Chapter 9 Estimating the Value of a Parameter

Section 9.1 Estimating a Population Proportion

Objectives

1. Obtain a point estimate for the population proportion
2. Construct and interpret a confidence interval for the population proportion
3. Determine the sample size necessary for estimating a population proportion within a specified margin of error

❶ Obtain a Point Estimate for the Population Proportion

A **point estimate** is the value of a statistic that estimates the value of a parameter.

Example Obtaining a Point Estimate of a Population Proportion

In a survey of 2019 adult Americans aged 18 years or older, 1252 stated that they frequently worry about their financial situation. Find the sample proportion of adult Americans aged 18 years or older who frequently worry about their financial situation.

Note: Round proportions to three decimal places.

❷ Construct and Interpret a Confidence Interval for the Population Proportion

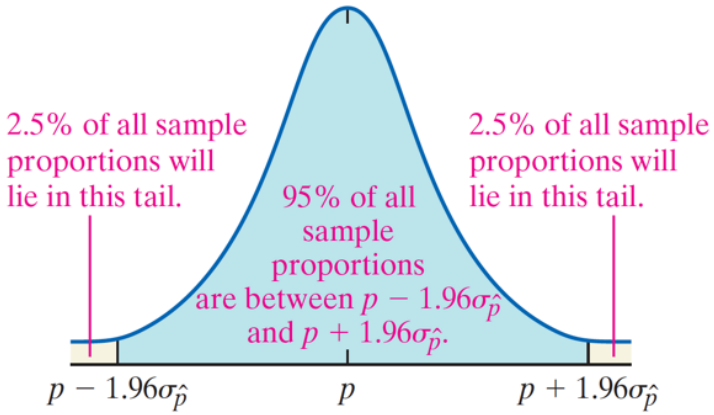
- A **confidence interval** for an unknown parameter consists of an interval of numbers based on a point estimate.
- The **level of confidence** represents the expected proportion of intervals that will contain the parameter if a large number of different samples is obtained. The level of confidence is denoted $(1-\alpha)\cdot 100\%$.

Two questions.

1. Why does the level of confidence represent the expected proportion of intervals that contain the parameter if a large number of different samples is obtained?
2. How is the margin of error determined?

A review of the sampling distribution of the sample proportion, $\hat{p}$.

- The shape of the distribution of all possible sample proportions is approximately normal provided $np(1-p) \geq 10$ and the sample size is no more than 5% of the population size.
- The mean of the distribution of the sample proportion equals the population proportion. That is, $\mu_{\hat{p}} = p$.
- The standard deviation of the distribution of the sample proportion (the standard error) is $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$.



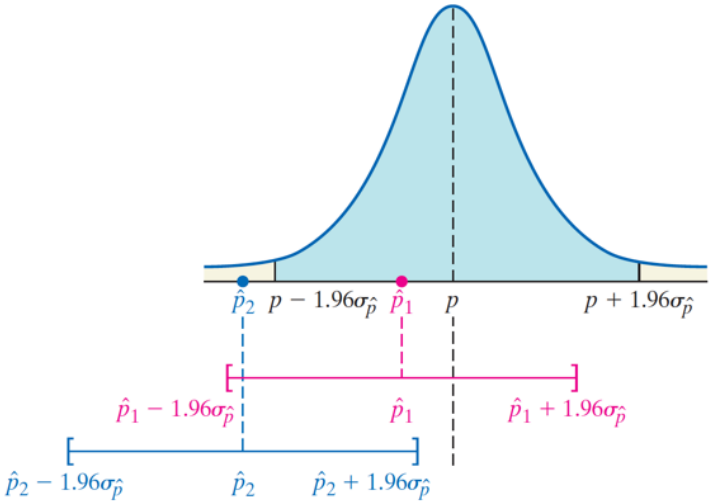
95% of all sample proportions are in the following inequality:

Rewrite the inequality with the population proportion, p, in the middle:

So, 95% of *all* sample proportions will result in confidence interval estimates that contain the population proportion, while 5% of *all* sample proportions (those in the tails of the distribution above), will result in confidence interval estimates that do not contain the population proportion.

Write the 95% confidence interval in the form *point estimate* $\pm$ *margin of error*:

What is the margin of error for a 95% confidence interval?



Activity Illustrating the Meaning of Level of Confidence

Go to StatCrunch and select Applets > Confidence intervals > for a proportion. Check the "Proportion with characteristic" radio button and set p to 0.3. Click Compute! Or go to www.pearsonhigher.com/sullivanstats and select the "Confidence Intervals for a Proportion with p = 0.3" applet. Change the sample size to 150 (this is to ensure the sampling distribution of the sample proportion is approximately normal).
(a) Click "100 intervals" two times to generate confidence intervals for 200 independent simple random samples of size n = 150 from a population with p = 0.3. What proportion of the 200 intervals include the population proportion?

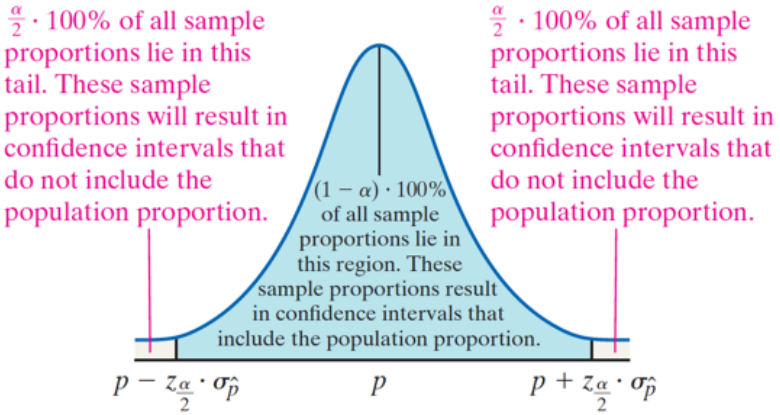
(b) What causes an interval to not include the population proportion?

Whether a confidence interval contains the population parameter depends solely on the value of the sample statistic. Any sample statistic that is in the tails of the sampling distribution will result in a confidence interval that does not include the population parameter.

CAUTION!

A 95% confidence interval does *not* mean that there is a 95% probability that the interval contains the parameter (such as p or μ). Remember, probability describes the likelihood of undetermined events. Therefore, it does not make sense to talk about the probability that the interval contains the parameter since the parameter is a fixed value. Think of it this way: I flip a coin and obtain a head. If I ask you to determine the probability that the flip resulted in a head, it would not be 0.5, because the outcome has already been determined. Instead, the probability is 0 or 1. Confidence intervals work the same way. Because p or μ is already determined, we do not say that there is a 95% probability that the interval contains p or μ.

Constructing Any $(1-\alpha)\cdot 100\%$ Confidence Interval



The value $z_{\frac{\alpha}{2}}$ is called the **critical value**.

Level of Confidence, $(1-\alpha)\cdot 100\%$	Area in Each Tail, $\frac{\alpha}{2}$	Critical Value, $z_{\frac{\alpha}{2}}$
90%	0.05	1.645
95%	0.025	1.96
99%	0.005	2.575

Interpretation of a Confidence Interval

A $(1-\alpha)\cdot 100\%$ confidence interval indicates that $(1-\alpha)\cdot 100\%$ of all simple random samples of size n from the population whose parameter is unknown will result in an interval that contains the parameter.

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

Suppose that a simple random sample of size n is taken from a population or the data are the result of a randomized experiment. A $(1-\alpha)\cdot 100\%$ confidence interval for p is given by the following quantities:

$$\text{Lower bound: } \hat{p} - z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad \text{Upper bound: } \hat{p} + z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{2}$$

Note: It must be the case that $n\hat{p}(1-\hat{p}) \geq 10$ and $n \leq 0.05N$ to construct this interval.

The **margin of error, *E***, in a $(1-\alpha)\cdot 100\%$ confidence interval for a population proportion is given by

$$E = z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{3}$$

Example Constructing a Confidence Interval for a Population Proportion

In a survey of 2019 adult Americans aged 18 years or older, 1252 stated that they frequently worry about their financial situation. Obtain a 90% confidence interval for the proportion of adult Americans who frequently worry about their financial situation.

Example The Role of the Level of Confidence on the Margin of Error

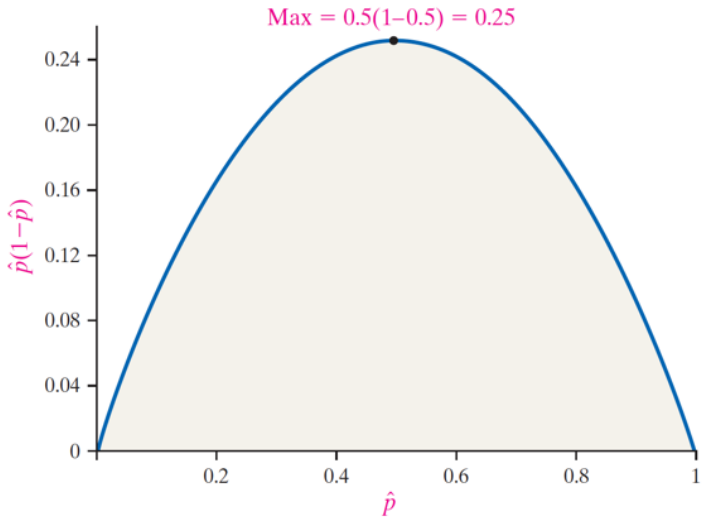
Redo the previous example by increasing the level of confidence from 90% to 95%.

Example The Role of the Sample Size on the Margin of Error

Suppose the number of individuals surveyed is increased four-fold to 8076 (but the sample proportion remains at 0.620. Compute the margin of error. Compare this margin of error to that when the sample size is 2019.

❸ Determine the Sample Size Necessary for Estimating a Population Proportion within a Specified Margin of Error

Solve $E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ for n.



Sample Size Needed for Estimating the Population Proportion p

The sample size required to obtain a $(1 - \alpha) \cdot 100\%$ confidence interval for p with a margin of error E is given by

$$n = \hat{p}(1 - \hat{p})\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 \quad \textbf{(4)}$$

rounded up to the next integer, where $\hat{p}$ is a prior estimate of p.

If a prior estimate of p is unavailable, the sample size required is

$$n = 0.25\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 \quad \textbf{(5)}$$

rounded up to the next integer.

The margin of error should always be expressed as a decimal when using Formulas (4) and (5).

Example Determining Sample Size

A social worker wants to know the proportion of the U.S. population over age 16 who provide eldercare – unpaid care for someone with a condition related to aging – to others. What size sample should be obtained if the social worker wants an estimate within 3 percentage points of the true proportion with 95% confidence if

(a) the social worker uses the 2015 estimate from the American Time Use Survey of 16%.
(b) the social worker does not use any prior estimates?

9.2 Estimating a Population Mean

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 the population mean within a specified margin of error

The point estimate of the population mean, μ , is the sample mean, $\bar{x}$.

Example Computing a Point Estimate of the Population Mean

The following data represent the amount spent on Valentine's Day among males who purchased gifts for the holiday. The data is based on reports from fundivo.com.

$216	$156	$125	$133	$229	$155
$146	$120	$189	$201	$137	$262

Obtain a point estimate of the population mean amount spent on Valentine's Day among males who purchased gifts for the holiday.

❷ State Properties of Student's *t*-distribution

Using the logic used in constructing a confidence interval about a population proportion, the formula for a confidence interval about a population mean should be of the form

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

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

The problem with this formula is that we do not know the population standard deviation, σ . The solution to this problem would be to replace σ with the sample standard deviation, *s*. The "new" formula for the confidence interval will be

$$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \quad \textbf{(1)}$$

The problem with this approach is that the sample standard deviation is a random variable. We cannot use the critical value $z_{\frac{\alpha}{2}}$ in this formula because $\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ does not follow the normal distribution with mean 0 and standard deviation 1. The solution is to use a new probability distribution called *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

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

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.

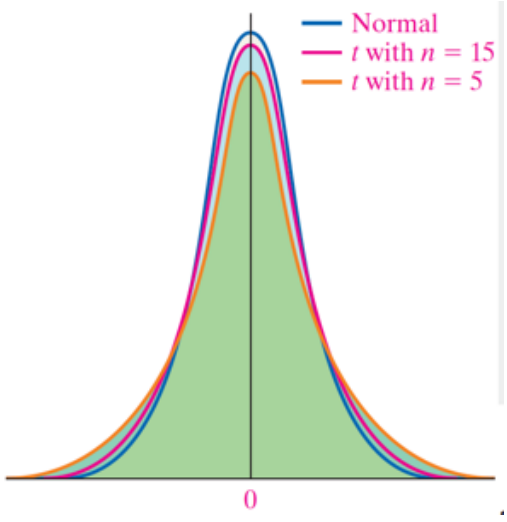
Activity Comparing the Standard Normal Distribution to the *t*-Distribution Using Simulation

(a) Use StatCrunch to obtain 2000 simple random samples of size n = 9 from a normal population with μ = 100 and standard deviation σ= 15. Calculate the sample mean and sample standard deviation for each sample. Compute $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x}-100}{\frac{15}{\sqrt{9}}} = \frac{\bar{x}-100}{5}$ and $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x}-100}{\frac{s}{\sqrt{9}}}$ for each sample. Draw histograms fro both z and t.

(b) Repeat part (a) for 2000 simple random samples of size $n = 16$.

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.



❸ Determine t-Values

The notation α is the *z*-score such that the area under the standard normal curve to the right of z_α is α. Similarly, α will be the *t*-value such that the area under the *t*-distribution to the right of t_α will be α.

Example Finding *t*-Values

Find the *t*-value such that the area under the *t*-distribution to the right of the *t*-value is 0.05, assuming 20 degrees of freedom. That is, find $t_{0.05}$ with 20 degrees of freedom.

❹ Construct and Interpret 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 $(n \le 0.05N)$, 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

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

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

The procedure for constructing a confidence interval about the population mean using Student's *t*-distribution is **robust** – meaning it is accurate despite minor departures from normality.

Example Constructing a Confidence Interval about a Population Mean

The following data represent the amount spent on Valentine's Day among males who purchased gifts for the holiday. The data is based on reports from fundivo.com. Construct a 90% confidence interval for the mean amount spent on Valentine's Day among males who purchased gifts for the holiday.

$216	$156	$125	$133	$229	$155
$146	$120	$189	$201	$137	$262

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

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

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

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

Example Determining Sample Size

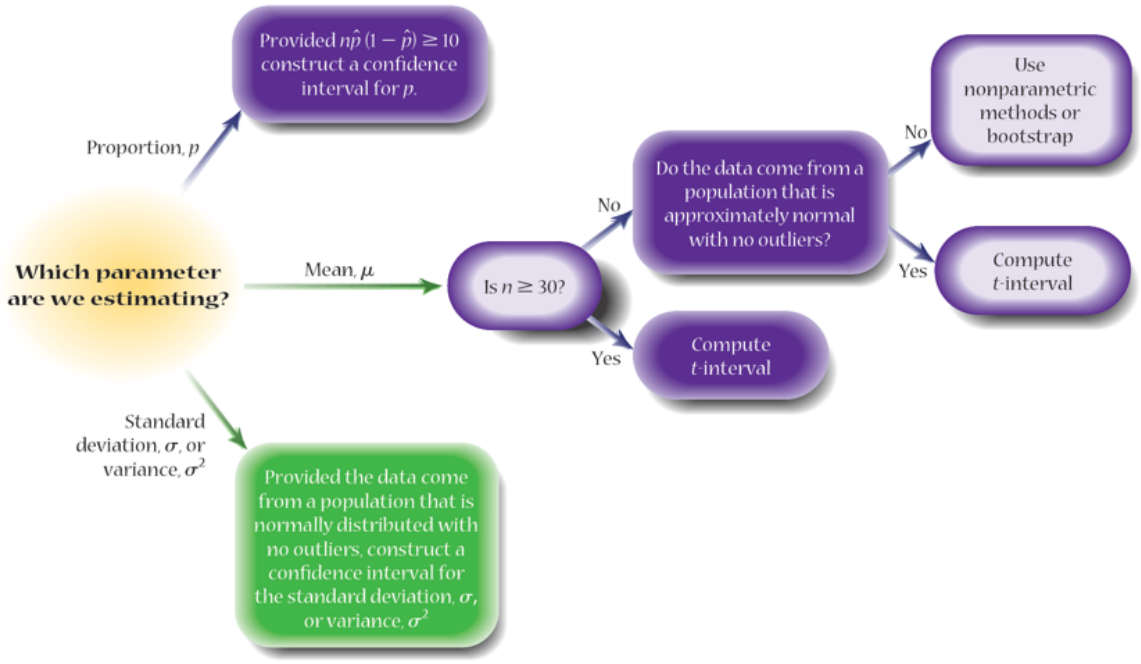
Again consider the Valentine's Day gift data. How large a sample is required to estimate the mean amount spent on Valentine's Day among males who purchased gifts for the holiday within $10 with 90% confidence.

9.4 Putting It Together: Which Procedure Do I Use?

Objective

1. Determine the Appropriate Confidence Interval to Construct

❶ Determine the Appropriate Confidence Interval to Construct



Example Constructing a Confidence Interval

In a recent survey of 2221 adult Americans, 1666 indicated that they turn off lights, televisions, or other appliances when not in use.

(a) What is the variable of interest in this study? Is it qualitative or quantitative?
(b) What type of confidence interval would make sense to construct for this variable of interest?
(c) Verify the model requirements to construct the confidence interval are satisfied.
(d) Construct a 95% confidence interval for the variable of interest.

Example Constructing a Confidence Interval

The following data represent the pitch speed, in miles per hour, of a random sample of 12 sliders thrown by Jake Arrietta, pitcher for the Chicago Cubs.

88.09	89.01	90.45	89.19	89.74	86.71
88.58	88.82	86.92	87.44	89.07	89.00

Source: BrooksBaseball.net

(a) What is the variable of interest in this study? Is it qualitative or quantitative?
(b) What type of confidence interval would make sense to construct for this variable of interest?
(c) Verify the model requirements to construct the confidence interval are satisfied.
(d) Construct a 90% confidence interval for the variable of interest.

Bootstrapping

Objective

1. Estimate a parameter using the bootstrap method

The methods for estimating parameters (such as μ or p) by constructing confidence intervals rely on parametric methods. **Parametric statistics** use estimates of parameters along with probability distributions to make inferences about the parameter. For example, we use the normal probability distribution to make inferences about the population proportion by constructing a confidence interval. To make inferences about a parameter using parametric statistics certain conditions (such as a normality requirement) must be satisfied. What if the conditions for using parametric statistics are not satisfied? In this case, **nonparametric statistics**, which do not require sample data follow a specified distribution, may be used.

DEFINITION

Bootstrapping is a computer-intensive approach to statistical inference whereby parameters are estimated by treating a set of sample data as a population. A computer is used to resample with replacement *n* observations from the sample data. The process is repeated many times. For each resample, the statistic (such as the sample mean) is obtained.

The bootstrap method was invented by Bradley Efron, from Stanford University, in 1979.

Bootstrapping is a method in which sample data may be used to build a sampling distribution of a sample statistic without having access to the entire population. To use bootstrapping, however, two basic requirements must be satisfied.

1. The "center" of the bootstrap distribution must be close to the "center" of the original sample data. For example, the mean of all bootstrap means must be close to the mean of the original data.
2. The distribution of the bootstrap sample statistic must be symmetric.

If using the bootstrap method to obtain a confidence interval for a population mean, we can use the distribution of the sample means obtained from the *B* resamples as an approximation for the actual sampling distribution. A $(1-\alpha)\cdot 100\%$ confidence interval may be obtained using the **percentile method**. With this method, determine the $\frac{\alpha}{2}\cdot 100\text{th}$ percentile (the lower bound) and the $\left(1-\frac{\alpha}{2}\right)\cdot 100\text{th}$ percentile (the upper bound) of the distribution. For example, the lower bound of a 95% confidence interval would be the 2.5th percentile and the upper bound would be the 97.5th percentile.

The Bootstrap Algorithm

1. Select *B* independent bootstrap samples of size *n* with replacement. Note that *n* is the number of observations in the original sample. So, if the original sample has 10 observations, each bootstrap sample will also have 10 observations. Sampling with replacement means that once an observation is selected to be in the sample, its value is "put back into the hat" so it may be sampled again.
2. Determine the value of the statistic of interest, such as the mean, for each of the *B* samples.
3. Use the distribution of the *B* statistics to make a judgement about the value of the parameter. For example, find the 2.5^{th} and 97.5^{th} percentiles to determine the lower and upper bound of a 95% confidence interval.

EXAMPLE 1 Using the Bootstrap Method to Construct a 95% Confidence Interval

The following data represent the pitch speed, in miles per hour, of a random sample of 12 sliders thrown by Jake Arrietta, pitcher for the Chicago Cubs. Construct a 95% confidence interval for the mean pitch speed, in miles per hour, of a Jake Arrietta slider.

88.09	89.01	90.45	89.19	89.74	86.71
88.58	88.82	86.92	87.44	89.07	89.00

Source: BrooksBaseball.net

Bootstrap Confidence Intervals using Simulation

1. Enter the raw data into column var1. Name the column.
2. Select the **Data** menu, highlight **Sample**. Select the variable from column var1. In the "Sample Size:" cell, enter the number of observations, *n.* In the "Number of samples:" cell, enter the number of resamples. Check the "Sample with replacement" box. Check the radio button "Stacked with a sample id". Click Compute!.
3. Select the **Stat** menu, highlight **Summary Stats**, and select **Columns**. Select the Sample(var1) under "Select Columns:" For example, if the variable name in column 1 is MPG, select Sample(MPG). In the dropdown menu "Group by:" select "Sample". Select only the Mean (or statistic of interest) in the Statistics menu. Check the box "Store output in data table". Click Compute!. A popup window will appear that says, "Whoa! Lots of numeric. . ." Click Cancel.
4. Select the **Stat** menu, highlight **Summary Stats**, and select **Columns**. Select the column "Mean" (or statistic of interest). Highlight "Mean" (or statistic of interest) in the "Statistics:" cell. Enter the percentiles corresponding to the lower and upper bound in the "Percentiles" cell. For example, enter 2.5, 97.5 for a 95% confidence interval. Click Compute!.

What was the 95% confidence interval for an Arrietta slider using Student's *t*-distribution?

The approach used in Example 1 illustrates the bootstrap method and logic, but many statistical spreadsheets (such as StatCrunch) have built-in algorithms that do bootstrapping.

EXAMPLE 2 Using StatCrunch's Resample Command and Bootstrap Applet to Obtain a Bootstrap Confidence Interval

Use StatCrunch to estimate a 95% confidence interval for the mean speed of an Arrietta slider using StatCrunch's Resample Command and Bootstrap applet.

Bootstrap Confidence Intervals Using the Resample Command

Select the **Stat** menu, highlight **Resample**, and select **Statistic**. Under "Columns to resample:" select the variable you wish to estimate. Enter the statistic you wish to estimate in the "Statistic:" cell. For example, enter "mean(variable name)" for the mean, "median(variable name)" for the median. Select the "Bootstrap" radio button. Select the "Univariate" radio button. Enter *B*, the number of resamples, in the "Number of resamples:" cell. Enter the percentiles corresponding to the confidence interval you wish to determine. Check any boxes you wish. Click Compute!.

Bootstrap Confidence Intervals Using the Bootstrapping Applet

1. Select the **Applets** menu, highlight **Resampling**, and select **Bootstrap a statistic.** Select the "From data table:" radio button. In the pull-down menu "Samples in:", choose the column that contains the original sample data. In the pull-down menu "Statistic:", select the statistic of interest. Click Compute!.

Click 1000 times for as many bootstrap samples as you want (at least 1000). Determine the lower and upper bounds of the confidence interval from the percentiles.

Bootstrapping to Estimate a Population Proportion

To construct a bootstrap confidence interval for a proportion, we need raw data using 0 for a failure and 1 for a success. The sample proportion is then estimated using the mean of the 0s and 1s.

Some Final Thoughts

Bootstrapping appears to be a powerful method for constructing confidence intervals without the need for underlying model requirements (such as the normality condition for constructing confidence intervals for the mean when sample sizes are small). However, there are some items to be aware of when constructing confidence intervals using the bootstrap approach.

1. If you have small samples, it is possible that the bootstrap distribution will not accurately reflect the true sampling distribution. This could happen if the sample data has less variability than the underlying population.
2. The standard error of bootstrap distributions tends to be narrower by a factor of $\sqrt{\frac{n-1}{n}}$ than those obtained using $\frac{s}{\sqrt{n}}$.
3. Although Efron originally stated that 1000 resamples is enough, it is better to have 10,000 resamples.