Chapter 9

The Confidence Game: Estimation

In This Chapter

- Introducing sampling distributions
- Understanding standard error
- Simulating the sampling distribution of the mean
- Attaching confidence limits to estimates

Populations and samples are pretty straightforward ideas. A population is a huge collection of individuals, from which you draw a sample. Assess the members of the sample on some trait or attribute, calculate statistics that summarize that sample, and you're in business.

In addition to summarizing the scores in the sample, you can use the statistics to create estimates of the population parameters. This is no small accomplishment. On the basis of a small percentage of individuals from the population, you can draw a picture of the population.

A question emerges, however: How much confidence can you have in the estimates you create? In order to answer this, you have to have a context in which to place your estimates. How probable are they? How likely is the true value of a parameter to be within a particular lower bound and upper bound?

In this chapter, I introduce the context for estimates, show how that plays into confidence in those estimates, and describe an Excel function that enables you to calculate your confidence level.

What is a Sampling Distribution?

Imagine that you have a population, and you draw a sample from this population. You measure the individuals of the sample on a particular attribute and calculate the sample mean. Return the sample members to the population. Draw another sample, assess the new sample's members, and then calculate *their* mean. Repeat this process again and again, always using the same number of individuals as you had in the original sample. If you could do this an infinite amount of times (with the same-size sample each time), you'd have

an infinite amount of sample means. Those sample means form a distribution of their own. This distribution is called *the sampling distribution of the mean.*

For a sample mean, this is the context I mention at the beginning of this chapter. Like any other number, a statistic makes no sense by itself. You have to know where it comes from in order to understand it. Of course, a statistic *comes from* a calculation performed on sample data. In another sense, a statistic is part of a sampling distribution.

In general, *a sampling distribution is the distribution of all possible values of a statistic for a given sample size*.

I italicize that definition for a reason: It's extremely important. After many years of teaching statistics, I can tell you that this concept usually sets the boundary line between people who understand statistics and people who don't.

So . . . if you understand what a sampling distribution is, you'll understand what the field of statistics is all about. If you don't, you won't. It's almost that simple.

If you don't know what a sampling distribution is, statistics will be a cookbook type of subject for you: Whenever you have to apply statistics, you'll plug numbers into formulas and hope for the best. On the other hand, if you're comfortable with the idea of a sampling distribution, you'll grasp the big picture of inferential statistics.

To help clarify the idea of a sampling distribution, take a look at Figure 9-1. It summarizes the steps in creating a sampling distribution of the mean.

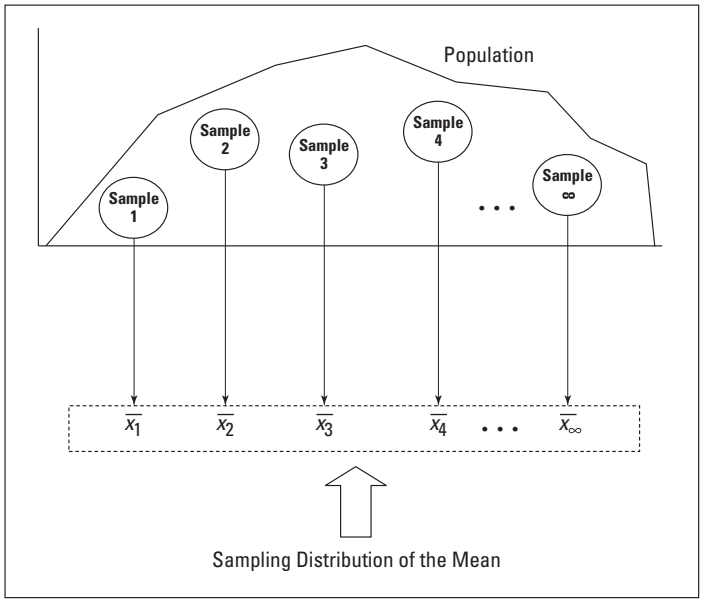

Figure 9-1: The sampling distribution of the mean.

A sampling distribution — like any other group of scores — has a mean and a standard deviation. The symbol for the mean of the sampling distribution of the mean (yes, I know that's a mouthful) is $\mu_{\bar{x}}$.

The standard deviation of a sampling distribution is a pretty hot item. It has a special name — *standard error*. For the sampling distribution of the mean, the standard deviation is called *the standard error of the mean*. Its symbol is $\sigma_{\bar{x}}$.

An EXTREMELY Important Idea: The Central Limit Theorem

The situation I ask you to imagine is one that never happens in the real world. You never take an infinite amount of samples and calculate their means, and you never create a sampling distribution of the mean. Typically, you draw one sample and calculate its statistics.

So if you have only one sample how can you ever know anything about a sampling distribution — a theoretical distribution that encompasses an infinite number of samples? Is this all just a wild-goose chase?

No, it's not. You can figure out a lot about a sampling distribution because of a great gift from mathematicians to the field of statistics. This gift is called *the Central Limit Theorem*.

According to the Central Limit Theorem

- The sampling distribution of the mean is approximately a normal distribution if the sample size is large enough.

  Large enough means about 30 or more.

- The mean of the sampling distribution of the mean is the same as the population mean.

  In equation form that's

  $$\mu_{\bar{x}} = \mu$$

- The standard deviation of the sampling distribution of the mean (also known as the standard error of the mean) is equal to the population standard deviation divided by the square root of the sample size.

  The equation here is

  $$\sigma_{\bar{x}} = \sigma \Big/ \sqrt{N}$$

Notice that the Central Limit theorem says nothing about the population. All it says is that if the sample size is large enough, the sampling distribution of

the mean is a normal distribution, with the indicated parameters. The population that supplies the samples doesn't have to be a normal distribution for the Central Limit Theorem to hold.

What if the population is a normal distribution? In that case, the sampling distribution of the mean is a normal distribution regardless of the sample size.

Figure 9-2 shows a general picture of the sampling distribution of the mean, partitioned into standard error units.

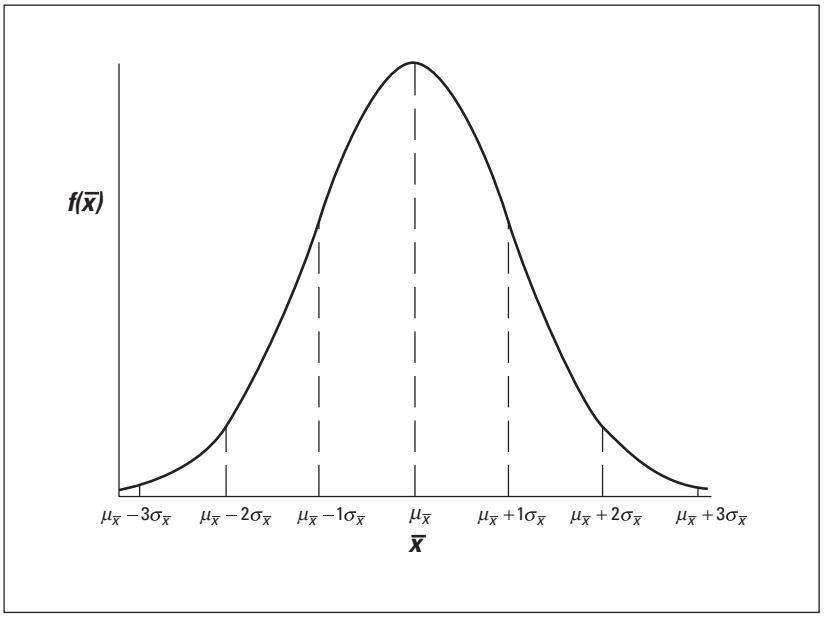

Figure 9-2: The sampling distribution of the mean.

Simulating the Central Limit Theorem

It almost doesn't sound right. How can a population that's not normally distributed result in a normally distributed sampling distribution?

To give you an idea of how the Central Limit Theorem works, I created a simulation. This simulation creates something like a sampling distribution of the mean for a very small sample, based on a population that's not normally distributed. As you'll see, even though the population is not a normal distribution, and even though the sample is small, the sampling distribution of the mean looks quite a bit like a normal distribution.

Imagine a huge population that consists of just three scores — 1, 2, and 3 and each one is equally likely to appear in a sample. (That kind of population is definitely *not* a normal distribution.) Imagine also that you can randomly select a sample of three scores from this population. Table 1 shows all the possible samples and their means.

Table 9-1 All Possible Samples of Three Scores (And Their Means) From a Population Consisting of the Scores 1, 2, and 3

Sample	*Mean*	*Sample*	*Mean*	*Sample*	*Mean*
1,1,1	1.00	2,1,1	1.33	3,1,1	1.67
1,1,2	1.33	2,1,2	1.67	3,1,2	2.00
1,1,3	1.67	2,1,3	2.00	3,1,3	2.33
1,2,1	1.33	2,2,1	1.67	3,2,1	2.00
1,2,2	1.67	2,2,2	2.00	3,2,2	2.33
1,2,3	2.00	2,2,3	2.33	3,2,3	2.67
1,3,1	1.67	2,3,1	2.00	3,3,1	2.33
1,3,2	2.00	2,3,2	2.33	3,3,2	2.67
1,3,3	2.33	2,3,3	2.67	3,3,3	3.00

If you look closely at the table, you can almost see what's about to happen in the simulation. The sample mean that appears most frequently is 2.00. The sample means that appear least frequently are 1.00 and 3.00. Hmmm . . .

In the simulation, I randomly select a score from the population, and then randomly select two more. That group of three scores is a sample. Then I calculate the mean of that sample. I repeat this process for a total of 60 samples, resulting in 60 sample means. Finally, I graph the distribution of the sample means.

What does the simulated sampling distribution of the mean look like? Figure 9-3 shows a worksheet that answers that question.

In the worksheet, each row is a sample. The columns labeled x1, x2, and x3 show the three scores for each sample. Column G shows the average for the sample in each row. Column I shows all the possible values for the sample mean, and column J shows how often each mean appears in the 60 samples. Columns I and J, and the graph, show that the distribution has its maximum frequency when the sample mean is 2.00. The frequencies tail off as the sample means get farther and farther away from 2.00.

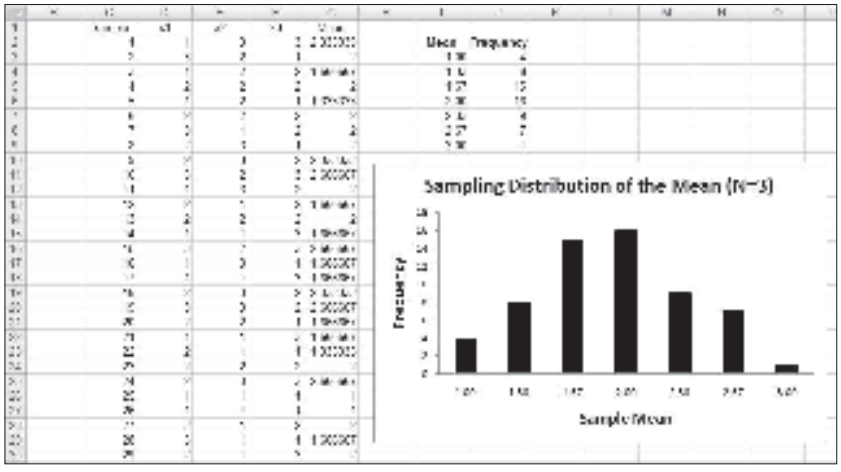

Figure 9-3: Simulating the sampling distribution of the mean (N=3) from a population consisting of the scores 1, 2, and 3. The simulation consists of 60 samples.

The point of all this is that the population looks nothing like a normal distribution and the sample size is very small. Even under those constraints, the sampling distribution of the mean based on 60 samples begins to look very much like a normal distribution.

What about the parameters the Central Limit Theorem predicts for the sampling distribution? Start with the population. The population mean is 2.00, the population variance is .67, and the population standard deviation is .82. (This kind of population requires some slightly fancy mathematics for figuring out the parameters. The math is a little beyond where we are, so I'll leave it at that.)

On to the sampling distribution. The mean of the 60 means is 1.91, and their standard deviation (an estimate of the standard error of the mean) is .48. Those numbers closely approximate the Central Limit Theorem–predicted parameters for the sampling distribution of the mean, 2.00 (equal to the population mean) and .47 (the population standard deviation, .82, divided by the square root of 3, the sample size).

In case you're interested in doing this simulation, here are the steps:

1. **Select a cell for your first randomly selected number.**

   I selected cell D2.

2. **Use the worksheet function RANDBETWEEN to select 1, 2, or 3.**

   This simulates drawing a number from a population consisting of the numbers 1, 2, and 3 where you have an equal chance of selecting each number. You can either select Formulas | Math & Trig | RANDBETWEEN and use the Function Arguments dialog box, or just type

   ```
   =RANDBETWEEN(1,3)
   ```

in D2 and press Enter. The first argument is the smallest number RANDBETWEEN returns, and the second argument is the largest number.

3. **Select the cell to the right of the original cell and pick another random number between one and three. Do this again for a third random number in the cell to the right of the second one**.

   The easiest way to do this is to autofill the two cells to the right of the original cell. In my worksheet those two cells are E2 and F2.

4. **Consider these three cells to be a sample and calculate their mean in the cell to the right of the third cell.**

   The easiest way to do this is just type

   ```
   =AVERAGE(D2:F2)
   ```

   in cell G2 and press Enter.

5. **Repeat this process for as many samples as you want to include in the simulation. Have each row correspond to a sample.**

   I used 60 samples. The quick and easy way to get this done is to select the first row of three randomly selected numbers and their mean, and then autofill the remaining rows. The set of sample means in column G is the simulated sampling distribution of the mean. Use AVERAGE and STDEVP to find its mean and standard deviation.

To see what this simulated sampling distribution looks like, use the array function FREQUENCY on the sample means in column G. Follow these steps:

1. **Enter the possible values of the sample mean into an array.**

   I used column I for this. I expressed the possible values of the sample mean in fraction form (3/3, 4/3, 5/3, 6/3, 7/3, 8/3, and 9/3) as I entered them into the cells I3 through I9. Excel converts them to decimal form.

2. **Select an array for the frequencies of the possible values of the sample mean.**

   I used column J to hold the frequencies, selecting cells J3 through J9.

3. **From the Statistical Functions menu, select FREQUENCY to open the Function Arguments dialog box for FREQUENCY.**

4. **In the Function Arguments dialog box, enter the appropriate values for the arguments.**

   In the Data_array box, I entered the cells that hold the sample means. In this example, that's G2:G61.

5. **Identify the array that holds the possible values of the sample mean.**

   FREQUENCY holds this array in the Bins_array box. For my worksheet, I3:I9 goes into the Bins_array box. After you identify both arrays, the Function Arguments dialog box shows the frequencies inside a pair of curly brackets. (See Figure 9-4.)

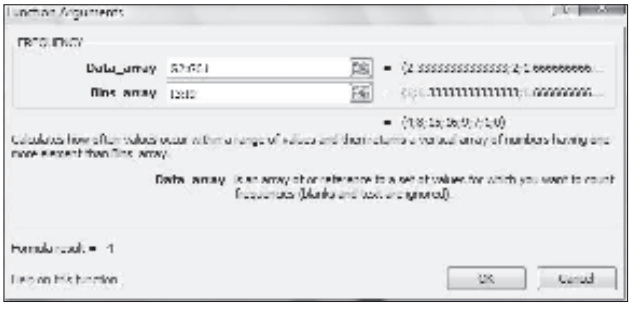

Figure 9-4: The Function Arguments dialog box for FREQUENCY in the Simulated Sampling Distribution Worksheet.

6. **Press Ctrl+Shift+Enter to close the Function Arguments dialog box and show the frequencies.**

   Use this keystroke combination because FREQUENCY is an array function. (For more on FREQUENCY, see Chapter 7.)

Finally, with I3:I9 highlighted, select

Insert | Column

and choose the Clustered Column layout to produce the graph of the frequencies. (See Chapter 3.) Your graph will probably look somewhat different from mine.

By the way, Excel repeats the random selection process whenever you do something that causes Excel to recalculate the worksheet. The effect is that the numbers can change as you work through this. For example, if you go back and autofill one of the rows again, the numbers change and the graph changes.

The Limits of Confidence

I told you about sampling distributions because they help you answer the question I pose at the beginning of this chapter: How much confidence can you have in the estimates you create?

The idea is to calculate a statistic, and then use that statistic to establish upper and lower bounds for the population parameter with, say, 95 percent

confidence. You can only do this if you know the sampling distribution of the statistic and the standard error. In the next section, I show how to do this for the mean.

Finding confidence limits for a mean

The FarBlonJet Corporation, a manufacturer of navigation systems, has developed a new battery to power their portable model. To help market their system, FarBlonJet wants to know how long, on average, each battery lasts before it burns out.

They'd like to estimate that average with 95 percent confidence. They test a sample of 100 batteries, and find that the sample mean is 60 hours, with a standard deviation of 20 hours. The Central Limit Theorem, remember, says that with a large enough sample (30 or more), the sampling distribution of the mean approximates a normal distribution. The standard error of the mean (the standard deviation of the sampling distribution of the mean) is

$$\sigma_{\bar{x}} = \sigma / \sqrt{N}$$

The sample size, *N*, is 100. What about σ? That's unknown, so you have to estimate it. If you know σ, that would mean you know μ, and establishing confidence limits would be unnecessary.

The best estimate of σ is the standard deviation of the sample. In this case that's 20. This leads to an estimate of the standard error of the mean

$$s_{\bar{x}} = s / \sqrt{N} = 20 / \sqrt{100} = 20 / 10 = 2$$

The best estimate of the population mean is the sample mean, 60. Armed with this information, — estimated mean, estimated standard error of the mean, normal distribution — you can envision the sampling distribution of the mean, which I've done in Figure 9-5. Consistent with Figure 9-2, each standard deviation is a standard error of the mean.

Now that you have the sampling distribution, you can establish the 95 percent confidence limits for the mean. This means that, starting at the center of the distribution, how far out to the sides do you have to extend until you have 95 percent of the area under the curve? (For more on area under the normal distribution and what it means, see Chapter 8.)

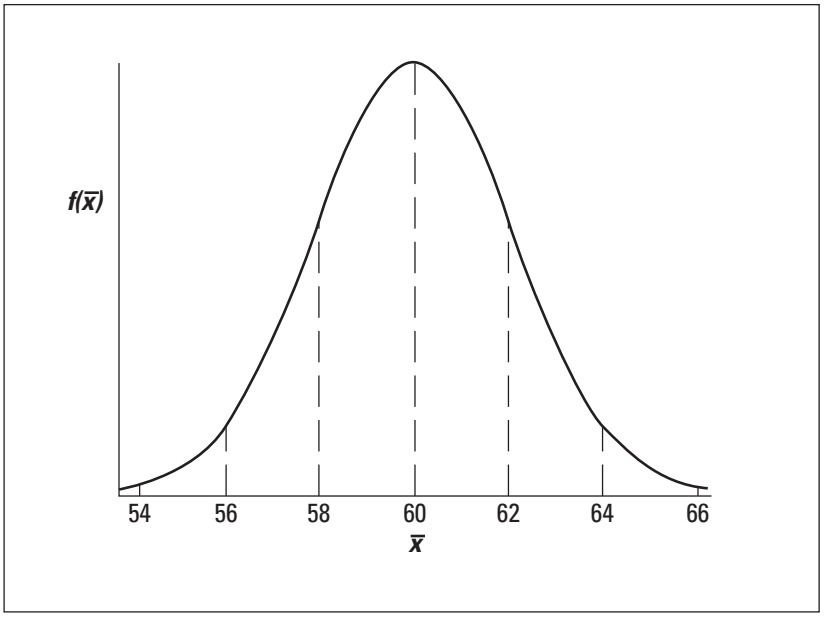

Figure 9-5: The sampling distribution of the mean for the FarBlonJet battery.

One way to answer this question is to work with the standard normal distribution and find the z-score that cuts off 47.5 percent on the right side and 47.5 percent on the left side (yes, Chapter 8 again). The one on the right is a positive z-score, the one on the left is a negative z-score. Then multiply each z-score by the standard error. Add each result to the sample mean to get the upper confidence limit and the lower confidence limit.

It turns out that the z-score is 1.96 for the boundary on the right side of the standard normal distribution, and –1.96 for the boundary on the left. You can calculate those values (difficult), get them from a table of the normal distribution that you typically find in a statistics textbook (easier), or use the Excel worksheet function I describe in the next section to do all the calculations (much easier). The point is that the upper bound in the sampling distribution is 63.92 ($60 + 1.96s_{\bar{x}}$), and the lower bound is 56.08 ($60 - 1.96s_{\bar{x}}$). Figure 9-6 shows these bounds on the sampling distribution.

This means you can say with 95 percent confidence that the FarBlonJet battery lasts, on the average, between 56.08 hours and 63.92 hours. Want a narrower range? You can either reduce your confidence level (to, say, 90 percent) or test a larger sample of batteries.

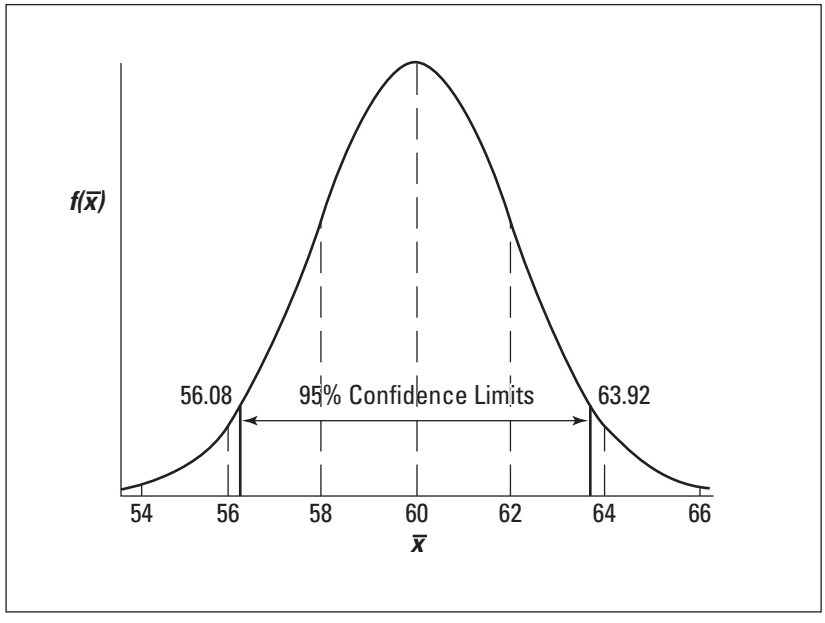

Figure 9-6: The 95 percent confidence limits on the FarBlonJet sampling distribution.

CONFIDENCE

The CONFIDENCE worksheet function does the lion's share of the work in constructing confidence intervals. You supply the confidence level, the standard deviation, and the sample size. CONFIDENCE returns the result of multiplying the appropriate z-score by the standard error of the mean. To determine the upper bound of the confidence limit, you add that result to the sample mean. To determine the lower bound, you subtract that result from the sample mean.

To show you how it works, I'll go through the FarBlonJet batteries example again. Here are the steps:

1. **Select a cell.**
2. **From the Statistical Functions menu, select CONFIDENCE to open the Function Arguments dialog box for CONFIDENCE.** (See Figure 9-7.)

   The Alpha box holds the result of subtracting the desired confidence level from 1.00.

Figure 9-7: The Function Arguments dialog box for CONFIDENCE.

Yes, that's a little confusing. Instead of typing .95 for the 95 percent confidence limit, I have to type .05. Think of it as the percentage of area *beyond* the confidence limits rather than the area *within* the confidence limits. And why is it labeled "Alpha"? I get into that in Chapter 10.

3. **In the Standard_dev box, I typed the standard deviation of the sample. For this example, the standard deviation is 20.**

   The Size box holds the number of individuals in the sample. The example specifies 100 batteries tested. After I typed that number, the answer (3.919928) appears in the dialog box.

4. **Click OK to put the answer into your selected cell.**

To finish things off, I add the answer to the sample mean (60) to determine the upper confidence limit (63.92) and subtract the answer from the mean to determine the lower confidence limit (56.08).

Fit to a t

The Central Limit Theorem specifies (approximately) a normal distribution for large samples. Many times, however, you don't have the luxury of large sample sizes, and the normal distribution isn't appropriate. What do you do?

For small samples, the sampling distribution of the mean is a member of a family of distributions called the *t-distribution*. The parameter that distinguishes members of this family from one another is called *degrees of freedom*.

Think of degrees of freedom as the denominator of your variance estimate. For example, if your sample consists of 25 individuals, the sample variance that estimates population variance is

$$s^2 = \frac{\sum\left(x-\bar{x}\right)^2}{N-1} = \frac{\sum\left(x-\bar{x}\right)^2}{25-1} = \frac{\sum\left(x-\bar{x}\right)^2}{24}$$

The number in the denominator is 24, and that's the value of the degrees of freedom parameter. In general, degrees of freedom (df) = N– 1 (N is the sample size) when you use the t-distribution the way I'm about to in this section.

Figure 9-8 shows two members of the t-distribution family (df =3 and df = 10), along with the normal distribution for comparison. As the Figure shows, the greater the df, the more closely t approximates a normal distribution.

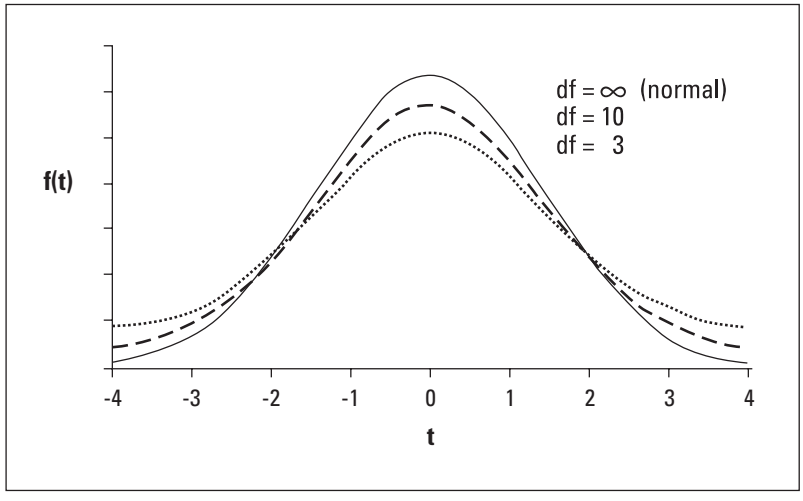

Figure 9-8: Some members of the t-distribution family.

So to determine the 95 percent confidence level if you have a small sample, work with the member of the t-distribution family that has the appropriate df. Find the value that cuts off 47.5 percent of the area on the right side of the distribution and 47.5 percent of the area on the left side of the distribution. The one on the right is a positive value, the one on the left is negative. Then multiply each value by the standard error. Add each result to the mean to get the upper confidence limit and the lower confidence limit.

In the FarBlonJet batteries example, suppose the sample consists of 25 batteries, with a mean of 60 and a standard deviation of 20. The estimate for the standard error of the mean is

$$s_{\bar{x}} = s/\sqrt{N} = 20/\sqrt{25} = 20/5 = 4$$

The df = N – 1 = 24. The value that cuts off 47.5 percent of the area on the right of this distribution is 2.064, and on the left it's –2.064. As I said earlier, you can calculate these values (difficult), look them up in a table that's in statistics textbooks (easier), or use the Excel function I describe in the next section (much easier).

The point is that the upper confidence limit is 68.256 (60 + 2.064$s_{\bar{x}}$) and the lower confidence limit is 51.744 (60 – 2.064). With a sample of 25 batteries, you can say with 95 percent confidence that the average life of a FarBlonJet battery is between 51.744 hours and 68.256 hours. Notice that with a smaller sample, the range is wider for the same level of confidence that I used in the previous example.

TINV

Excel's TINV worksheet function finds the value in the t-distribution that cuts off the desired area. Working with it is short and sweet:

1. **Select a cell.**
2. **From the Statistical Functions menu, select TINV to open the Function Arguments dialog box for TINV.**
3. **In the Function Arguments dialog box, enter the appropriate values for the arguments.**
4. **In the Probability box, enter the result of subtracting your confidence level from 1.00.**

   As I say in the description of the CONFIDENCE function, that's a bit confusing. Instead of typing .95 for the 95 percent confidence limit, I typed .05 in the Probability box. Think of it as the percentage of area *beyond* the confidence limits rather than the area *within* the confidence limits.
5. **In the Deg_freedom box I type the degrees of freedom.**

   For this example, df = 24. The answer appears in the dialog box.
6. **Click OK to close the dialog box and put the answer in the selected cell.** (See Figure 9-9.)

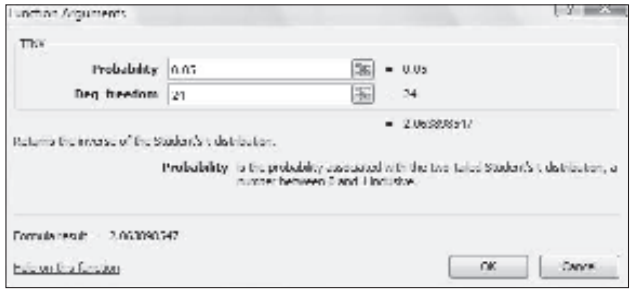

Figure 9-9: The Function Arguments dialog box for TINV.

You still have to multiply TINV's answer by the standard error of the mean and do the arithmetic to find the upper and lower limits.

I advise against using the CONFIDENCE worksheet function if your sample size is less than 30 and if you can't assume your population is a normal distribution. Why? CONFIDENCE always assumes a normally distributed sampling distribution, and that's not always appropriate. So if your confidence level is 95 percent, for example, CONFIDENCE multiplies the standard error by 1.96 regardless of the sample size. The result is that the confidence interval is too narrow for a small sample size.