Preparing for This Section Before getting started, review the following:

- Parameter versus statistic (Section 1.1, p. 33)
- z-scores (Section 3.4, p. 183–184)
- Standard normal distribution (Section 7.2, p. 394)
- Simple random sampling (Section 1.3, pp. 49–53)
- Degrees of freedom (Section 3.2, p. 163)
- Normal probability plots (Section 7.3, pp. 405–408)
- Distribution of the sample mean (Section 8.1, pp. 423–431)

Objectives
1. Obtain a point estimate for the population mean
2. State properties of Student's t-distribution
3. Determine t-values
4. Construct and interpret a confidence interval for a population mean
5. Determine the sample size needed to estimate a population mean within a specified margin of error

1 Obtain a Point Estimate for the Population Mean

Remember, the goal of statistical inference is to use information obtained from a sample and generalize the results to the population being studied. As with estimating the population proportion, the first step is to obtain a point estimate of the parameter. The point estimate of the population mean, μ, is the sample mean, $\bar{x}$.

EXAMPLE 1 Computing a Point Estimate of the Population Mean

Problem The website fueleconomy.gov allows drivers to report the miles per gallon of their vehicle. The data in Table 2 show the reported miles per gallon of 2011 Ford Focus automobiles for 16 different owners. Obtain a point estimate of the population mean miles per gallon of a 2011 Ford Focus.

Table 2

35.7	37.2	34.1	38.9
32.0	41.3	32.5	37.1
37.3	38.8	38.2	39.6
32.2	40.9	37.0	36.0

Source: www.fueleconomy.gov

Approach Treat the 16 entries as a simple random sample of all 2011 Ford Focus automobiles. To find a point estimate of the population mean, compute the sample mean miles per gallon of the 16 cars.

Solution The sample mean is

$$\bar{x} = \frac{35.7 + 32.0 + \cdots + 36.0}{16} = \frac{588.8}{16} = 36.8 \text{ miles per hour}$$

The point estimate of μ is 36.8 miles per hour. Remember, round statistics to one more decimal point than the raw data if necessary. ●

2 State Properties of Student's t-distribution

In Example 1, a different random sample of 16 cars would likely result in a different point estimate of μ. For this reason, we want to construct a confidence interval for the population mean, just as we did for the population proportion.

A confidence interval for the population mean is of the form point estimate ± margin of error (just like the confidence interval for a population proportion). To determine the margin of error, we need to know the sampling distribution of the sample mean. Recall that the distribution of $\bar{x}$ is approximately normal if the population from which the sample is drawn is normal or the sample size is sufficiently large. In addition, the distribution of $\bar{x}$ has the same mean as the parent population, $\mu_{\bar{x}} = \mu$, and a standard deviation equal to the parent population's standard deviation divided by the

square root of the sample size, $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$. Following the same logic used in constructing a confidence interval about a population proportion, our confidence interval would be

$$\text{point estimate} \ \pm \ \text{margin of error}$$

$$\bar{x} \ \pm \ z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

This presents a problem because we need to know the population standard deviation to construct this interval. It does not seem likely that we would know the population standard deviation but not know the population mean. So what can we do? A logical option is to use the sample standard deviation, s, as an estimate of σ. Then the standard deviation of the distribution of $\bar{x}$ would be estimated by $\dfrac{s}{\sqrt{n}}$ and our confidence interval would be

$$\bar{x} \ \pm \ z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \tag{1}$$

Unfortunately, there is a problem with this approach. The sample standard deviation, s, is a statistic and therefore will vary from sample to sample. Using the normal model to determine the critical value, $z_{\frac{\alpha}{2}}$, in the margin of error does not take into account the additional variability introduced by using s in place of σ. This is not much of a problem for large samples because the variability in the sample standard deviation decreases as the sample size increases (Law of Large Numbers), but for small samples, we have a real problem. Put another way, the z-score of $\bar{x}$, $\dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$, is normally distributed with mean 0 and standard deviation 1 (provided $\bar{x}$ is normally distributed). However, $\dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$ is *not* normally distributed with mean 0 and standard deviation 1. So a new model must be used to determine the margin of error in a confidence interval that accounts for this additional variability. This leads to the story of William Gosset.

In the early 1900s, William Gosset worked for the Guinness brewery. Gosset was in charge of conducting experiments at the brewery to identify the best barley variety. When working with beer, Gosset was limited to small data sets. At the time, the model used for constructing confidence intervals about a mean was the normal model, regardless of whether the population standard deviation was known. Gosset did not know the population standard deviation, so he simply substituted the sample standard deviation for the population standard deviation as suggested by Formula (1). While doing this, he was finding that his confidence intervals did not include the population mean at the rate expected. This led Gosset to develop a model that accounts for the additional variability introduced by using s in place of σ when determining the margin of error. Guinness would not allow Gosset to publish his results under his real name (Guinness was very secretive about its brewing practices), but did allow the results to be published under a pseudonym. Gosset chose Student. So we have Student's t-distribution.

Student's t-Distribution

Suppose that a simple random sample of size n is taken from a population. If the population from which the sample is drawn follows a normal distribution, the distribution of

$$t = \frac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$$

follows Student's t-distribution with $n - 1$ degrees of freedom,* where $\bar{x}$ is the sample mean and s is the sample standard deviation.

* The reader may wish to review the discussion of degrees of freedom in Section 3.2 on p. 163.

The interpretation of t is the same as that of the z-score. The t-statistic represents the number of *sample* standard errors $\bar{x}$ is from the population mean, μ. It turns out that the shape of the t-distribution depends on the sample size, n.

To help see how the t-distribution differs from the standard normal (or z-) distribution and the role that the sample size n plays, we will go through the following simulation.
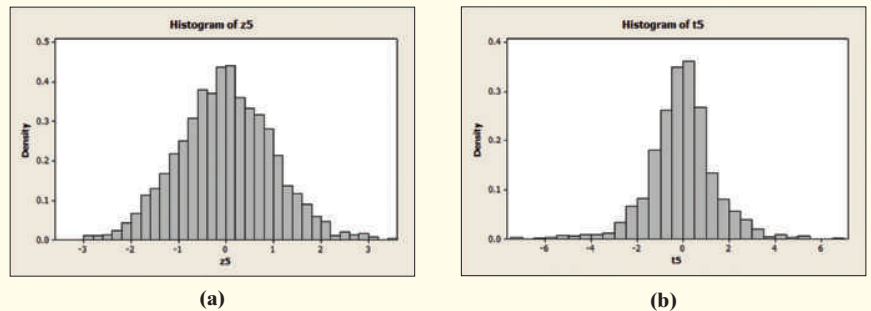
EXAMPLE 2   Comparing the Standard Normal Distribution to the t-Distribution Using Simulation

(a) Use statistical software such as Minitab or StatCrunch to obtain 1500 simple random samples of size $n = 5$ from a normal population with $\mu = 50$ and $\sigma = 10$. Calculate the sample mean and sample standard deviation for each sample. Compute $z = \dfrac{\bar{x} - \mu_{\bar{x}}}{\dfrac{\sigma}{\sqrt{n}}}$ and $t = \dfrac{\bar{x} - \mu_{\bar{x}}}{\dfrac{s}{\sqrt{n}}}$ for each sample. Draw histograms for both z and t.

(b) Repeat part (a) for 1500 simple random samples of size $n = 10$.

Solution

(a) We use Minitab to obtain the 1500 simple random samples and compute the 1500 sample means and sample standard deviations. We then compute $z = \dfrac{\bar{x} - \mu_{\bar{x}}}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{\bar{x} - 50}{\dfrac{10}{\sqrt{5}}}$ and $t = \dfrac{\bar{x} - \mu_{\bar{x}}}{\dfrac{s}{\sqrt{n}}} = \dfrac{\bar{x} - 50}{\dfrac{s}{\sqrt{5}}}$ for each of the 1500 samples. Figure 9(a) shows the histogram for z, and Figure 9(b) shows the histogram for t.
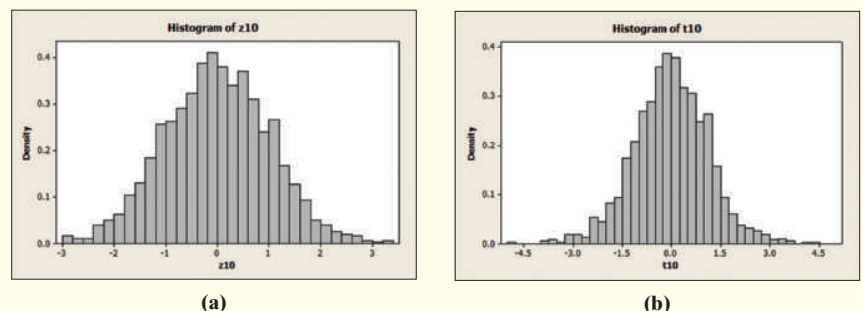
Figure 9

(a)   (b)

Notice that the histogram in Figure 9(a) is symmetric and bell shaped, with the histogram centered at 0, and virtually all the rectangles lying between -3 and 3. In other words, z with $n = 5$ follows a standard normal distribution. The histogram of t with $n = 5$ is also symmetric, bell shaped, and centered at 0, but the histogram of t has longer tails (that is, t is more dispersed), so it is unlikely that t follows a standard normal distribution. This additional spread is due to the fact that we divided by $\dfrac{s}{\sqrt{n}}$ to find t instead of by $\dfrac{\sigma}{\sqrt{n}}$.

(b) Repeat part (a) for samples of size $n = 10$. Figure 10(a) shows the histogram for z, and Figure 10(b) shows the histogram for t. What do you notice?

Figure 10

(a)   (b)

(continued)

The histogram in Figure 10(a) is symmetric, bell shaped, centered at 0, and virtually all the rectangles lie between -3 and 3. In other words, z with $n = 10$ follows a standard normal distribution. The histogram of t with $n = 10$ is also symmetric, bell shaped, and centered at 0, but the histogram has longer tails (that is, t is more dispersed) than the histogram of z with $n = 10$. So t with $n = 10$ also does not appear to follow the standard normal distribution.

One very important distinction must be made. The distribution of t with $n = 10$ (Figure 10(b)) is less dispersed than the distribution of t with $n = 5$ (Figure 9(b)).

We conclude that there are different t distributions for different sample sizes. In addition, the spread in the distribution of t decreases as the sample size n increases. In fact, it can be shown that as the sample size n increases, the distribution of t behaves more and more like the standard normal distribution.   •

Properties of the t-Distribution

1. The t-distribution is different for different degrees of freedom.
2. The t-distribution is centered at 0 and is symmetric about 0.
3. The area under the curve is 1. The area under the curve to the right of 0 equals the area under the curve to the left of 0, which equals $\frac{1}{2}$.
4. As t increases or decreases without bound, the graph approaches, but never equals, zero.
5. The area in the tails of the t-distribution is a little greater than the area in the tails of the standard normal distribution, because we are using s as an estimate of σ, thereby introducing further variability into the t-statistic.
6. As the sample size n increases, the density curve of t gets closer to the standard normal density curve. This result occurs because, as the sample size increases, the values of s get closer to the value of σ, by the Law of Large Numbers.
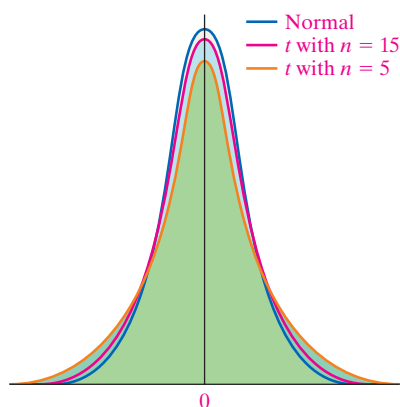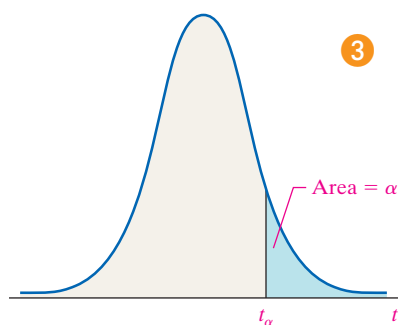
Figure 11

- Normal
- t with $n = 15$
- t with $n = 5$

0

In Figure 11, we show the t-distribution for the sample sizes $n = 5$ and $n = 15$, along with the standard normal density curve.

Figure 12

Area = α

t_α t

③ Determine t-Values

Recall that the notation z_α is used to represent the z-score whose area under the normal curve to the right of z_α is α. Similarly let t_α represent the t-value whose area under the t-distribution to the right of t_α is α. See Figure 12.

The shape of the t-distribution depends on the sample size, n. Therefore, the value of t_α depends not only on α, but also on the degrees of freedom, $n - 1$. In Table VII in Appendix A, the far left column gives the degrees of freedom. The top row represents the area under the t-distribution to the right of some t-value.

EXAMPLE 3   Finding t-Values

Problem Find the t-value such that the area under the t-distribution to the right of the t-value is 0.10, assuming 15 degrees of freedom (df). That is, find $t_{0.10}$ with 15 degrees of freedom.

Approach

Step 1 Draw a t-distribution with the unknown t-value labeled. Shade the area under the curve to the right of the t-value, as in Figure 12.

Step 2 Find the row in Table VII corresponding to 15 degrees of freedom and the column corresponding to an area in the right tail of 0.10. Identify where the row and column intersect. This is the unknown t-value.
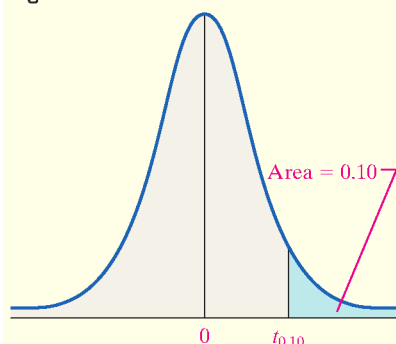
Figure 13

Solution

Step 1 Figure 13 shows the graph of the *t*-distribution with 15 degrees of freedom. The unknown value of *t* is labeled, and the area under the curve to the right of *t* is shaded.

Step 2 A portion of Table VII is shown in Figure 14. We have enclosed the row that represents 15 degrees of freedom and the column that represents the area 0.10 in the right tail. The value where the row and column intersect is the *t*-value we are seeking. The value of $t_{0.10}$ with 15 degrees of freedom is 1.341; that is, the area under the *t*-distribution to the right of $t = 1.341$ with 15 degrees of freedom is 0.10.

Figure 14

df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
							Area in Right Tail					
1	1.000	1.376	1.963	3.078	6.314	12.706	15.894	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015

Notice that the critical value of *z* with an area of to the right of 0.10 is smaller—approximately 1.28. This is because the *t*-distribution has more spread than the *z*-distribution. •

• **Now Work Problem 7**

🖥 Using Technology

The TI-84 Plus graphing calculator has an invT feature, which finds the value of *t* given an area to the left of the unknown *t*-value and the degrees of freedom.

If the degrees of freedom we desire are not listed in Table VII, choose the closest number in the "df" column. For example, if we have 43 degrees of freedom, we use 40 degrees of freedom from Table VII. In addition, the last row of Table VII lists the *z*-values from the standard normal distribution. Use these values when the degrees of freedom are more than 1000 because the *t*-distribution starts to behave like the standard normal distribution as *n* increases.

④ Construct and Interpret a Confidence Interval for a Population Mean

We are now ready to construct a confidence interval for a population mean.

Constructing a $(1 - \alpha) \cdot 100\%$ Confidence Interval for μ

Provided

- sample data come from a simple random sample or randomized experiment,
- sample size is small relative to the population size ($n \leq 0.05N$), and
- the data come from a population that is normally distributed, or the sample size is large.

A $(1 - \alpha) \cdot 100\%$ confidence interval for μ is given by

$$\text{Lower bound: } \bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \qquad \text{Upper bound: } \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \tag{2}$$

where $t_{\frac{\alpha}{2}}$ is the critical value with $n - 1$ degrees of freedom.

Because this confidence interval uses the *t*-distribution, it is often referred to as a ***t*-interval**.

The **margin of error** for constructing confidence intervals about a mean is

$$E = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

Notice that a confidence interval about μ can be computed for non-normal populations even though Student's *t*-distribution requires a normal population. This is because the procedure for constructing the confidence interval is **robust**—it is accurate despite minor departures from normality. If a data set has outliers, the confidence interval is not accurate because neither the sample mean nor the sample standard deviation is resistant to outliers. Sample data should always be inspected for serious departures from normality and for outliers. This is easily done with normal probability plots and boxplots.

EXAMPLE 4   Constructing a Confidence Interval about a Population Mean

Table 3

35.7	37.2	34.1	38.9
32.0	41.3	32.5	37.1
37.3	38.8	38.2	39.6
32.2	40.9	37.0	36.0

Source: www.fueleconomy.gov

Problem The website fueleconomy.gov allows drivers to report the miles per gallon of their vehicle. The data in Table 3 show the reported miles per gallon of 2011 Ford Focus automobiles for 16 different owners. Treat the sample as a simple random sample of all 2011 Ford Focus automobiles. Construct a 95% confidence interval for the mean miles per gallon of a 2011 Ford Focus. Interpret the interval.

Approach

Step 1 Verify the data are obtained randomly and the sample size is small relative to the population size. Because the sample size is small, draw a normal probability plot to verify the data come from a population that is normally distributed and a boxplot to verify that there are no outliers.

By-Hand Approach

Step 2 Compute the value of $\bar{x}$ and s.

Step 3 Determine the critical value $t_{\frac{\alpha}{2}}$ with $n - 1$ degrees of freedom.

Step 4 Use Formula (2) to determine the lower and upper bounds of the confidence interval.
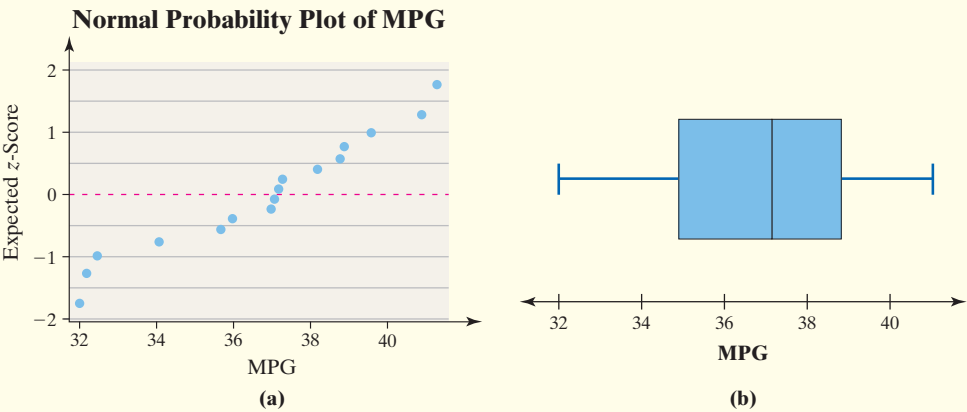
Step 5 Interpret the result.

Technology Approach

Step 2 Use a statistical spreadsheet or graphing calculator with advanced statistical features to obtain the confidence interval. We will use Minitab to construct the confidence interval. The steps for constructing confidence intervals using the TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on page 470.

Step 3 Interpret the result.

Solution

Step 1 The data are obtained from a simple random sample. In addition, there are likely thousands of 2011 Ford Focus vehicles on the road, so the sample size is small relative to the population size. Figure 15 shows a normal probability plot and boxplot for the data in Table 3. The correlation between MPG and the expected *z*-scores is 0.979. Because 0.979 > 0.941 (Table VI), it is reasonable to conclude the sample data come from a population that is normally distributed. The boxplot does not reveal any outliers. The requirements for constructing the confidence interval are satisfied.

Figure 15

(a)

(b)

By-Hand Solution

Step 2 We determined the sample mean in Example 1 to be $\bar{x} = 36.8$ mpg. Using a calculator, the sample standard deviation is $s = 2.92$ mpg.

Step 3 Because we want a 95% confidence interval, we have $\alpha = 1 - 0.95 = 0.05$. The sample size is $n = 16$. So we find $t_{\frac{\alpha}{2}} = t_{\frac{0.05}{2}} = t_{0.025}$ with $16 - 1 = 15$ degrees of freedom. Table VII shows that $t_{0.025} = 2.131$.

Step 4 Substituting into Formula (2), we obtain:

Lower bound:

$$\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 36.8 - 2.131 \cdot \frac{2.92}{\sqrt{16}} = 35.24$$

Upper bound:

$$\bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 36.8 + 2.131 \cdot \frac{2.92}{\sqrt{16}} = 38.36$$

Step 5 We are 95% confident that the mean miles per gallon of all 2011 Ford Focus cars is between 35.24 and 38.36 mpg.

• **Now Work Problem 31**

Technology Solution

Step 2 Figure 16 shows the results from Minitab.

Figure 16

```
One-Sample T: MPG

Variable   N    Mean  StDev  SE Mean       95% CI
MPG        16  36.800  2.917   0.729  (35.246, 38.354)
```

Minitab presents confidence intervals in the form (*lower bound, upper bound*). The lower bound is 35.25 and the upper bound is 38.35.

Step 3 We are 95% confident that the mean miles per gallon of all 2011 Ford Focus cars is between 35.25 and 38.35 mpg.

Notice that $t_{0.025} = 2.131$ for 15 degrees of freedom, while $z_{0.025} = 1.96$. The t-distribution gives a larger critical value, so the width of the interval is wider. Remember, this larger critical value is necessary to account for the increased variability due to using s as an estimate of σ.

Remember, 95% confidence refers to our confidence in the method. If we obtained 100 samples of size $n = 16$ from the population of 2011 Ford Focuses, we would expect about 95 of the samples to result in confidence intervals that include μ. We do not know whether the interval in Example 4 includes μ or does not include μ.

What should we do if the requirements to compute a t-interval are not met? We could increase the sample size beyond 30 observations, or we could try to use *nonparametric procedures*. **Nonparametric procedures** typically do not require normality, and the methods are resistant to outliers. A third option is to use resampling methods, such as bootstrapping, introduced in Section 9.5.

⑤ Determine the Sample Size Needed to Estimate a Population Mean within a Specified Margin of Error

The margin of error in constructing a confidence interval about the population mean is $E = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$. Solving this for n, we obtain $n = \left(\frac{t_{\frac{\alpha}{2}} \cdot s}{E} \right)^2$. The problem with this formula is that the critical value $t_{\frac{\alpha}{2}}$ requires that we know the sample size to determine the degrees of freedom, $n - 1$. Obviously, if we do not know n we cannot know the degrees of freedom. The solution to this problem lies in the fact that the t-distribution approaches the standard normal z-distribution as the sample size increases. To convince yourself of this, look at the last few rows of Table VII and compare them with the corresponding z-scores for 95% or 99% confidence. Now if we use z in place of t and a sample standard deviation, s, from previous or pilot studies, we can write the margin of error formula as $E = z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$ and solve it for n to obtain a formula for determining sample size.

Determining the Sample Size *n*

The sample size required to estimate the population mean, μ, with a level of confidence $(1 - \alpha) \cdot 100\%$ within a specified margin of error, E, is given by

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot s}{E} \right)^2 \qquad\qquad \textbf{(3)}$$

where *n* is *rounded up* to the nearest whole number.

EXAMPLE 5   Determining Sample Size

Using Technology

StatCrunch has the ability to determine sample size. See the Technology Step-by-Step below.

Problem We again consider the problem of estimating the miles per gallon of a 2011 Ford Focus. How large a sample is required to estimate the mean miles per gallon within 0.5 mile per gallon with 95% confidence?

Approach Use Formula (3) with $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, $s = 2.92$, and $E = 0.5$ to find the required sample size.

Solution Substitute the values of z, s, and E into Formula (3) and obtain

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot s}{E} \right)^2 = \left(\frac{1.96 \cdot 2.92}{0.5} \right)^2 = 131.02$$

• **Now Work Problem 41**

Round 131.02 up to 132. A sample size of $n = 132$ results in an interval estimate of the population mean miles per gallon of a 2011 Ford Focus with a margin of error of 0.5 mile per gallon with 95% confidence.   •

Technology Step-by-Step   Confidence Intervals for μ

TI-83/84 Plus
1. If necessary, enter raw data in L1.
2. Press STAT, highlight TESTS, and select 8:TInterval.
3. If the data are raw, highlight DATA. Make sure List is set to L1 and Freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics.
4. Enter the confidence level following C-Level:.
5. Highlight Calculate; press ENTER.

Minitab
1. If you have raw data, enter them in column C1.
2. Select the **Stat** menu, then **Basic Statistics**, then highlight **1-Sample *t* . . . .**
3. If you have raw data, select "One or more samples, each in a column" from the pull-down menu. Place the cursor in the box, highlight the column containing the raw data, and click "Select." If you have summarized data, select "Summarized data" from the pull-down menu and enter the summarized data. Select **Options . . .** and enter a confidence level. Click OK twice.

Excel
1. Load the XLSTAT Add-in.
2. Enter the raw data in Column A.
3. Select the **XLSTAT** menu, highlight **Parametric tests**. Select **One-sample t-test and z-test**.
4. Place the cursor in the Data cell. Highlight the raw data in the spreadsheet. Be sure the box for Student's *t* test is checked and the radio button for One sample is selected.

Click the Options tab. For a 90% confidence interval, let the Significance level (%) equal 10; for a 95% confidence interval, let the Significance level (%) equal 5, and so on. Click OK.

StatCrunch
Constructing a Confidence Interval for the Population Mean
1. If necessary, enter the raw data into column var1. Name the column.
2. Select **Stat**, highlight **T Stats**, highlight **One Sample**. Choose **With Data** if you have raw data, choose **With Summary** if you have summarized data.
3. If you chose **With Data**, highlight the column that contains the data in "Select column(s):". If you chose **With Summary**, enter the sample mean, sample standard deviation, and sample size. Choose the confidence interval radio button. Enter the level of confidence. Click Compute!.

Determining Sample Size When Estimating a Population Mean
1. Select **Stat**, highlight **Z Stats**, highlight **One Sample**, and highlight **Power/Sample Size**. Note: You may also highlight **T Stats** and follow the same steps.
2. Click on the "Confidence Interval Width" tab. Enter the Confidence level and standard deviation. The width is the difference between the lower bound and the upper bound in the confidence interval. Therefore, the width is two times the margin of error. Clear any entry in the sample size cell. Click Compute.