96 \times 8 = 15.68$
- $75 - 15.68 = 59.32$
- $75 + 15.68 = 90.68$
- Therefore, 95% of the scores would be found between the values of 59.32 and 90.68.

To grasp the point more fully, consider these additional examples, assuming a normally distributed population in each case.

With a mean of 40 and a standard deviation of 5:

What values would encompass 95% of the scores?

Answer: 30.20 to 49.80

What values would encompass 99% of the scores?

Hint: Use a Z value of 2.58 for a 99% confidence interval.

Answer: 27.10 to 52.90

With a mean of 100 and a standard deviation of 10:

What values would encompass 95% of the scores?

Answer: 80.40 to 119.60

What values would encompass 99% of the scores?

Answer: 74.20 to 125.80

The key step in each of these examples had to do with the standard deviation of your distribution of scores. In each case, you multiplied the standard deviation by a particular Z value.

Exercises like these are interesting, and they demonstrate how useful the normal curve can be, but how does all of that come into play when we're trying to construct a confidence interval? As it turns out, we'll rely on the same sort of method. We'll calculate the value we add to and subtract from our sample mean by multiplying a Z value by an expression of standard deviation units. That brings us to the question of what Z value to use.

Z Values and the Width of the Interval

To determine the right *Z* value, we first decide how wide we want our interval to be. Statisticians routinely make a choice between a 95% confidence interval and a 99% confidence interval (Pyrczak, 1995). It's possible to construct an 80% confidence interval, or a 60% confidence interval, for that matter, but statisticians typically aim for either 95% or 99%. Without worrying right now about why they do that, just focus on the fundamental difference between the two types of intervals.

In the situation we're considering—one in which (σ) is known—a 95% confidence interval is built by using a *Z* value of 1.96 in the formula. A 99% confidence interval, in turn, is built by using a *Z* value of 2.58. By now, these should be very familiar values to you. If you're unclear as to why they should be familiar, take the time to reread Chapter 4.

LEARNING CHECK

Question: What *Z* value is associated with a 95% confidence interval? What *Z* value is associated with a 99% confidence interval?

Answer: A *Z* value of 1.96 is used for a 95% confidence interval. A *Z* value of 2.58 is used for a 99% confidence interval.

Now we deal with the question of how to put the *Z* values such as 1.96 or 2.58 to use. In other words, what's the rest of the formula all about—the question mark (?) that follows the *Z* value? Just so it will be clear in your mind, here's the formula again:

$$\text{Confidence Interval, or CI} = \text{Sample Mean} \pm Z\,(\,?\,)$$

Typically Z = 1.96 or 2.58
1.96 for a 95% confidence interval
2.58 for a 99% confidence interval

Bringing in the Standard Error of the Mean

To understand what the question mark represents, take a moment or two to review what we know so far. Indulge yourself in the repetition, if necessary. The logic involved in where we've been is central to the logic of where we're going.

Returning to our example, we're attempting to estimate the mean SAT score for our total customer base of 10,000 customers, based on a sample of 225 customers. The mean math SAT score ($\bar{X}$) for the sample was 606, and we know that the SAT math section has a standard deviation (σ) of 100. It's that last bit of information (σ = 100) that allows us to approach the problem as the construction of a confidence interval with σ (the standard deviation of the population) known.

Our sample of 225 students may have produced a mean ($\bar{X}$ = 606) that equals the population mean (the mean or μ of all of our customers), but there's also a possibility it didn't. Maybe our sample mean varied just a little bit from the true population mean; maybe it varied a lot. We have no way of knowing.

The key to grasping all of this is to think back to the notion of a sampling distribution of sample means. As you know, a sampling distribution of sample means is what you would get if you took a large number of samples, calculated the mean of each sample, and plotted the means. You should also remember that most of those sample means would be located toward the center of the distribution, but some of them would be located in the outer regions—the more extreme means.

If you put our sample mean in the context of all of that, here's what you should be thinking:

> I've got a sample mean here, but I don't know where it falls in relationship to all possible sample means. A different sample could have yielded a different mean. Maybe the sample (just by chance) included mostly customers with extremely high SAT math scores, or maybe it's a sample that (just by chance) included mostly customers with extremely low scores. The probability of something like that happening is small (if a random sample was selected), but anything is possible.

In other words, there's no way to know how far the sample mean deviates from the mean (μ) of the population of 10,000 customers, if at all. In a case like that, we're left with no choice except to take into account some overall average of how far different sample means would deviate from the true population mean (μ). Of course, that's exactly what the *standard error of the mean* is—it's an overall expression of how far the various sample means deviate from the mean of the sampling distribution of sample means.

To understand this point, take some time for a dark room moment, if necessary. Just as before, imagine that you're taking an infinite number of samples, and imagine all the different means you get. Imagine a plot of all those different sample means. Most of those sample means are close to the center, but a lot of them aren't. Some deviate from the mean a little; some deviate a lot. Now begin to think about the fact that there's an overall measure of that deviation—in essence, a standard deviation for the sampling distribution. Focus on that concept—the standard deviation of a sampling distribution of sample means. Now focus on the fact that we have a special name for the standard deviation of the sampling distribution—the *standard error.* If, for some reason, that doesn't sound familiar to you, go through the dark room moment exercise again.

Assuming you're comfortable with the concept of the standard error of the mean, you can begin to think of it as analogous to what you encountered earlier in this chapter—the examples in which you were dealing with a population of scores. In those earlier situations, you multiplied the standard deviation of the distribution by 1.96 to determine the values or scores that would encompass

95% of the cases. Similarly, you multiplied the standard deviation of distribution by 2.58 if you wanted to determine the values that encompassed 99% of the cases.

In our present situation, we'll do essentially the same thing. The only change is that we'll be using the standard error instead of the standard deviation. To better understand this, take a minute or two to really focus on the illustration shown in Figure 6-1.

If you truly digested that illustration, and you realized you were looking at a sampling distribution of sample means, you noticed something very important: 95% of the *possible means* would fall between ±1.96 standard error units from the mean of the sampling distribution of sample means. By the same token, 99% of the possible means would fall between ±2.58 standard error units from the mean of the sampling distribution of sample means. None of this should surprise you. After all, the Central Limit Theorem tells us that the sampling distribution of sample means will approach the shape of a normal distribution.

Now we move toward the final stage of our solution to the problem. Remember what the task is: We want to estimate the mean math score on the SAT for our entire customer base. All we know is that the mean math SAT score for a random sample of 225 customers is 606 and that the test in question has a standard deviation (σ) of 100.

As we launch into this, let's throw in the assumption that we want to be on fairly solid ground—in other words, we want to have a substantial amount of confidence in our estimate. For this reason, we decide to construct a 99% confidence interval. For a 99% confidence interval, and taking the mean of our sample ($\overline{X}$ = 606) as our starting point, we simply add 2.58 standard error units to our sample mean and subtract 2.58 standard error units from our sample mean. That will produce the interval that we're trying to construct.

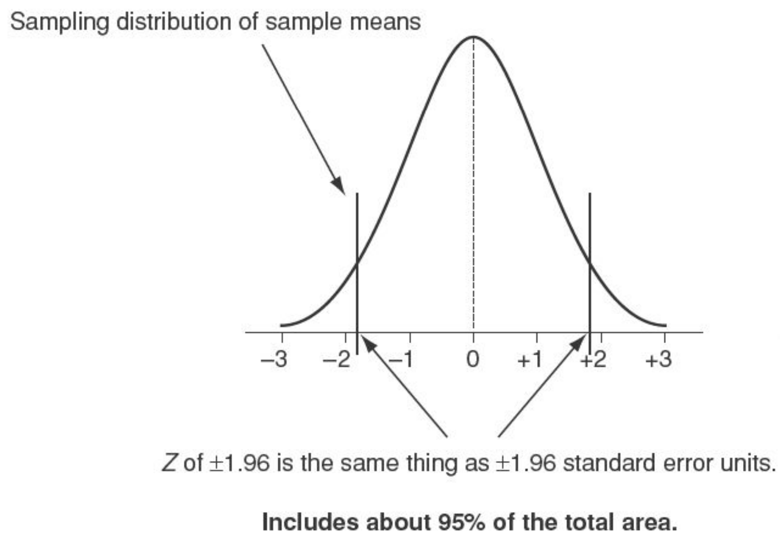

Figure 6-1 The Concept of the Standard Error of the Mean

But wait just a minute, you may be thinking. I understand that we're adding and subtracting 2.58 standard error units, but how much is a standard error unit? Indeed, that's the central question. To find the answer, all we have to do is return to the Central Limit Theorem. Think for a moment about what the Central Limit Theorem told us. Here it is once again:

> If repeated random samples of size n are taken from a population with a mean or mu (μ) and a standard deviation (σ), the sampling distribution of sample means will have a mean equal to mu (μ) and a standard error equal to $\frac{\sigma}{\sqrt{n}}$. Moreover, as n increases, the sampling distribution will approach a normal distribution.

LEARNING CHECK

Question: According to the Central Limit Theorem, what is the relationship between the standard deviation of the population (σ) and the standard error (the standard deviation of the sampling distribution of sample means)?

Answer: The standard error is equal to σ divided by the square root of the sample size.

The Relevance of the Central Limit Theorem and the Standard Error

The Central Limit Theorem tells us that the standard error of the sampling distribution (the missing value that we've been looking for) will equal the standard deviation of the population divided by the square root of our sample size. In the case we're considering here, we know that the standard deviation for the general population is 100. Thus, we divide 100 by the square root of our sample size (the square root of 225, or 15) to get the value of the standard error.

At this point, let me emphasize that what we're doing is calculating the value of the standard error. We can calculate it in a direct fashion because the Central Limit Theorem tells us how to do that. It tells us that the standard error is calculated by dividing σ by the square root of n:

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

Note that symbol for the standard error of the mean is $\sigma_{\bar{x}}$. Remember: We're working with a situation in which the standard deviation on the test (the math portion of the SAT) is 100 points. We obtain the standard error of the mean ($\sigma_{\bar{x}}$) by dividing σ (the standard deviation of the population, or 100) by the square root of our sample size (square root of 225, or 15):

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

$$\sigma_{\bar{x}} = \frac{100}{\sqrt{225}}$$

$$\sigma_{\bar{x}} = \frac{100}{15}$$

$$\sigma_{\bar{x}} = 6.67$$

In other words, the standard error of the mean ($\sigma_{\bar{x}}$) = 6.67.

Now that we have the standard error at hand, along with a grasp of the fundamental logic, we can appreciate the complete formula for the construction of a confidence interval with σ (the standard deviation of the population) known:

$$\text{CI} = \bar{X} \pm Z\,(\sigma_{\bar{x}}) \text{ where } (\sigma_{\bar{x}}) = \frac{\sigma}{\sqrt{n}}$$

LEARNING CHECK

Question: How is the standard error calculated when the standard deviation of the population (σ) is known?

Answer: The standard deviation of the population (σ) is divided by the square root of the sample size (n).

All that remains to construct a 99% confidence interval is to multiply the standard error (6.67) the appropriate or associated Z value (2.58), and wrap that product around our sample mean (add it to our mean and subtract it from our mean). As it turns out, 6.67×2.58 equals 17.21. Therefore, we add 17.21 to our sample mean ($\bar{X} = 606$) and subtract 17.21 from our sample mean to get our interval. Following through with all of that, we obtain the following:

- 606 − 17.21 = 588.79
- 606 + 17.21 = 623.21
- Therefore, our confidence interval is 588.79 to 623.21.
- We can estimate that the true mean math SAT score for our customer base is located between 588.79 and 623.21.

As a review of the entire process, here are all the calculations again, laid out from start to finish, in the context of the formula for the construction of a confidence interval for the mean (with σ known).

$$CI = \overline{X} \pm Z(\sigma_{\overline{x}})$$

$$CI = 606 \pm 2.58\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$CI = 606 \pm 2.58\left(\frac{100}{\sqrt{225}}\right)$$

$$CI = 606 \pm 2.58\left(\frac{100}{15}\right)$$

$$CI = 606 \pm 2.58(6.67)$$

$$CI = 606 \pm 17.21$$

$$CI = 588.79 \text{ to } 623.21$$

Is it possible that we missed the mark? Is it possible that the true mean math SAT score for our 10,000 customers doesn't fall between 588.79 and 623.21? You bet it's possible. Is it probable? No, it isn't very probable. The method we used will produce an interval that contains the true mean of the population 99 times out of 100 (99% of the time). Let me repeat that: The method we used will generate an interval that contains the true mean of the population 99 times out of 100. Since I repeated that, it's obviously important, so you deserve an explanation.

Think of it this way: If the previous exercise were repeated 100 times, (100 different samples of 225 students), we'd find ourselves working with many different sample means. These different sample means would result in different final answers. We would always be wrapping the same amount around our sample mean (adding the same amount of sampling error and subtracting the same amount of sampling error), but different means would result in different final answers (different intervals). In 99 of the 100 trails, our result (our confidence interval) would contain the true mean of the population.

The method would produce an interval containing the population mean 99 times out of 100 because of what lies beneath the application—random sampling, the Central Limit Theorem, and the normal curve. Statisticians have tested the method. The method works. To fully understand this idea, take a look at the illustration shown in Figure 6-2. You'll probably find it to be very helpful.

Because the central element in all of this has to do with the method we used, let me emphasize something about the way I think an interpretation of a confidence interval should be structured. Obviously, there are different ways to make a concluding statement about a confidence interval, but here is the one that I prefer (let's assume the case involves a 99% confidence interval):

> **I estimate that the true mean of the population falls somewhere between ____ and ____ (fill in the blanks with the correct values), and I have used a method that will generate a correct estimate 99 times out of 100.**

In other words, the heart of your final interpretation goes back to the method that was used. You have *confidence* in the estimate because of the *method* that was used.

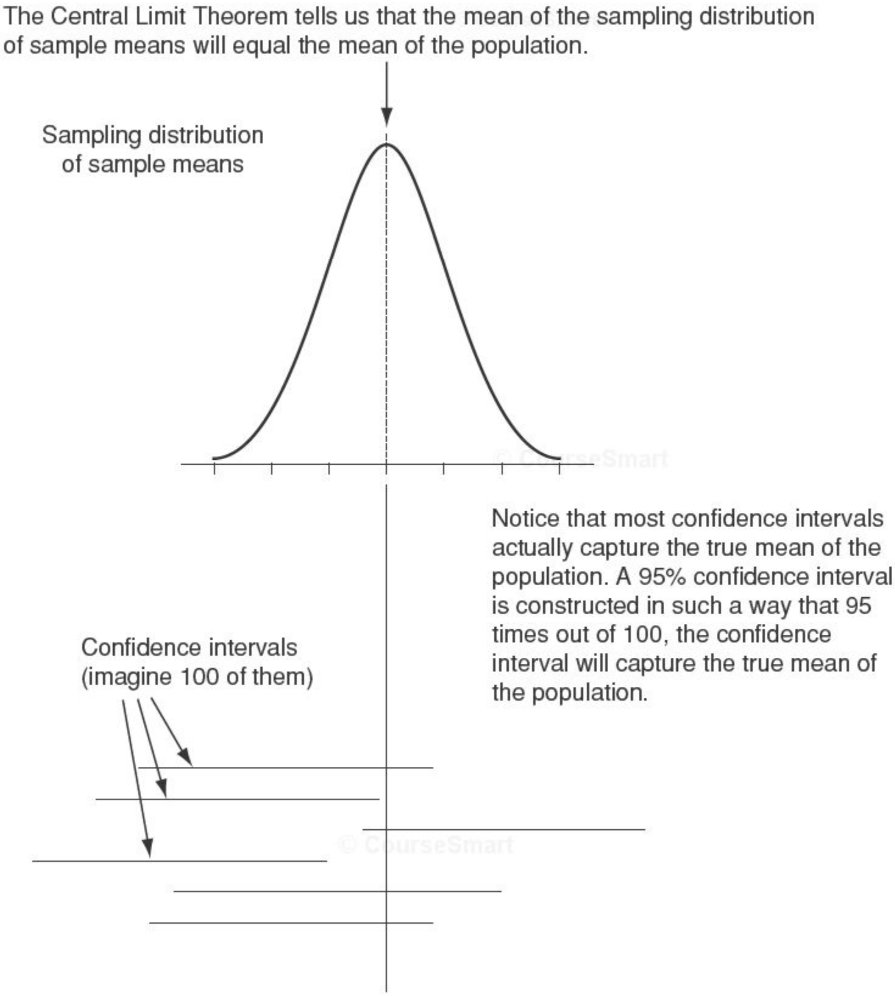

Figure 6-2 The Method Underlying the Construction of a Confidence Interval for the Mean (Why the Method Works)

Confidence and Interval Width

Now let's tackle a 95% confidence interval for the same problem. Everything will stay the same in the application, with one exception. In this instance, we'll multiply the standard error by 1.96 (instead of 2.58). Once again, here's the formula we'll be using:

$$CI = \overline{X} \pm Z(\sigma_{\overline{x}})$$

If we apply the same procedure we used before, changing only the Z value (using 1.96 instead of 2.58), we'll get an interval that is slightly smaller in width—something we would expect since we're multiplying the standard error by a slightly smaller value. Here is how the calculation would unfold:

$$CI = \overline{X} \pm Z(\sigma_{\bar{x}})$$
$$CI = 606 \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$
$$CI = 606 \pm 1.96\left(\frac{100}{\sqrt{225}}\right)$$
$$CI = 606 \pm 1.96\left(\frac{100}{15}\right)$$
$$CI = 606 \pm 1.96(6.67)$$
$$CI = 606 \pm 13.07$$
$$CI = 592.93 \text{ to } 619.07$$

Given those calculations, the appropriate conclusion or interpretation would be as follows:

> **I estimate that the true mean of the population falls between 592.93 and 619.07, and I have used a method that will produce a correct estimate 95 times out of 100.**

At this point, you should take note of the relationship between the level of confidence (95% versus 99%) and the width of the interval. The 95% confidence interval will, by definition, be narrower than the 99%. To convince yourself of this, compare our two sets of results:

For the 99% level of confidence, our interval is 588.79 to 623.21.

For the 95% level of confidence, our interval is 592.93 to 619.07.

In other words, all factors being equal, a 95% interval will produce a more precise estimate—an estimate that has a narrower range. By the same token, a 99% confidence interval will be wider than a 95% interval—it will produce a less precise estimate.

A word of clarification is probably in order at this point. To say that one estimate is more *precise* than another is to say that one estimate has a narrower range than the other. For example, an estimate that asserts that the mean of the population falls between 20 and 30 is a more precise estimate than one that asserts that the mean is somewhere between 10 and 40. It's particularly easy to get thrown off track on this topic, particularly if you're inclined to confuse *precision* with *accuracy*. Although the two terms can be used synonymously in some instances, the present context is not one of them.

If you want to understand the difference between the two (when thinking about confidence intervals), just consider the following statement: I estimate that the true mean age of the population of students falls somewhere between zero and a billion. That statement would obviously have a high degree of accuracy—it's very likely to be a correct statement and, therefore, accurate. We would also have a great deal of confidence in the estimate, just because of the width of the interval. The estimate reflected in that statement, however, is anything but precise—the range of the estimate is anything but narrow.

All of this is another way of saying that there is an inverse relationship between the level of confidence and precision. As our confidence increases, the precision of our estimate decreases. Alternatively, as our precision increases, our confidence decreases. For example, we could, if we wanted to, construct a 75% confidence interval. It would be a fairly narrow interval (at least compared to a 95% or 99% interval). It would be fairly narrow, and therefore rather precise, but we wouldn't have a lot of confidence in our estimate.

LEARNING CHECK

Question: What is the relationship between the level of confidence and the precision of an estimate when constructing a confidence interval for the mean?

Answer: Level of confidence and precision are inversely related. As one increases, the other decreases.

It's also possible to affect the precision of an estimate by changing the sample size—something that should make a certain amount of intuitive sense to you if you think about it for a minute or two. Given a constant level of confidence (let's say, a 95% level), you can increase the precision of an estimate by increasing the size of the sample. The problems presented in the next section should give you an adequate demonstration of that point.

A Brief Recap

Just to make certain that you are comfortable with all of this, let me suggest that you work through the problems that follow—typical problems that call for a 95% and a 99% confidence interval. Follow the same procedure we just used.

Assume the following: $\bar{X} = 50$ $\sigma = 8$ $n = 100$
Calculate a 95% confidence interval. Answer: 48.43 to 51.57
Calculate a 99% confidence interval. Answer: 47.94 to 52.06

Assume the following: $\bar{X} = 50$ $\sigma = 8$ $n = 400$
Calculate a 95% confidence interval. Answer: 49.22 to 50.78
Calculate a 99% confidence interval. Answer: 48.97 to 51.03

Assume the following: $\overline{X} = 85$ $\sigma = 16$ $n = 25$

Calculate a 95% confidence interval. Answer: 78.73 to 91.27

Calculate a 99% confidence interval. Answer: 76.74 to 93.26

Assume the following: $\overline{X} = 85$ $\sigma = 16$ $n = 225$

Calculate a 95% confidence interval. Answer: 82.90 to 87.10

Calculate a 99% confidence interval. Answer: 82.24 to 87.76

As before, you may want to take a moment to focus on how the width of a confidence interval varies with level of confidence and how it varies with sample size.

LEARNING CHECK

Question: What effect does increasing the size of a sample have on the width of the confidence interval and the precision of the estimate?

Answer: It decreases the width of the interval and, therefore, increases the precision of the estimate.

Confidence Interval for the Mean With σ Unknown

With the previous section as a foundation, we now take up the more typical applications of confidence interval construction—those involving an estimate of the mean of a population when the standard deviation of the population is unknown. For the most part, the logic involved is identical to what you've just encountered. There are just two hitches. I've already mentioned the first one—it has to do with the fact that you don't know the value of the population standard deviation (σ). The second hitch arises because you can't rely on the normal curve, so you can't rely on those familiar values such as 1.96 (for a 95% confidence interval) or 2.58 (for a 99% confidence interval). Rather than jumping into an application straightaway, let's take some time to really examine how the two approaches differ.

Estimating the Standard Error of the Mean

Let's start with the first hitch—the fact that you don't know the standard deviation of the population (σ). If you think back to the previous section, you were able to determine the standard error—the standard deviation of the sampling distribution of sample means—because you knew the standard deviation of the population. The Central Limit Theorem told you that all you had to do was divide the standard deviation of the population (σ) by the square root of your sample size, and the result would be standard error (the standard deviation of the sampling distribution of sample means). But now we're considering situations in which we don't know the value of the standard deviation (σ), so we can't rely on a direct calculation to get the standard error. Instead, we'll have to *estimate* it. That's the first difference in a nutshell. Remember: When you know the value of the standard deviation of the population (σ)—which you rarely do—you can make a direct calculation of the standard error of the mean. When you don't know the value of the standard deviation of the population (σ)—which is usually the case—you'll have to estimate the standard error.

LEARNING CHECK

Question: When constructing a confidence interval for the mean, how do you approach the standard error? How does the approach differ, depending on whether you know the value of the standard deviation of the population (σ)?

Answer: If σ is known, you make a direct calculation of the value of the standard error. If σ is unknown, you have to estimate the value of the standard error.

As it turns out, there's a very reliable estimate of the standard error of the mean, and it's easy to calculate. All we have to know is the standard deviation of our sample (s) and our sample size (n). Assuming we have the standard deviation of our sample at hand, we simply divide it by the square root of our sample size. We designate the **estimate of the standard error of the mean** as $s_{\overline{X}}$. The formula for the estimate is as follows:

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

For example, let's say we're interested in the average expenditure per customer in a bookstore. A sample of 100 sales receipts reveals that the mean ($\overline{X}$) expenditure is \$31.50 with a standard deviation (s) of \$4.75. To estimate the standard error of the mean, we would simply divide the standard deviation of the sample (s = \$4.75) by the square root of the sample size (n = 100).

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

$$s_{\overline{X}} = \frac{4.75}{\sqrt{100}}$$

$$s_{\bar{x}} = \frac{4.75}{10}$$

$$s_{\bar{x}} = .475$$

$$s_{\bar{x}} = .48$$

In other words, the standard error of the mean would be .475 (rounded to $.48).

Let me mention one minor point here. If the standard deviation of our sample was derived using the $n - 1$ correction factor discussed in Chapter 2, we will do just as I outlined above. We'll divide the sample standard deviation (s) by the square root of our sample size (n). If, on the other hand, the standard deviation of the sample (s) was obtained *without* using the $n - 1$ correction factor, we'll obtain the estimate of the standard error by dividing the sample standard deviation (s) by the square root of $n - 1$. This point is demonstrated in Table 6-1, which should help you understand why different texts approach the estimate of the standard error in different ways.

Since the approach taken throughout this book is to assume that the sample standard deviation was calculated using the $n - 1$ correction faction, all we

Table 6-1 Two Approaches to Estimating the Standard Error of the Mean ($s_{\bar{x}}$), and an Important Note

When the sample standard deviation (s) has been calculated using $n - 1$ in the denominator, the estimate of the standard error ($s_{\bar{x}}$) is computed as follows: $\frac{s}{\sqrt{n}}$	When the sample standard deviation (s) has been calculated using n in the denominator, the estimate of the standard error ($s_{\bar{x}}$) is computed as follows: $\frac{s}{\sqrt{n-1}}$

AN IMPORTANT NOTE: Just in Case You're a Little Bit Confused . . .

Always remember that different statisticians and different resources may approach the same topic in different fashions. The examples above provide a case in point. Some statisticians calculate the standard deviation of a sample using only n in the denominator when they simply want to know the sample standard deviation, but switch to $n - 1$ in the denominator when they're using the sample standard deviation as an estimate of the population standard deviation.

There's no reason to let all of this confuse you. Just remember that some of the fundamentals of statistical analysis aren't carved in stone, despite what you might have thought. If you encounter different symbols, notations, or approaches, don't let them throw you. A little bit of time and effort will, I suspect, unravel any mysteries.

had to do was divide 4.75 (the sample standard deviation, or s) by the square root of 100 (the sample size). The result was 4.75 divided by 10, or .48. That value of .48 becomes our estimate of the standard error—an estimate of the standard deviation of the sampling distribution of sample means. Just to make certain you're on the right track with all of this, consider the following examples:

Given		**Estimate of the standard error of the mean ($s_{\bar{x}}$)**
$s = 8$	$n = 100$	Answer: 0.80
$s = 20$	$n = 25$	Answer: 4.00
$s = 6$	$n = 36$	Answer: 1.00
$s = 50$	$n = 225$	Answer: 3.33

LEARNING CHECK

Question: How do you estimate the value of the standard error of the mean ($s_{\bar{x}}$)?

Answer: The standard error of the mean is estimated by dividing the sample standard deviation (s) by the square root of the sample size ($\sqrt{n}$).

Now we turn to the second hitch—the fact that we can't rely on the normal curve or the sampling distribution of Z, with its familiar values such as 1.96 or 2.58. The why behind this problem, which can be found in a more advanced statistical text, is something you shouldn't concern yourself with at this point. What's important is what we can use as an alternative to the normal curve distribution. Instead of relying on the normal distribution and its familiar Z values, we'll rely on what's referred to as the **family of *t* distributions.**

The Family of t Distributions

As the expression implies, the family of t distributions is made up of several distributions. Like the normal curve, each t distribution is symmetrical, and each curve has a mean of 0, located in the middle. Positive t values, or deviation units, lie to the right of 0, and negative t values lie to the left—just like Z scores on the normal curve. But there are many different t distributions, and the exact shape of each distribution is based on sample size (n). It was William Gosset, an early-day statistician and employee of the Guinness Brewery, who developed the notion of the t distribution.

Without going into the mathematics behind Gossett's contribution, it's useful to consider what it tells us—namely, that the shape of a sampling distribution depends on the number of cases in each of the samples that make up the sampling distribution. When the number of cases is small, the distribution will be relatively flat. As the number of cases in each sample increases, however, the middle portion of the curve will begin to grow. As the middle portion of the curve grows, the curve begins to take on more height.

To understand what happens with an increase in sample size, take a look at Figure 6-3. Think of each curve as a sampling distribution of sample means. Notice how the curve begins to grow in the middle as you move from a sampling distribution based on small samples to a sampling distribution based on samples with a larger number of cases. The curves presented here are exaggerated or stylized (they're not based on the construction of actual sampling distributions), but they serve to illustrate the point.

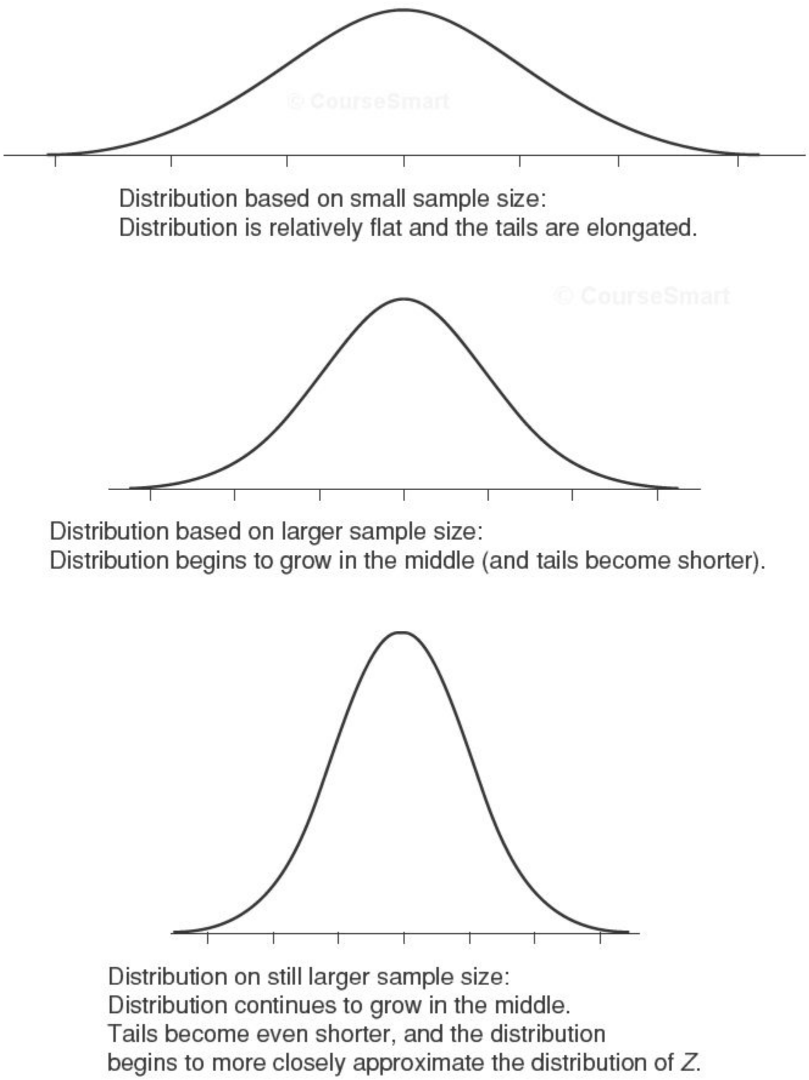

Figure 6-3 Shape of *t* Distribution in Relationship to Sample Size

Assuming you've grasped the idea that the shape of the sampling distribution is a function of the size of the samples used in constructing it, we can now move on toward a more precise understanding of the specific shapes. As a first step in that direction, let me ask you to start thinking in terms of *t* values in the same way that you've thought of *Z* values. A *t* value (like a *Z* value) is just a point along the baseline of a distribution (or, more correctly, a sampling distribution). Now think back to a couple of points I mentioned earlier.

First, there are many different sampling distributions of *t*, and each one has a slightly different shape. A *t* distribution built on the basis of small samples will be flatter than one based on underlying samples that are larger. When a distribution is flat, you'll have to go out a greater distance above and below the mean to encompass a given percentage of cases or area under the curve. To better grasp this point, consider Figure 6-4 (as before, the distributions are somewhat stylized to make the point).

Remember: We're dealing with the confidence intervals for the mean when the standard deviation of the population (σ) is unknown. Since you're not going to be able to use the normal curve and its familiar values such as 1.96 or 2.58, it's time you take a look at Gossett's family of *t* distributions.

The Table for the Family of t Distributions

You'll find the family of *t* distributions presented twice—once in Appendix B and again in Appendix C. For the application we're considering here (the construction of a confidence interval for the mean), you'll be working with Appendix B. Before you turn to Appendix B, though, let me give you an overview of what you'll encounter.

First, you'll notice a column on the far left of the table. It is labeled Degrees of Freedom (*df*). The concept of *degrees of freedom* is something that comes up throughout inferential statistics and in many different applications. The exact meaning of the concept, in a sense, varies from application to application. At this point, you'll need to know a little about degrees of freedom in the context of a mean.

Here's an easy way to think of it: Given the mean of a distribution of n scores, $n - 1$ of the scores are free to vary. Let me give you a translation of that. Assume you have a sample of five incomes ($n = 5$) and the mean income of the sample is \$26,354. In this situation, four of the incomes could be any numbers you might choose, but given a mean of \$26,354, the fifth income would then be predetermined. In other words, only four of the five cases ($n - 1$) are free to vary.

Here's another example of how and why that works out. Let's say we have a sample of seven scores on a current events test with a maximum possible score of 10, and we know that the mean score is 5. With seven cases and a mean of 5, we know that the total of all the scores must equal 35. Six of the values ($n - 1$) are free to vary. Let's just make up some values—for example, 1, 2, 3, 3, 7, and 10. The total of these six values is 26. So what must the missing score be (the one that isn't free to vary)? We already know that the sum of all the

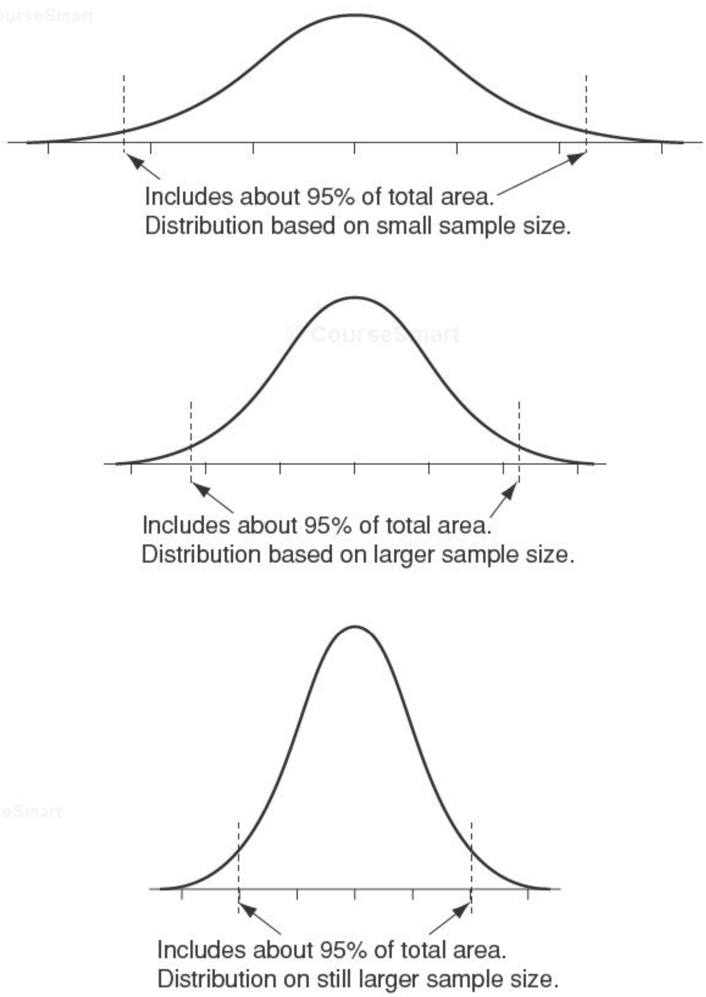

Figure 6-4 Relationship Between Area Under the Curve (t Distribution) and Sample Size

scores must equal 35 (35/7 = 5, our mean). If we have to reach a final total of 35, and these six values add up to 26, the missing score must be 35 – 26, or 9.

Here is another example to illustrate the point:

Total of five scores (n = 5), mean = 8. Degrees of freedom (n – 1) = 4.

If the mean is 8 and there are five scores, the total of all scores must be 40 (40/5 = 8).

Pick any four scores (let four vary); let's say the scores are 8, 8, 10, and 10.

The total of those four scores is 36.

The total of all the scores must equal 40; therefore, the missing score has to be 4.

8 + 8 + 10 + 10 + **4** = 40

The missing score—the one that is predetermined. Only four scores are free to vary; the fifth score is predetermined.

Just for good measure, here's yet another example:

Total of six scores ($n = 6$), mean = 8. Degrees of freedom ($n - 1$) = 5.

If the mean is 8 and there are six scores, the total of all scores must be 48 (48/6 = 8).

Pick any five scores (let five vary); let's say the scores are 6, 7, 7, 10, and 10.

The total of those five scores is 40.

The total of all the scores must equal 48; therefore, the missing score has to be 8.

6 + 7 + 7 + 10 + 10 + **8** = 48

The missing score—the one that is predetermined. Only five scores are free to vary; the sixth score is predetermined.

All of that is what lies behind the left-hand column of the table in Appendix B. If you're attempting to construct a confidence interval of the mean, and you have a sample size of 22, you'd be working at 21 degrees of freedom ($n - 1$, or 22 – 1). If you were working with a sample of 15, you'd be working with 14 degrees of freedom ($n - 1$, or 15 – 1). And so it goes. Now let's turn our attention to another part of the table.

LEARNING CHECK

Question: When using the *t* table and constructing a confidence interval for the mean (with σ unknown), how is the number of degrees of freedom computed?

Answer: The number of degrees of freedom will equal $n - 1$ (the size of the sample, minus 1).

At the top of the table, you'll see the phrase Level of Significance. Later on we'll take up the exact meaning of that phrase in greater detail. For the moment, though, I'll just ask you to make a slight mental conversion in using the table. If you want to construct a 95% confidence interval, just look at the section for the .05 (5%) level of significance. You can simply think of it this way: 1 minus the level of significance will equal the **level of confidence.** If you want to construct a 99% confidence interval, you'll go to the section for the .01 (1%) level of significance. (Remember: 1 minus the level of significance equals the level of confidence.)

> Use the .05 level of significance for a 95% confidence interval (1 – .05 = .95).
>
> Use the .01 level of significance for a 99% confidence interval (1 – .01 = .99).
>
> Use the .20 level of significance for an 80% confidence interval (1 – .20 = .80).

Before you turn to Appendix B, let me mention one last thing about how the table has been constructed and how it differs from the Table of Areas Under the Normal Curve. Recall for a moment that the Table of Areas Under the Normal Curve was one table for one curve. What you're going to see in Appendix B is really one table for many different curves. Therefore, the Table for the Family of *t* Distributions is constructed in a different fashion.

Instead of the *Z* values that you're accustomed to seeing in the Table of Areas Under the Normal Curve, you'll see *t* values. The *t* values are directly analogous to *Z* values—you can think of the *t* values as points along the baseline of the different *t* distributions. The *t* values, however, won't be listed in columns (as was the case with the *Z* values in the Table of Areas Under the Normal Curve); instead, they will appear in the body of the table. Finally, all the different proportions (or percentages of areas under the curve) that you're accustomed to seeing in the Normal Curve Table won't appear the same way in Appendix B. As noted previously, you'll only see a few of the proportions (or percentages). What's more, the percentages that you'll see appear in an indirection fashion. The percentages values are there—for example, 80%, 90%, 95%, 99%—but they're found by looking at the column headings labeled Level of Significance (.20, .10, .05, .01). Remember: 1 minus the level of significance equals the level of confidence.

You've had enough preparation to take a serious look at Appendix B. Let me urge you to approach it the way I suggest students approach any table. Instead of simply glancing at the table and saying "OK, I've looked at it," take a few moments to thoroughly digest the material. Consider the following statements and questions as you study the table. They're designed to make you more familiar with the content of the table and how it's structured. Don't worry that you're still not making a direct application of the material. Remember what the objective is: The idea is to understand how the table is structured. Just to make sure you do, take a look at the following.

> If you're going to construct a 95% confidence interval for the mean, you'll be working with values found in the .05 Level of Significance column. Remember: The confidence level is 1 minus the level of

significance. Locate the appropriate column for a 95% confidence interval.

If you want a 99% confidence interval, you'll be working with values in the .01 Level of Significance column. Locate the appropriate column for a 99% confidence interval.

What about an 80% confidence interval? What column would you focus on? (Answer: .20 Level of Significance)

If you're working with a sample of 35 cases, you'll be focusing on the row associated with 34 degrees of freedom. Remember: Degrees of freedom equals the number of cases minus 1. Locate the row for 34 degrees of freedom.

What about a sample of 30 cases? What row would you focus on? (Answer: The row associated with 29 degrees of freedom)

What about a sample of 25 cases? What row would you focus on? (Answer: The row associated with 24 degrees of freedom)

LEARNING CHECK

Question: When using the *t* table and constructing a confidence interval for the mean (with σ unknown), how do you find the level of confidence in the table? Give an example.

Answer: The level of confidence is expressed indirectly. It is equal to 1 minus the level of significance. For example, to work at the 95% level of confidence, use the column dedicated to the .05 level of significance (1 – .05 = .95).

An Application

Assuming you feel comfortable enough to move ahead, we can now tackle an application or two. Let's say that we have a random sample of 25 retirees, and we want to estimate the average number of emails retirees send out to friends or relatives each week. Let's further assume that our sample yields a mean of 12 (12 emails per week) with a standard deviation of 3 and that we've decided to construct a 95% confidence interval for our estimate of the mean. Those are the essential ingredients we need, so now the question is how to proceed.

First, we take the sample mean of 12 as a starting point. Then, we build our cushion by adding a certain amount to the mean and subtracting a certain amount from the mean. Here's the formula we'll be working with—one that's remarkably similar to the one you encountered earlier:

$$CI = \overline{X} \pm t(s_{\overline{x}})$$

Since all we have is the sample standard deviation (the population standard deviation, or σ, is unknown), we'll be working with the *t* distribution, and we'll have to estimate the standard error.

The value of *t* we'll use is found by locating the intersection of the appropriate degrees of freedom and confidence level. In this case, we have 24 degrees of freedom ($n - 1$, or $25 - 1$), so that's the row in the table that we'll focus on. We want to construct a 95% confidence interval, so we'll focus on the .05 Level of Significance column ($1 - .05 = .95$). The point in the body of the table at which the selected row and column intersect shows the appropriate *t* value of 2.064 (rounded to 2.06).

We'll have to multiply the *t* value (2.06) by our estimate of the standard error, so the next step is to calculate the estimate. We estimate the standard error ($s_{\bar{x}}$) by dividing our sample standard deviation (s) of 3 by the square root of our sample size (the square root of 25, or 5). The result (3/5, or .60) is our estimate of the standard error.

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

$$s_{\bar{x}} = \frac{3}{\sqrt{25}}$$

$$s_{\bar{x}} = \frac{3}{5}$$

$$s_{\bar{x}} = 0.60$$

We now have everything we need: our sample mean as a starting point, the appropriate *t* value, and our estimate of the standard error. When plugged into the formula (the mean, plus and minus a little bit of cushion), here's what we get:

$$\text{CI} = \overline{X} \pm t(s_{\bar{x}})$$

$$\text{CI} = 12 \pm 2.06(0.60)$$

$$\text{CI} = 12 \pm 1.24$$

$$\text{CI} = 10.76 \text{ to } 13.24$$

We can now say we estimate that the true mean of the population falls somewhere between 10.76 and 13.24 emails per week, and we have used a method that will produce a correct estimate 95 times out of 100.

Assuming all of that made sense, let's change the problem just a bit. Let's say that we're more concerned about confidence than precision, so we want to construct a 99% interval. The steps are the same, and so are all the values, except one—the appropriate *t* value. In this case, we're working with a 99% confidence interval, so our *t* value will be 2.80. As we have seen previously, our interval will now be a little wider. Our confidence will increase (from 95% to 99%), but our precision will decrease (the interval will be wider).

$$\text{CI} = \overline{X} \pm t(s_{\overline{X}})$$
$$\text{CI} = 12 \pm 2.80(0.60)$$
$$\text{CI} = 12 \pm 1.68$$
$$\text{CI} = 10.32 \text{ to } 13.68$$

Our confidence interval now ranges from 10.32 to 13.68—an interval that's slightly wider than the one we got when we constructed a 95% confidence interval.

Assuming you're getting the idea here, let's try a few more problems that should solidify your thinking. In each case, give some thought to each element that comes into play in the problem solution.

1. Given a mean of 100, a standard deviation of 12, and $n = 16$, construct a 95% confidence interval for the mean. Answer: 93.61 to 106.39
2. Given a mean of 54, a standard deviation of 15, and $n = 25$, construct a 99% confidence interval for the mean. Answer: 45.60 to 62.40
3. Given a mean of 6500, a standard deviation of 240, and $n = 16$, construct a 95% confidence interval for the mean. Answer: 6372.20 to 6627.80

Assuming you took the time to work through those problems, let me ask you to do one more thing—something similar to what you did in the last section. Pick any one of the problems you just worked, and change it by substituting a larger sample size. For example, focus on problem 2 and change the sample size from 25 to, let's say, 100. Before you even work through the reformulated problem, give some thought to what you expect will happen to the width of the interval when you construct it on the basis of $n = 100$. Consider that this would be a substantial increase in sample size. Notice what the increase in sample size does to the width of the interval (and, therefore, what it does to the precision of the estimate).

The principle involved is the same as the one you encountered earlier. Given a constant level of confidence (let's say, 95%), you can increase the precision of an estimate (or decrease the width of the interval) by increasing your sample size. To understand the logic behind this, think of the largest sample size you could possibly have. That, of course, would be the entire population. In that case, there would be no standard error, and your estimate would exactly equal the mean of the population—the narrowest interval you could possibly have!

A Final Comment About the Interpretation of a Confidence Interval for the Mean

At this point, it's probably a good idea to return the fundamental meaning of a confidence interval for the mean. Let's take the example of a sample mean of 108 and a corresponding confidence interval that ranges from 99.64 to

116.36 (Elifson, Runyon, & Haber, 1990). In interpreting those results (or any other for that matter), it is wise to remember what a confidence interval does and does not tell us.

> In establishing the interval within which we believe the population mean falls, we have *not* established any probability that our obtained mean is correct. In other words, we cannot claim that the chances are 95 in 100 (or 99 in 100) that the population mean is 108. Our statements are valid only with respect to the interval and not with respect to any particular value of the sample mean. (Elifson et al., 1990, pp. 367–368)

Translation? A confidence interval for the mean doesn't provide you with an exact estimate of the value of the population mean. Rather, it provides you with an interval—an interval of two values—that you believe contains the true mean of the population. If you were working at a 95% level of confidence, *and you went through the exercise of constructing a confidence interval 100 times,* 95 times your result would be a confidence interval that contains the true mean of the population. Do you ever know that you've produced an interval that does, in fact, contain the true mean of the population? No. On the other hand, you do know the probability that you've produced an interval containing the population mean. It's all about probability and the method—the probability that your method has generated a correct interval estimate.

A Final Comment About Z Versus t

In practice, some statisticians use the *Z* distribution (instead of *t*), even when σ is unknown, provided they are working with a large sample. Indeed, in many texts, you'll find an application based on the use of the *Z* distribution in such cases (σ unknown, but a large sample). The easiest way to understand why it's possible to use *Z* with a large sample, even if you don't know the value of σ, is to take a close look at Appendix B again and concentrate on what happens to the *t* values as the degrees of freedom increase. To fully comprehend this point, take a moment to look at Figure 6-5.

Keeping in mind that the number of degrees of freedom is an indirect statement of sample size, you'll see something rather interesting in Figure 6-5. Once you're beyond 120 degrees of freedom (see the entry for infinity, ∞), the values of *t* and *Z* are identical. For example, if you were working with a sample of 150 cases and constructing a 95% confidence interval for the mean, it really wouldn't make any difference if you relied on the value of *t* or *Z*. Both values would be 1.96. It may be a minor point, but explanations like this can go a long way when you're trying to understand why two texts or resources approach the same topic in a slightly different fashion.

Having dealt with that minor point, we can now turn our attention to a slightly different topic. Instead of dealing with means, we'll move to the topic of proportions.

Degrees of Freedom	LEVEL OF SIGNIFICANCE					
	.20	.10	.05	.02	.01	.001
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
60	1.296	1.671	2.000	2.390	2.660	3.460
80	1.292	1.664	1.990	2.374	2.639	3.416
100	1.290	1.660	1.984	2.364	2.626	3.390
120	1.289	1.658	1.980	2.358	2.617	3.373
Infinity	1.282	1.645	1.960	2.327	2.576	3.291

The value of t equals Z beyond 120 degrees of freedom. Note that t is equal to 1.96 for a 95% confidence interval (equivalent to the Z value of 1.96).

Figure 6-5 What Happens to t Beyond 120 Degrees of Freedom

Confidence Intervals for Proportions

The application we take up now may strike you as familiar, because it's the sort of thing you're apt to encounter in everyday life—something that's very common in the fields of public opinion and market research, as well as sociology and political science.

The purpose behind a **confidence interval for a proportion** parallels that of a confidence interval for the mean. We construct a confidence interval for a proportion on the basis of information about a proportion in a sample—for example, the proportion in a sample that favors capital punishment. Our ultimate purpose, however, is to estimate the proportion (in support of capital punishment) in the population.

When someone reports the results of a political poll or a survey, he/she frequently speaks in terms of proportions or percentages—for example, 57% responded this way and 43% responded that way, or 34 of the 60 respondents (an expression of a proportion) said this and 26 said that. To take another example, an opinion poll might report that 88% of voters in a community have a favorable attitude toward Councilman Brown. Maybe we're also told that the poll has a margin of error of ±3%. That simply means that somewhere between 85% and 91% of the voters hold a favorable attitude toward Brown (once again, an estimate expressed as an interval). In each instance, the purpose is to get an estimate of the relevant proportion in the population.

The question, of course, is how did the political pollster come up with that projection. I dare say that's a question that you've asked yourself at one time or another. As it turns out, the procedure is really quite simple, and it is based on the same logic that you encountered earlier in this chapter. The big difference is that in this instance the goal is to estimate a proportion by constructing a confidence interval for a proportion (as opposed to a mean).

LEARNING CHECK

Question: What is the purpose behind the construction of a confidence interval for a proportion?

Answer: A confidence interval for a proportion is constructed in an effort to estimate the proportion in a population, based upon a proportion in a sample.

An Application

Let's say that Candidate Groves is running for mayor, and he's asked us to survey a random sample of 200 likely voters. He wants us to find out what proportion of the vote he can expect to receive. Let's say that our survey results indicate that 55% of the likely voters intend to vote for Groves for mayor. Given what we know about sampling error, we know that we have to take into account the fact that we're working with only one sample of 200 voters. A different sample of 200 voters would likely yield slightly different results (it's just a matter of sampling error). Given that situation, it's clear that we'll have to come up with some measure of standard error. As before, we'll eventually use that measure, along with a Z value, to develop our interval or our projection of the eventual vote. If there's a hitch in all this, it has to do with how we estimate the standard error of the proportion. We'll eventually get to all of that, but for the moment, let's review the problem under consideration, in light of our now familiar logic.

The fundamental logic in this problem will be the same as before. We've determined that 55% of the respondents said that they plan to vote for Groves, so we use that as our starting point. We'll place our observed sample proportion (or percentage) in the middle of a sampling distribution of sample proportions (not sample means, but sample proportions). Using our observed proportion as a starting point, we'll then build in a cushion, just as we did before. To build the cushion, we'll add some standard error to our observed proportion, and we'll subtract some standard error from our observed proportion. The result will be a confidence interval—just as we had in the cases involving estimates of the population mean.

For the sake of this example, let's assume that we want to construct a 95% confidence interval for the proportion. Given our sample size (n = 200), we can use the Z score associated with 95% of the area under the normal curve as one of the elements in our computation. Note how remarkably similar the formula (stated in somewhat nonmathematical terms) is to what we encountered earlier:

Confidence Interval (CI) for a proportion = observed proportion ± Z (?)

We already know that our observed proportion is .55 (the proportion intending to vote for Groves), and we know that the Z value will be 1.96 (since we're constructing a 95% confidence interval). All that remains is to determine where we get the value to substitute for the question mark. As it turns out, the value that we're looking for is the estimate of the standard error of the proportion. I should tell you in advance that the formula for the estimate of the standard error of the proportion (s_p) is a little ominous at first glance, but it's also quite straightforward if you take the time to examine it. Here it is:

$$s_p = \sqrt{\frac{P(1 - P)}{n}}$$

As complex as this formula may appear, let me assure you that it is easy to understand if you take it apart, element by element. First, the P in the formula represents the value of the observed proportion (.55, or 55% if expressed as a percentage). The value of $1 - P$, therefore, represents the remaining proportion (.45, or 45% if expressed as a percentage). In other words, $P + (1 - P) =$ 100%. As before, we'll have to consider our sample size along the way. Substituting the appropriate values for the elements in the formula, we obtain the standard error of the proportion as follows:

$$s_p = \sqrt{\frac{0.55(1 - 0.55)}{200}}$$

$$s_p = \sqrt{\frac{0.55(0.45)}{200}}$$

$$s_p = \sqrt{\frac{.2475}{200}}$$

$$s_p = \sqrt{0.0012375}$$

$$s_p = 0.035$$

Armed with the value of the estimate of standard error of the proportion (0.035), and assuming we want to construct a 95% confidence interval for the proportion, we can now complete the problem as follows:

$$CI = P \pm Z\,(s_p)$$

$$CI = 0.55 \pm 1.96\ (0.035)$$
$$CI = 0.55 \pm 0.0686$$
$$CI = 0.4814 \text{ to } 0.6186$$
$$CI = 48.14\% \text{ to } 61.86\%$$

Thus, we're in a position to estimate that between 48.14% and 61.86% of the voters are likely to vote for Groves. As before, we could include a statement that we've used a method that generates a correct estimate 95 times out of 100.

Margin of Error

Public opinion poll results are rarely expressed in the form of an interval. Rather, the results are typically given with some reference to a **margin of error.** For example, a pollster may report that 34% approve of Proposition X, with a margin of error of ±4%. By now you should understand that the margin of error is, in effect, simply a statement of the interval width. Thinking back to the poll for Candidate Groves, we can say that the margin of error was 6.86%. After all, that was the amount that we were adding and subtracting to develop the confidence interval.

LEARNING CHECK

Question: In a confidence interval for a proportion, what is the margin of error? Give an example.

Answer: The margin of error is an indirect statement of the width of the interval. For example, the statement that the proportion in a population is estimated at 45%, with a margin of error of ±3%, is actually a statement that the interval of the estimate ranges from 42% to 48%.

For Candidate Groves' purposes, the margin of error (55%, plus or minus 6.86%) is so large that he can't take much comfort in the poll. He might capture as much as 61.86% of the vote, but he might receive only 48.14%. For a more precise estimate (at the same level of confidence), Groves would have to request a larger sample size. For example, we could follow through the same calculations again, but with the assumption that we're working with a sample of 750 likely voters. As you'll soon discover, the width of our confidence interval (and, therefore, the margin of error) would decrease quite a bit.

First, we'll recalculate the estimate of the standard error of the proportion with our new sample size:

$$s_p = \sqrt{\frac{0.55(1 - 0.55)}{750}}$$
$$s_p = \sqrt{\frac{0.55(0.45)}{750}}$$
$$s_p = \sqrt{\frac{.2475}{750}}$$
$$s_p = \sqrt{0.00033}$$
$$s_p = 0.018$$

Then, we'll use the new estimate of the standard error of the proportion to calculate our confidence interval:

$$\text{CI} = P \pm Z\,(s_p)$$
$$\text{CI} = 0.55 \pm 1.96\,(0.018)$$
$$\text{CI} = 0.55 \pm 0.0353$$
$$\text{CI} = 0.5147 \text{ to } 0.5853$$
$$\text{CI} = 51.47\% \text{ to } 58.53\%$$

Based on a sample of 750, then, our estimate would result in a projected vote between 51.47% to 58.53%. By the same token, we could legitimately report our results as a projected vote of 55% with a margin of error of 3.53%.

LEARNING CHECK

Question: Given a constant level of confidence, what is the effect on the margin of error of increasing the sample size when developing a confidence interval for a proportion?

Answer: Given a constant level of confidence, an increase in the size of a sample will decrease the margin of error.

As you're probably aware, pollsters commonly refer to a margin of error, but they rarely refer to the level of confidence that underlies their estimate. As a student of statistics, however, you're now aware that the two concepts are different. The two concepts are related, to be sure, but they are different in important ways. The margin of error is an indirect measure of the width of the interval, but the level of confidence actually goes to the method used in calculating the interval.

You'll find more examples of confidence intervals involving proportions at the conclusion of this chapter. They're presented in such a way that you'll be able to work through them in fairly quick fashion.

Chapter Summary

As we conclude this chapter, let's consider what you've covered. You've encountered a mountain of material. In the simplest of terms, you've entered the world of inferential statistics. You've learned how to construct confidence intervals. You've learned how to use sample characteristics (statistics) to make inferences about population characteristics (parameters).

You've learned, for example, about two basic approaches to constructing a confidence interval for the mean. You use one approach when you know the standard deviation of the population (σ) and a slightly different procedure when you don't know the standard deviation of the population (σ). You've also learned how to make a direct calculation of the standard error (when you know the value of σ) and how to estimate the standard error (when you don't know the value of σ).

Beyond all of that, you've learned how the survey results that you read or hear reported in the media are often derived—how a confidence interval for a proportion is constructed. You've also learned about margins of error and levels of confidence—how they're related, but how they are different.

Finally, and maybe most important, you've learned something about the world of inferential statistics in general. You've learned that there is no such thing as a direct leap from a sample to a population. You can't simply look at a sample mean (or a proportion, for that matter) and assume that it is equal to the true population parameter. You can use your sample value as a starting point, but you invariably have to ask yourself a central question in one form or another:

- Where did the sample value come from?
- Where did the sample value fall in relationship to all other values that might be possible?
- Where did the sample value fall along a sampling distribution of all possible values?
- What do I know about the sampling distribution, and how can I use that information to determine a reasonable estimate of the true population parameter?

Let me suggest that you take time out now for a dark room moment—one that might help you put a lot of this material into perspective. In this instance, I'm asking you to think about the construction of a confidence interval for the mean, but the same mental steps would be involved if you were constructing a confidence interval for a proportion.

Imagine that you've just surveyed a random sample of students, and you've calculated a mean age for your sample. This time I'm going to ask you to conjure up a mental image of a circle with some value in it—let's say 22.3. Treat that value as the mean age of your sample, and mentally focus on that circle with the value of 22.3 in the middle of it.

Now imagine a sampling distribution of sample means above the circle. Imagine that it looks something like a normal distribution sitting inside a cloud (since it's such a theoretical concept, the cloud image is probably appropriate). Think about what that sampling distribution represents—a distribution of all possible means, given repeated random sampling from a population. Now imagine that you're asking yourself a simple question. Where would my sample mean fall along the sampling distribution? Would it be at the upper end? Would it be at the lower end? Would it be somewhere in the middle?

Now imagine a rectangle above the sampling distribution. Imagine that the rectangle represents the population. Imagine a question mark in the middle of the rectangle—a question mark to convey the notion that you don't know what the value of the population mean is. That's as far as you have to go in this little mental exercise. Don't clutter your mind with the specifics of how you get from the circle on the bottom to the rectangle on the top. Simply take a mental step backward, and take in the entire view—circle, sampling distribution, and rectangle. Your image should look something like the one in Figure 6-6.

Imagine that you're first looking at the circle, then looking at the sampling distribution (moving through it, so to speak), and then moving to the rectangle. That's the essence of inferential statistics—from a sample, through a sampling distribution, and on to the population for a final answer or interpretation. As before, let me urge you to take the time to experience that dark room moment.

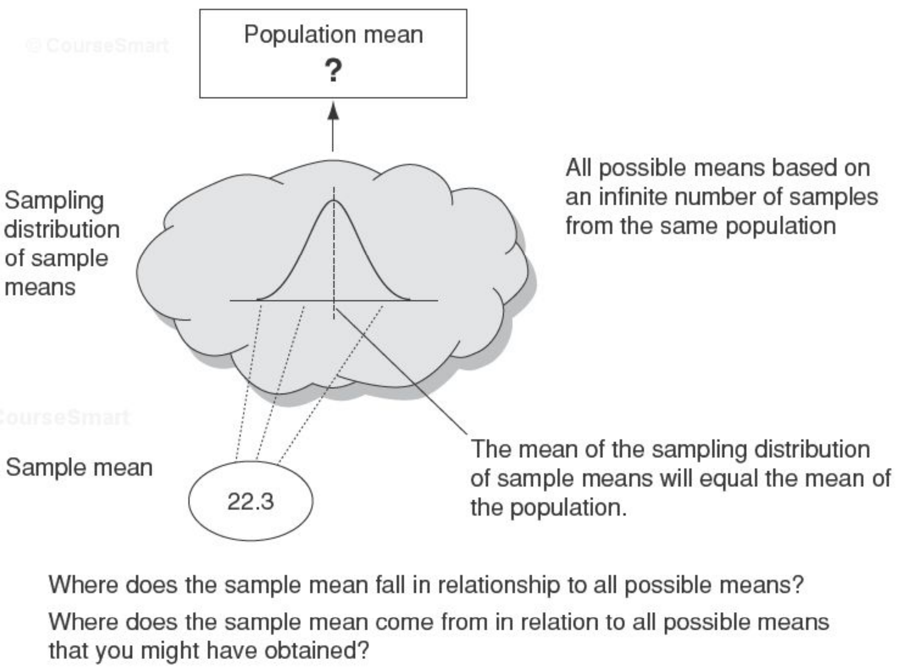

Figure 6-6 An Image of Inferential Statistics

The mental image should serve you well in the long run. What's more, it will help you prepare for our next topic—an introduction to hypothesis testing.

Some Other Things You Should Know

If there's one topic that demonstrates the matter of choice and personal preference when it comes to statistical applications, it's the topic of confidence interval construction. As I mentioned previously, some statisticians use the *Z* distribution (instead of *t*), even when σ is unknown, provided they're working with a large sample. For one statistician, though, "large" may be 60 cases; for another, it may be 100. You should always keep that in mind, particularly when you consult other resources. What may strike you as total confusion may be nothing more or less than personal preference on the part of the author.

While we're on the topic of personal preference and variation from resource to resource, you should be aware that the symbolic notation used in the field of statistical analysis is not carved in stone. For example, the notation for the estimate of the standard error used here ($s_{\bar{x}}$) is just one approach. Another text or resource may rely on a different notation (such as s_M).

Finally, you should be aware of a fundamental assumption that's involved when constructing a confidence interval for the mean with σ known. In truth, you have to make an assumption that your sample comes from a population that is equivalent to the population for which you have a known σ. Let me explain.

In the example we used at the beginning of this chapter, the assumption was made that the population of customers who had taken the SAT prep course was equivalent to the population of all students taking the SAT. We implicitly made that assumption when we took the approach that we knew the standard deviation of the population. In short, we made the assumption that our population of customers—would-be college students who enrolled in a SAT prep course—was equivalent to a population of *all* would-be college students who take the SAT. In truth, though, a population of those who enroll in a prep course may differ from the population at large in some important way (for example, maybe they are more motivated to do well, so they enroll in a prep course). For this reason, a researcher may prefer to frame the research question as though the population standard deviation (σ) were unknown, relying on a standard deviation to estimate the standard error. Once again, we're back to the matter of personal preferences.

At this point, let me encourage you to spend some time with additional resources. For example, you may want to take a look at other texts or tour the Cengage Web site www.cengage.com/psychology/caldwell. Learning to navigate your way through various approaches to the same type of question, different systems of symbolic notation, or encounters with personal preferences can provide an added boost to your overall level of statistical understanding.

Key Terms

confidence interval for the mean
confidence interval for a proportion
estimate of the standard error of the mean
family of t distributions
level of confidence
margin of error

Chapter Problems

Fill in the blanks, calculate the requested values, or otherwise supply the correct answer.

General Thought Questions

1. A confidence interval for the mean is calculated by adding and subtracting a value to and from the sample __________.
2. The purpose of constructing a confidence interval for the mean is to __________ the true value of the population mean, based upon the mean of a __________.
3. A confidence interval for the mean is an interval within which you believe the __________ of the population is located.
4. As the level of confidence increases, the precision of your estimate __________.
5. There is a(n) __________ relationship between level of confidence and precision of the estimate.
6. When constructing a confidence interval for a proportion, the margin of error is actually a reflection or statement of the __________ of the interval.
7. Whether constructing a confidence interval for a proportion or a mean, there are two ways to increase the precision of the estimate. You can __________ sample size, or you can __________ the level of confidence.
8. When constructing a confidence interval for the mean with σ known, how is the standard error of the mean calculated?
9. When constructing a confidence interval for the mean with σ unknown, how is the standard error of the mean estimated?

Application Questions/Problems: Confidence Interval for the Mean With σ Known

1. Compute the standard error of the mean, given the following values for σ (population standard deviation) and n (size of sample).
   a. $\sigma = 25$ $n = 4$
   b. $\sigma = 99$ $n = 49$
   c. $\sigma = 62$ $n = 50$
   d. $\sigma = 75$ $n = 25$

2. Given the following:
   $\overline{X} = 150$ $\sigma = 12$ $n = 25$
   a. Estimate the mean of the population by constructing a 95% confidence interval.
   b. Estimate the mean of the population by constructing a 99% confidence interval.
3. Given the following:
   $\overline{X} = 54$ $\sigma = 9$ $n = 60$
   a. Estimate the mean of the population by constructing a 95% confidence interval.
   b. Estimate the mean of the population by constructing a 99% confidence interval.
4. Given the following:
   $\overline{X} = 75$ $\sigma = 5$ $n = 100$
   a. Estimate the mean of the population by constructing a 95% confidence interval.
   b. Estimate the mean of the population by constructing a 99% confidence interval.
5. Assume you've administered a worker satisfaction test to a random sample of 25 workers at your company. The test is purported to have a population standard deviation or σ of 4.50. The test results reveal a sample mean ($\overline{X}$) of 78. Based on that information, develop an estimate of the mean score for the entire population of workers, using a 95% confidence interval.
6. The mean for the verbal component of the SAT is reported as 500, with a standard deviation (σ) of 100. A sample of 400 students throughout a particular school district reveals a mean score of 498. Estimate the mean score for all the students in the district, using a 95% confidence interval?
7. The mean for the verbal component of the SAT is reported as 500, with a standard deviation (σ) of 100. A sample of 900 students throughout a particular school district reveals a mean ($\overline{X}$) score of 522. Estimate the mean score for all the students in the district, using a 95% confidence interval.
8. Repeat Problem 7 using a 99% confidence interval.
9. The mean for the math component of the New Century Achievement Test is reported as 100, with a standard deviation (σ) of 15. A sample of 400 students throughout a particular school district reveals a mean ($\overline{X}$) score of 110. Estimate the mean score for all the students in the district, using a 99% confidence interval.

Application Questions/Problems: Confidence Interval for the Mean With σ Unknown

1. Estimate the standard error of the mean, given the following values for *s* (sample standard deviation) and *n* (sample size).
   a. $s = 5$ $n = 16$
   b. $s = 12.50$ $n = 25$

c. $s = 18.25 \quad n = 50$
d. $s = 35.50 \quad n = 30$

2. Given the following:

$\overline{X} = 26 \quad s = 5 \quad n = 30$

a. Estimate the mean of the population by constructing a 95% confidence interval.

b. Estimate the mean of the population by constructing a 99% confidence interval.

3. Given the following:

$\overline{X} = 402 \quad s = 110 \quad n = 30$

a. Estimate the mean of the population by constructing a 95% confidence interval.

b. Estimate the mean of the population by constructing a 99% confidence interval.

4. Given the following:

$\overline{X} = 80 \quad s = 15 \quad n = 25$

a. Estimate the mean of the population by constructing a 95% confidence interval.

b. Estimate the mean of the population by constructing a 99% confidence interval.

5. A sample of 25 program participants in an alcohol rehabilitation program are administered a test to measure their self-reported levels of alcohol intake prior to entering the program. Results indicate an average ($\overline{X}$) of 4.4 drinks per day for the sample of 25, with a sample standard deviation (s) of 1.75 drinks. Based on that information, develop a 95% confidence interval to provide an estimate of the mean intake level for the entire population of program participants (μ).

6. Information collected from a random sample of 29 visitors to a civic art fair indicates an average amount of money spent per person ($\overline{X}$) of \$38.75, with a sample standard deviation (s) of \$6.33. Based on that information, develop a 99% confidence interval to provide an estimate of the mean expenditure per person for the entire population of visitors.

7. A sample of 25 participants in a parenting skills class are administered a test to measure their skill levels on a 200 point skills test before entering the class. Results indicate that the mean ($\overline{X}$) skill level for the sample is 86, with a standard deviation (s) of 12. Based on that information, develop a 95% confidence interval to provide an estimate of the mean skill level for the entire population of program participants.

8. A sample of 25 participants in a parenting skills class are administered a test to measure their skill levels on a 200 point skills test before entering the class. Results indicate that the mean ($\overline{X}$) skill level for the sample is 101, with a standard deviation (s) of 16. Based on that information, develop a 95% confidence interval to provide an estimate of the mean skill level for the entire population of program participants.

9. Data are collected concerning the birth weights for a nation-wide sample of 30 Wimberley Terriers. Results indicate that the mean ($\overline{X}$) birth weight for the sample of pups equals 6.36 ounces, with a standard deviation (s) of 1.45 ounces. Based on that information, develop a 95% confidence interval to provide an estimate of the mean birth weight for the national population of Wimberley Terriers.

Confidence Interval Problems for a Proportion

1. In a sample of 200 freshmen at a state university, 40% report that they work at least 20 hours a week while in school. Estimate the proportion of all freshmen at the university working at least 20 hours per week. Develop your estimate on the basis of a 95% confidence interval.
2. From sample of 100 patients in a statewide drug rehabilitation program, you've determined that 20% of the patients were able to find employment within three months of entering the program. Estimate the percentage of patients throughout the program who were able to find employment within three months. Develop your estimate on the basis of a 99% confidence interval.
3. Of a sample of 200 registered voters, 32% report that they intend to vote in a school board election. Using a 95% confidence interval, estimate the percentage of all registered voters planning to vote.
4. Of a sample of 150 customers at a local bank, 15% report that they are likely to request a bank loan within the next year. Using a 99% confidence interval, estimate the percentage likely to request a loan within the population of all customers.
5. Results from a sample of 400 high school dropouts throughout the state reflect that 13% of the dropouts plan to return to school next year. Using a 99% confidence interval, estimate the percentage throughout the state planning to return to school next year.
6. An opinion poll based on a sample of 750 community residents indicates that 61% are in favor of a local civic redevelopment project. Estimate the level of support throughout the community, based on a 95% confidence interval.
7. An opinion poll based on responses from a sample of 250 community residents indicates that 61% are in favor of a local civic redevelopment project. Estimate the level of support throughout the community, based on a 95% confidence interval.
8. A poll, based upon a national sample of 1200 potential voters and focused on attitudes toward Social Security reform, indicates that 73.55% of the respondents oppose a proposal that would extend the minimum retirement age. Using a 95% confidence interval, estimate the proportion of opposition throughout the population of potential voters.
9. Repeat Problem 8 using a sample size of 750.