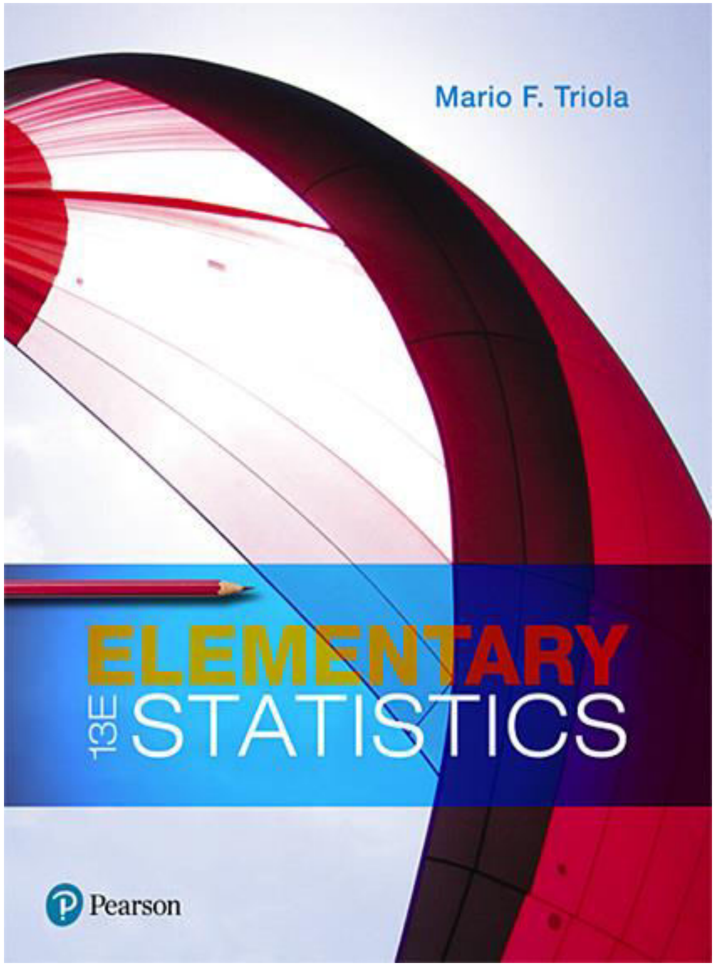
Elementary Statistics

Thirteenth Edition



Chapter 7

Estimating Parameters and Determining Sample Sizes

Estimating Parameters and Determining Sample Sizes

7−1 Estimating a Population Proportion

7−2 Estimating a Population Mean

7−3 Estimating a Population Standard Deviation or Variance

7−4 Bootstrapping: Using Technology for Estimates

Key Concept

This section presents methods for using a sample proportion to make an inference about the value of the corresponding population proportion. Here are the three main concepts included in this section:

- **Point Estimate:** The sample proportion ($\hat{p}$) is the best **point estimate** of the population proportion p.
- **Confidence Interval:** We can use a sample proportion to construct a **confidence interval** estimate of the true value of a population proportion.
- **Sample Size:** We should know how to find the sample size necessary to estimate a population proportion.

Point Estimate (1 of 2)

- Point Estimate
  - A **point estimate** is a single value used to estimate a population parameter. The sample proportion $\hat{p}$ is the best **point estimate** of the population proportion p.

Point Estimate (2 of 2)

Unbiased Estimator

We use $\hat{p}$ as the point estimate of p because it is unbiased and it is the most consistent of the estimators that could be used.

An unbiased estimator is a statistic that targets the value of the corresponding population parameter in the sense that the sampling distribution of the statistic has a mean that is equal to the corresponding population parameter.

The statistic $\hat{p}$ targets the population proportion p.

Example: Facebook (1 of 2)

A Gallup poll was taken in which 1487 adults were surveyed and 43% of them said that they have a Facebook page. Based on that result, find the best point estimate of the proportion of **all** adults who have a Facebook page.

Example: Facebook (2 of 2)

Solution

Because the sample proportion is the best point estimate of the population proportion, we conclude that the best point estimate of p is 0.43. (If using the sample results to estimate the **percentage** of all adults who have a Facebook page, the best point estimate is 43%.)

Confidence Interval

- Confidence Interval
  - A **confidence interval** (or **interval estimate**) is a range (or an interval) of values used to estimate the true value of a population parameter. A confidence interval is sometimes abbreviated as CI.

Confidence Level

- Confidence Level
  - The **confidence level** is the probability $1 - \alpha$ (such as 0.95, or 95%) that the confidence interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times. (The confidence level is also called the **degree of confidence**, or the **confidence coefficient**.)

Relationship Between Confidence Level and α

The following table shows the relationship between the confidence level and the corresponding value of α. The confidence level of 95% is the value used most often.

Most Common Confidence Levels	**Corresponding Values of α**
90% (or 0.90) confidence level:	$\alpha = 0.10$
95% (or 0.95) confidence level:	$\alpha = 0.05$
99% (or 0.99) confidence level:	$\alpha = 0.01$

Pearson

Interpreting a Confidence Interval (1 of 3)

We must be careful to interpret confidence intervals correctly. There is a correct interpretation and many different and creative incorrect interpretations of the confidence interval $0.405 < p < 0.455$.

Correct: “We are 95% confident that the interval from 0.405 to 0.455 actually does contain the true value of the population proportion p.”

Interpreting a Confidence Interval (2 of 3)

We must be careful to interpret confidence intervals correctly. There is a correct interpretation and many different and creative incorrect interpretations of the confidence interval $0.405 < p < 0.455$.

Wrong: "There is a 95% chance that the true value of p will fall between 0.405 and 0.455."

Interpreting a Confidence Interval (3 of 3)

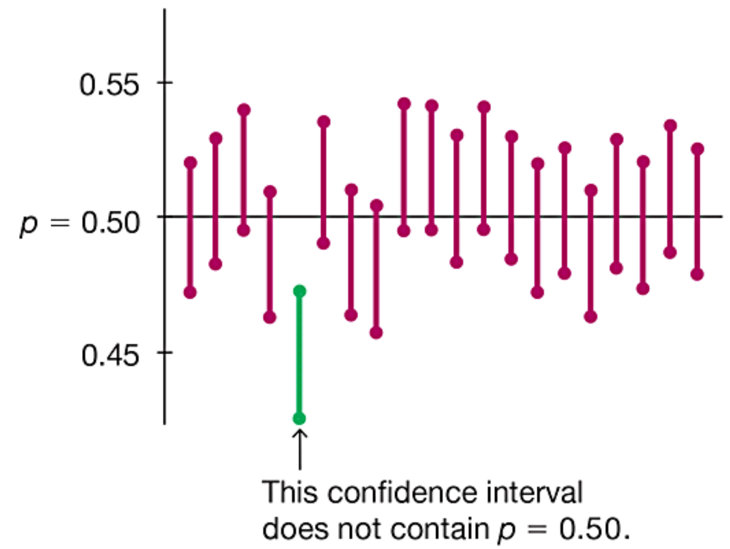
We must be careful to interpret confidence intervals correctly. There is a correct interpretation and many different and creative incorrect interpretations of the confidence interval $0.405 < p < 0.455$.

Wrong: "95% of sample proportions will fall between 0.405 and 0.455."

The Process Success Rate

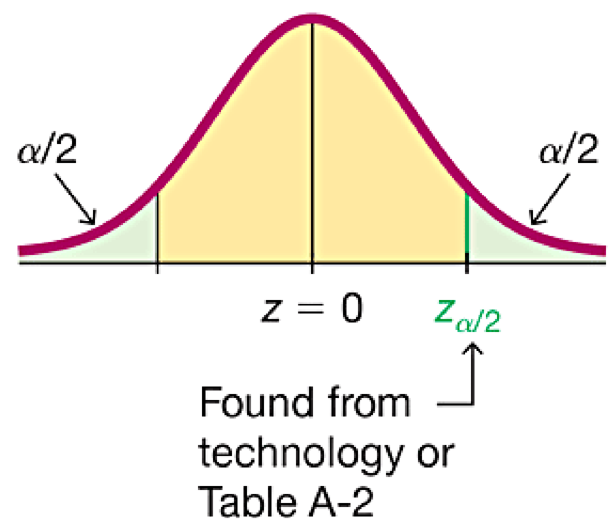
A confidence level of 95% tells us that the **process** we are using should, in the long run, result in confidence interval limits that contain the true population proportion 95% of the time.

Confidence Interval from 20 Different Samples



Critical Values

- Critical Values
  - A **critical value** is the number on the borderline separating sample statistics that are significantly high or low from those that are not significant. The number $z_{\frac{\alpha}{2}}$ is a critical value that is a z score with the property that it is at the border that separates an area of $\frac{\alpha}{2}$ in the right tail of the standard normal distribution.



Example: Finding a Critical Value (1 of 3)

Find the critical value $z_{\frac{\alpha}{2}}$ corresponding to a 95% confidence level.

Example: Finding a Critical Value (2 of 3)

Solution

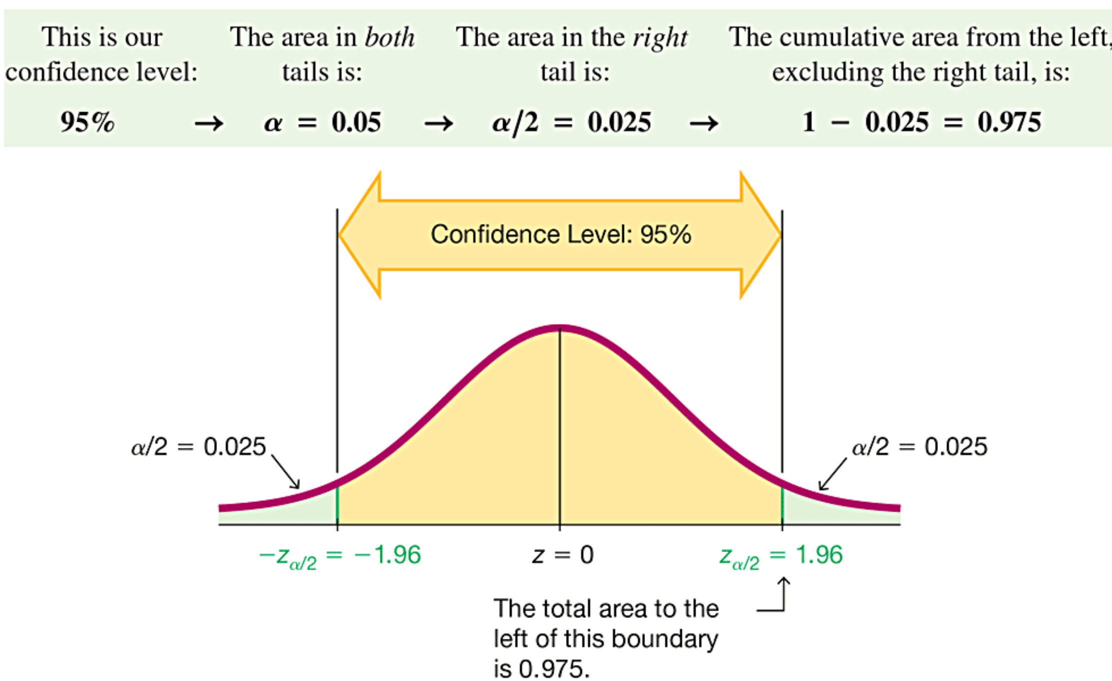
A 95% confidence level corresponds to $\alpha = 0.05$, so $\dfrac{\alpha}{2} = 0.025$.

The figure on the next slide shows that the area in each of the green-shaded tails is $\dfrac{\alpha}{2} = 0.025$. We find $z_{\frac{\alpha}{2}} = 1.96$ by noting that the cumulative area to its left must be 1 − 0.025, or 0.975.

We can use technology or refer to Table A-2 to find that the cumulative left area of 0.9750 corresponds to $z = 1.96$. For a 95% confidence level, the critical value is therefore $z_{\frac{\alpha}{2}} = 1.96$.

Example: Finding a Critical Value (3 of 3)

Note that when finding the critical z score for a 95% confidence level, we use a cumulative left area of 0.9750 (**not** 0.95). Think of it this way:



Common Critical Values

The previous example showed that a 95% confidence level results in a critical value of $z_{\frac{\alpha}{2}} = 1.96$.

This is the most common critical value, and it is listed with two other common values in the table that follows.

Confidence level	α	Critical Value, $z_{\frac{\alpha}{2}}$
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.575

Margin of Error

- Margin of Error

  When data from a simple random sample are used to estimate a population proportion *p*, the difference between the sample proportion $\hat{p}$ and the population proportion *p* is an error. The maximum likely amount of that error is the **margin of error,** denoted by ***E***. There is a probability of $1 - \alpha$ (such as 0.95) that the difference between $\hat{p}$ and *p* is *E* or less. The margin of error *E* is

  The margin of error *E* is also called the **maximum error of the estimate** and can be found by multiplying the critical value and the estimated standard deviation of sample proportions.

Margin of Error for Proportions

Formula

$$E = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$ margin of error for proportions

Critical value

Estimated standard deviation of sample proportions

Pearson

Confidence Interval for Estimating a Population Proportion *p*: Objective

Construct a confidence interval used to estimate a population proportion p.

Confidence Interval for Estimating a Population Proportion *p*: Notation

p = **population** proportion

$\hat{p}$ = **sample** proportion

n = number of sample values

E = margin of error

$z_{\frac{\alpha}{2}}$ = critical value: the z score separating an area of $\frac{\alpha}{2}$ in the right tail of the standard normal distribution

Confidence Interval for Estimating a Population Proportion *p*: Requirements

1. The sample is a simple random sample.
2. The conditions for the binomial distribution are satisfied: There is a fixed number of trials, the trials are independent, there are two categories of outcomes, and the probabilities remain constant for each trial.
3. There are at least 5 successes and at least 5 failures.

Confidence Interval for Estimating a Population Proportion *p*: Confidence Interval Estimate of *p*

$$\hat{p} - E < p < \hat{p} + E \quad \text{where} \quad E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

The confidence interval is often expressed in the following formats:

$$\hat{p} \pm E \quad \text{or} \quad \left(\hat{p} - E,\ \hat{p} + E\right)$$

Confidence Interval for Estimating a Population Proportion p: Round−Off Rule for Confidence Interval Estimates of *p*

Round the confidence interval limits for *p* to three significant digits.

Procedure for Constructing a Confidence Interval for *p* (1 of 2)

1. Verify that the requirements in the preceding slides are satisfied.
2. Use technology or Table A-2 to find the critical value $z_{\frac{\alpha}{2}}$ that corresponds to the desired confidence level.
3. Evaluate the margin of error $E = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$.

Procedure for Constructing a Confidence Interval for *p* (2 of 2)

4. Using the value of the calculated margin of error E and the value of the sample proportion $\hat{p}$, find the values of the **confidence interval limits** $\hat{p} - E$ and $\hat{p} + E$. Substitute those values in the general format for the confidence interval.
5. Round the resulting confidence interval limits to three significant digits.

Example: Constructing a Confidence Interval: Poll Results (1 of 8)

We noted that a Gallup poll of 1487 adults showed that 43% of the respondents have Facebook pages.

The sample results are $n = 1487$ and $\hat{p} = 0.43$.

a. Find the margin of error *E* that corresponds to a 95% confidence level.

b. Find the 95% confidence interval estimate of the population proportion *p*.

c. Based on the results, can we safely conclude that fewer than 50% of adults have Facebook pages? Assuming that you are a newspaper reporter, write a brief statement that accurately describes the results and includes all of the relevant information.

Example: Constructing a Confidence Interval: Poll Results (2 of 8)

Requirement Check

(1) The polling methods used by the Gallup organization result in samples that can be considered to be simple random samples.

(2) The conditions for a binomial experiment are satisfied because there is a fixed number of trials (1487), the trials are independent (because the response from one person doesn’t affect the probability of the response from another person), there are two categories of outcome (subject has a Facebook page or does not), and the probability remains constant, because *P*(having a Facebook page) is fixed for a given point in time.

(3) With 43% of the respondents having Facebook pages, the number with Facebook pages is 639 (or 43% of 1487). If 639 of the 1487 subjects have Facebook pages, the other 848 do not, so the number of successes (639) and the number of failures (848) are both at least 5.

The check of requirements has been successfully completed.

Example: Constructing a Confidence Interval: Poll Results (3 of 8)

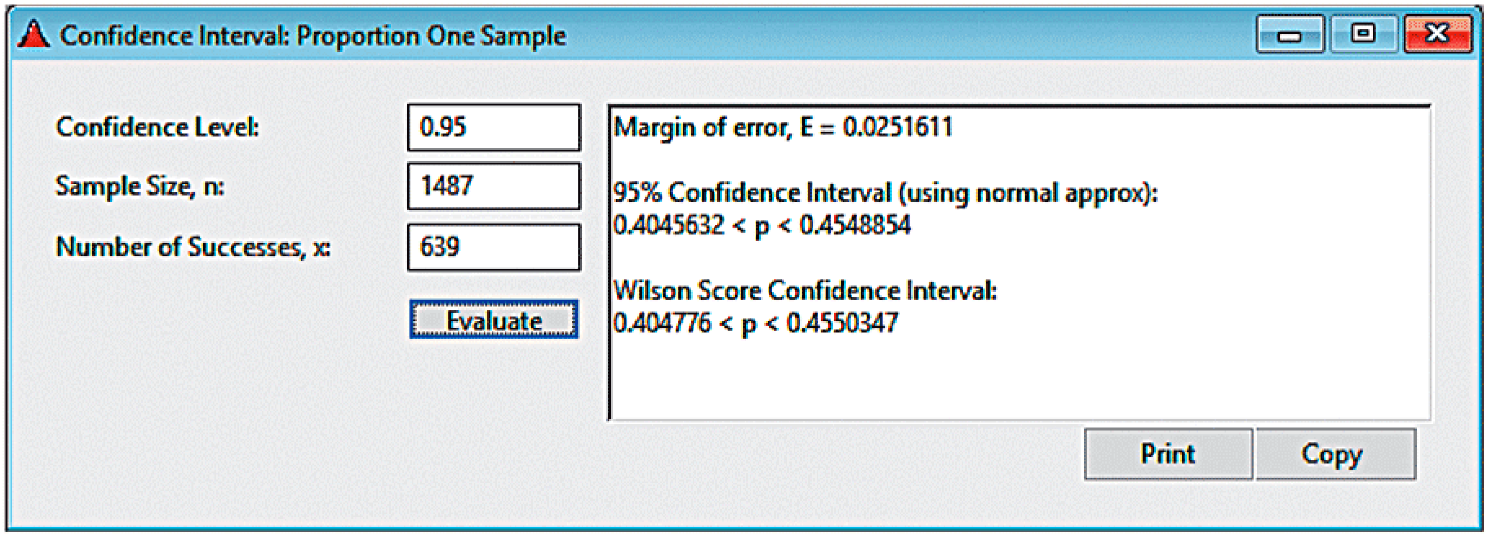
Technology

The confidence interval and margin of error can be easily found using technology. From the Statdisk display on the next slide, we can see the required entries on the left and the results displayed on the right. Like most technologies, Statdisk requires a value for the number of successes, so we simply find 43% of 1487 and round the result of 639.41 to the whole number 639. The results show that the margin of error is $E = 0.025$ (rounded) and the confidence interval is $0.405 < p < 0.455$ (rounded).

Example: Constructing a Confidence Interval: Poll Results (4 of 8)

Technology

Statdisk



Example: Constructing a Confidence Interval: Poll Results (5 of 8)

Solution (Manual Calculation)

a. The margin of error is found by using Formula 7-1 with $z_{\frac{\alpha}{2}} = 1.96$, $\hat{p} = 0.43$, $\hat{q} = 0.57$, and $n = 1487$.

$$E = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96\sqrt{\frac{(0.43)(0.57)}{1487}} = 0.0251636$$

Example: Constructing a Confidence Interval: Poll Results (6 of 8)

Solution

b. Constructing the confidence interval is really easy now that we know that $\hat{p} = 0.43$ and $E = 0.0251636$. Simply substitute those values to obtain this result:

$$\hat{p} - E < p < \hat{p} + E$$

$$0.43 - 0.0251636 < p < 0.43 + 0.0251636$$

$$0.405 < p < 0.455 \text{ (rounded)}$$

Example: Constructing a Confidence Interval: Poll Results (7 of 8)

Solution

c. Based on the confidence interval obtained in part (b), it does appear that fewer than 50% of adults have a Facebook page because the interval of values from 0.405 to 0.455 is an interval that is completely below 0.50.

Example: Constructing a Confidence Interval: Poll Results (8 of 8)

Summary of results

Here is one statement that summarizes the results: 43% of adults have Facebook pages. That percentage is based on a Gallup poll of 1487 randomly selected adults in the United States. In theory, in 95% of such polls, the percentage should differ by no more than 2.5 percentage points in either direction from the percentage that would be found by interviewing all adults.

Analyzing Polls

When analyzing results from polls, consider the following:

1. The sample should be a simple random sample, not an inappropriate sample.
2. The confidence level should be provided.
3. The sample size should be provided.
4. Except for relatively rare cases, the quality of the poll results depends on the sampling method and the size of the sample, but the size of the **population** is usually not a factor.

Caution

Never think that poll results are unreliable if the **sample size** is a small percentage of the **population size.** The population size is usually not a factor in determining the reliability of a poll.

Finding the Point Estimate and *E* from a Confidence Interval

Point estimate of *p*:

$$\hat{p} = \frac{(\text{upper confidence interval limit}) + (\text{lower confidence interval limit})}{2}$$

Margin of error:

$$E = \frac{(\text{upper confidence interval limit}) - (\text{lower confidence interval limit})}{2}$$

Example: Finding a Sample Proportion and Margin of Error (1 of 3)

The article “High−Dose Nicotine Patch Therapy,” by Dale, Hurt, et al. (**Journal of the American Medical Association,** Vol. 274, No. 17) includes this statement: “Of the 71 subjects, 70% were abstinent from smoking at 8 weeks (95% confidence interval [CI], 58% to 81%).” Use that statement to find the point estimate $\hat{p}$ and the margin of error E.

Example: Finding a Sample Proportion and Margin of Error (2 of 3)

Solution

We get the 95% confidence interval of $0.58 < p < 0.81$ from the given statement of “58% to 81%.”

The point estimate $\hat{p}$ is the value midway between the upper and lower confidence interval limits, so we get

$$\hat{p} = \frac{(\text{upper confidence interval limit}) + (\text{lower confidence interval limit})}{2}$$

$$= \frac{0.81 + 0.58}{2} = 0.695$$

Example: Finding a Sample Proportion and Margin of Error (3 of 3)

Solution

The margin of error can be found as follows:

$$E = \frac{(\text{upper confidence interval limit}) - (\text{lower confidence interval limit})}{2}$$

$$= \frac{0.81 - 0.58}{2} = 0.115$$

Determining Sample Size: Finding the Sample Size Required to Estimate a Population Proportion: Objective

Finding the Sample Size Required to Estimate a Population Proportion

Determining Sample Size: Finding the Sample Size Required to Estimate a Population Proportion: Notation

p = **population** proportion

$\hat{p}$ = **sample** proportion

n = number of sample values

E = margin of error

$z_{\frac{\alpha}{2}}$ = critical value: the z score separating an area of $\frac{\alpha}{2}$ in the right tail of the standard normal distribution

Determining Sample Size: Finding the Sample Size Required to Estimate a Population Proportion: Requirements

The sample must be a simple random sample of independent sample units.

When an estimate $\hat{p}$ in known:

$$n = \frac{\left[z_{\frac{\alpha}{2}}\right]^2 \hat{p}\hat{q}}{E^2}$$

When no estimate $\hat{p}$ is known:

$$n = \frac{\left[z_{\frac{\alpha}{2}}\right]^2 0.25}{E^2}$$

Determining Sample Size: Finding the Sample Size Required to Estimate a Population Proportion: Round−Off Rule for Determining Sample Size

If the computed sample size *n* is not a whole number, round the value of *n* up to the next **larger** whole number, so the sample size is sufficient instead of being slightly insufficient. For example, round 1067.11 to 1068.

Example: What Percentage of Adults Make Online Purchases? (1 of 4)

When the author was conducting research for this chapter, he could find no information about the percentage of adults who make online purchases, yet that information is extremely important to online stores as well as brick and mortar stores. If the author were to conduct his own survey, how many adults must be surveyed in order to be 95% confident that the sample percentage is in error by no more than three percentage points?

a. Assume that a recent poll showed that 80% of adults make online purchases.

b. Assume that we have no prior information suggesting a possible value of the population proportion.

Example: What Percentage of Adults Make Online Purchases? (2 of 4)

Solution

a. With a 95% confidence level, we have $\alpha = 0.05$, so $z_{\frac{\alpha}{2}} = 1.96$. Also, the margin of error is $E = 0.03$. The prior survey suggests that $\hat{p} = 0.80$, so $\hat{q} = 0.20$ (found from $\hat{q} = 1 - 0.80$). Because we have an estimated value of $\hat{p}$, we use:

$$n = \frac{\left[z_{\frac{\alpha}{2}}\right]^2 \hat{p}\hat{q}}{E^2} = \frac{[1.96]^2 (0.80)(0.20)}{0.03^2}$$

$$= 682.951 = 683 \text{ (rounded)}$$

We must obtain a simple random sample that includes at least 683 adults.

Example: What Percentage of Adults Make Online Purchases? (3 of 4)

Solution

b. With no prior knowledge of *pn* (or *qn*), we use Formula 7−3 as follows:

$$n = \frac{\left[z_{\frac{\alpha}{2}}\right]^2 \cdot 0.25}{E^2} = \frac{[1.96]^2 \cdot 0.25}{0.03^2}$$

$$= 1067.11 = 1068 \text{ (rounded up)}$$

We must obtain a simple random sample that includes at least 1068 adults.

Example: What Percentage of Adults Make Online Purchases? (4 of 4)

Interpretation

To be 95% confident that our sample percentage is within three percentage points of the true percentage for all adults, we should obtain a simple random sample of 1068 adults, assuming no prior knowledge. By comparing this result to the sample size of 683 found in part (a), we can see that if we have no knowledge of a prior study, a larger sample is required to achieve the same results as when the value of $\hat{p}$ can be estimated.

Better Performing Confidence Intervals (1 of 3)

Plus Four Method

The **plus four confidence interval** performs better than the Wald confidence interval in the sense that its coverage probability is closer to the confidence level that is used.

The plus four confidence interval uses this very simple procedure: Add 2 to the number of successes x, add 2 to the number of failures (so that the number of trials n is increased by 4), and then find the Wald confidence interval as described in Part 1 of this section.

Better Performing Confidence Intervals (2 of 3)

Wilson Score

The Wilson score confidence interval performs better than the Wald CI in the sense that the coverage probability is closer to the confidence level. With a confidence level of 95%, the Wilson score confidence interval would get us closer to a 0.95 probability of containing the parameter *p*. However, given its complexity, it is easy to see why this superior Wilson score confidence interval is not used much in introductory statistics courses.

$$\frac{\hat{p}+\dfrac{z_{\frac{\alpha}{2}}^{2}}{2n}\pm z_{\frac{\alpha}{2}}\sqrt{\dfrac{\hat{p}\hat{q}+\dfrac{z_{\frac{\alpha}{2}}^{2}}{4n}}{n}}}{1+\dfrac{z_{\frac{\alpha}{2}}^{2}}{n}}$$

Better Performing Confidence Intervals (3 of 3)

Clopper−Pearson Method

The Clopper−Pearson method is an “exact” method in the sense that it is based on the exact binomial distribution instead of an approximation of a distribution. It is criticized for being **too conservative** in this sense: When we select a specific confidence level, the coverage probability is usually greater than or equal to the selected confidence level. Select a confidence level of 0.95, and the actual coverage probability is usually 0.95 or greater, so that 95% or more of such confidence intervals will contain *p*. Calculations with this method are too messy to consider here.

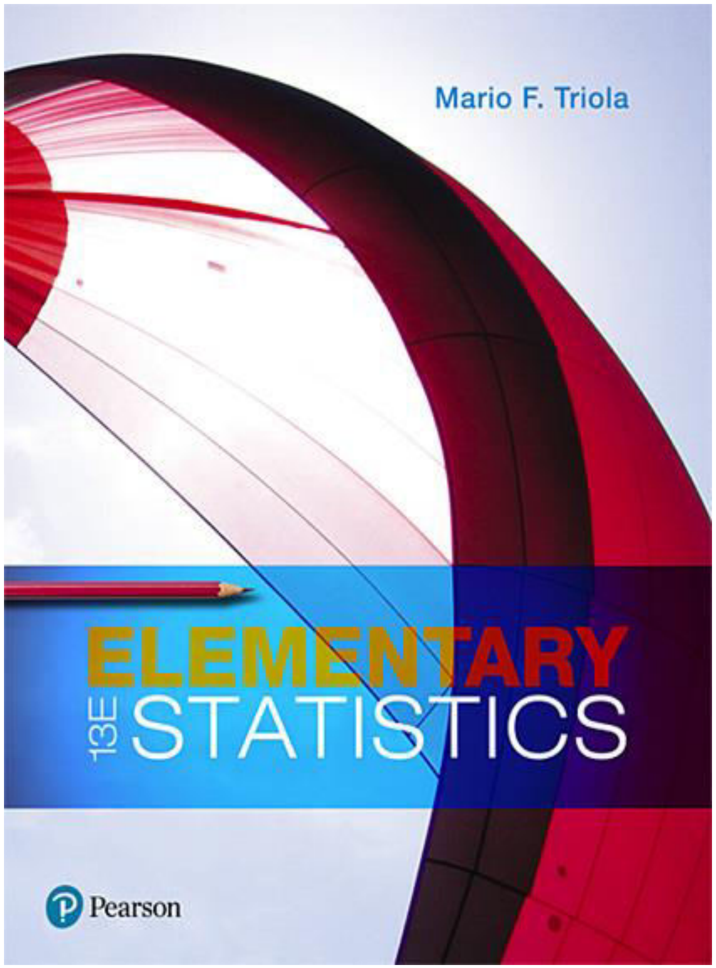
Which Method is Best?

There are other methods for constructing confidence intervals that are not discussed here. There isn't universal agreement on which method is best for constructing a confidence interval estimate of *p.*

- The Wald confidence interval is best as a teaching tool for introducing students to confidence intervals.
- The plus four confidence interval is almost as easy as Wald and it performs better than Wald by having a coverage probability closer to the selected confidence level.

Elementary Statistics

Thirteenth Edition



Chapter 7

Estimating Parameters and Determining Sample Sizes

Estimating Parameters and Determining Sample Sizes

7-1 Estimating a Population Proportion

7-2 Estimating a Population Mean

7-3 Estimating a Population Standard Deviation or Variance

7-4 Bootstrapping: Using Technology for Estimates

Key Concept (1 of 2)

The main goal of this section is to present methods for using a sample mean $\bar{x}$ to make an inference about the value of the corresponding population mean μ.

Key Concept (2 of 2)

There are three main concepts included in this section:

- **Point Estimate:** The sample mean $\bar{x}$ is the best **point estimate** (or single value estimate) of the population mean μ.
- **Confidence Interval:** Use sample data to construct and interpret a **confidence interval** estimate of the true value of a population mean μ.
- **Sample Size:** Find the sample size necessary to estimate a population mean.

Confidence Interval for Estimating a Population Mean with σ Not Known: Objective

Construct a confidence interval used to estimate a population mean.

Confidence Interval for Estimating a Population Mean with σ Not Known: Notation

μ = population mean

n = number of sample values

$\bar{x}$ = sample mean

E = margin of error

s = sample standard deviation

Confidence Interval for Estimating a Population Mean with σ Not Known: Requirements

1. The sample is a simple random sample.
2. Either or both of these conditions are satisfied: The population is normally distributed or $n > 30$.

Confidence Interval for Estimating a Population Mean with σ Not Known: Confidence Interval (1 of 2)

Margin of Error:

$$E = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \quad (\textbf{Use df} = n - 1)$$

Confidence Interval: The confidence interval is associated with a confidence level, such as 0.95 (or 95%), and α is the complement of the confidence level. For a 0.95 (or 95%) confidence level, $\alpha = 0.05$.

Confidence Interval for Estimating a Population Mean with σ Not Known: Confidence Interval (2 of 2)

Critical Value: $t_{\frac{\alpha}{2}}$ is the critical t value separating an area of $\frac{\alpha}{2}$ in the right tail of the Student t distribution.

Degrees of Freedom: df = $n - 1$ is the number of degrees of freedom used when finding the critical value.

Confidence Interval for Estimating a Population Mean with σ Not Known: Round-Off Rule

1. **Original Data:** When using an **original set of data** values, round the confidence interval limits to one more decimal place than is used for the original set of data.

2. **Summary Statistics:** When using the **summary statistics** of n, x, and s, round the confidence interval limits to the same number of decimal places used for the sample mean.

Key Points about the Student *t* Distribution (1 of 5)

- **Student *t* Distribution** If a population has a normal distribution, then the distribution of

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

is a **Student *t* distribution** for all samples of size *n*. A Student *t* distribution is commonly referred to as a ***t* distribution.**

Key Points about the Student *t* Distribution (2 of 5)

- **Degrees of Freedom** Finding a critical value $t_{\frac{\alpha}{2}}$ requires a value for the **degrees of freedom** (or **df**).

  In general, the number of degrees of freedom for a collection of sample data is the number of sample values that can vary after certain restrictions have been imposed on all data values. For the methods of this section, the number of degrees of freedom is the sample size minus 1.

$$\textbf{Degrees of freedom} = n - 1$$

Key Points about the Student t Distribution (3 of 5)

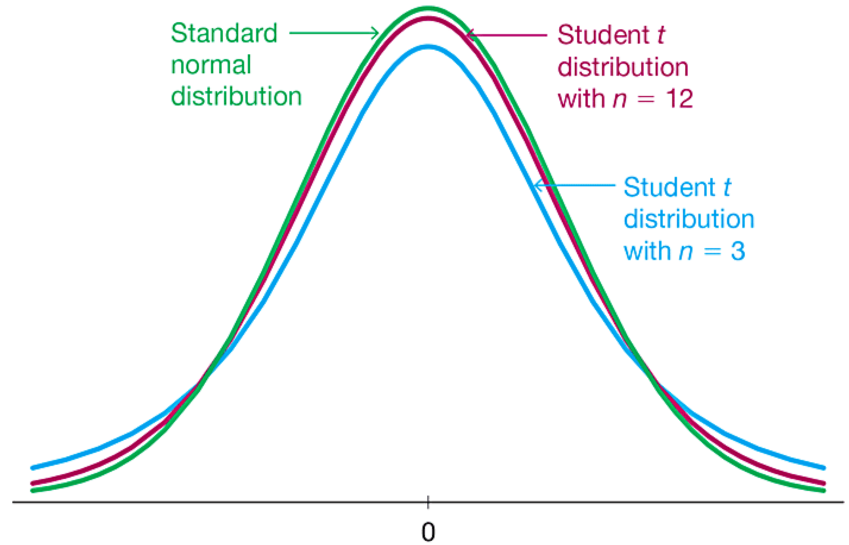
Finding Critical Value $t_{\frac{\alpha}{2}}$

A critical value $t_{\frac{\alpha}{2}}$ can be found using technology or Table A-3. Technology can be used with any number of degrees of freedom, but Table A-3 can be used for select numbers of degrees of freedom only. If using Table A-3 to find a critical value of $t_{\frac{\alpha}{2}}$, but the table does not include the exact number of degrees of freedom, you could use the closest value, or you could be conservative by using the next lower number of degrees of freedom found in the table, or you could interpolate.

Key Points about the Student *t* Distribution (4 of 5)

- The Student *t* distribution is different for different sample sizes. See the figure for the cases $n = 3$ and $n = 12$.

The Student *t* distribution has the same general shape and symmetry as the standard normal distribution, but it has the greater variability that is expected with small samples.



Key Points about the Student *t* Distribution (5 of 5)

- The Student *t* distribution has the same general symmetric bell shape as the standard normal distribution, but has more variability (with wider distributions), as we expect with small samples.
- The Student *t* distribution has a mean of $t = 0$ (just as the standard normal distribution has a mean of $z = 0$).
- The standard deviation of the Student *t* distribution varies with the sample size, but it is greater than 1 (unlike the standard normal distribution, which has s = 1).
- As the sample size *n* gets larger, the Student *t* distribution gets closer to the standard normal distribution.

Procedure for Constructing a Confidence Interval for μ (1 of 2)

1. Verify that the two requirements are satisfied: The sample is a simple random sample and the population is normally distributed or $n > 30$.
2. With σ unknown (as is usually the case), use $n - 1$ degrees of freedom and use technology or a t distribution table (such as Table A-3) to find the critical value $t_{\frac{\alpha}{2}}$ that corresponds to the desired confidence level.
3. Evaluate the margin of error using $E = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$.

Procedure for Constructing a Confidence Interval for μ (2 of 2)

4. Using the value of the calculated margin of error E and the sample mean $\bar{x}$, substitute the values in one of the formats for CI:

$$\bar{x} - E < \mu < \bar{x} + E \text{ or } \bar{x} \pm E \text{ or } (\bar{x} - E, \bar{x} + E).$$

5. With an **original set of data** values, round the confidence interval limits to one more decimal place than used for the original set of data, but when using the **summary statistics** of n, $\bar{x}$, and s, round the confidence interval limits to the same number of decimal places used for the sample mean.

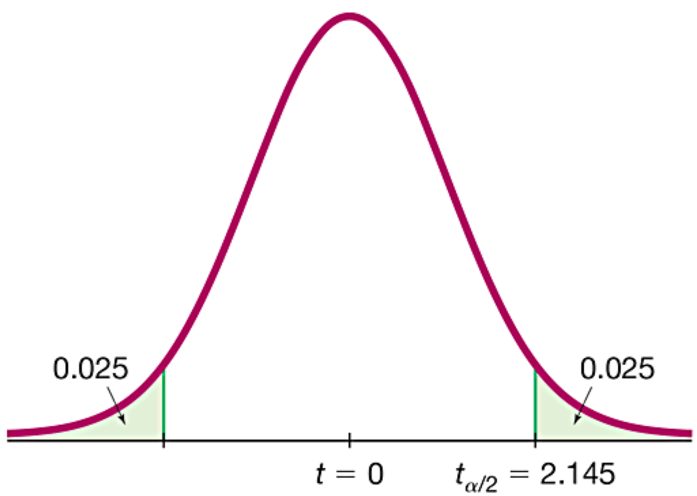
Example: Finding a Critical Value t alpha by 2 (1 of 4)

Find the critical value $t_{\frac{\alpha}{2}}$ corresponding to a 95% confidence level, given that the sample has size $n = 15$.

Example: Finding a Critical Value t alpha by 2 (2 of 4)

Solution

Because n = 15, the number of degrees of freedom is $n - 1 = 14$. The 95% confidence level corresponds to $\sigma = 0.05$, so there is an area of 0.025 in each of the two tails of the t distribution, as shown.

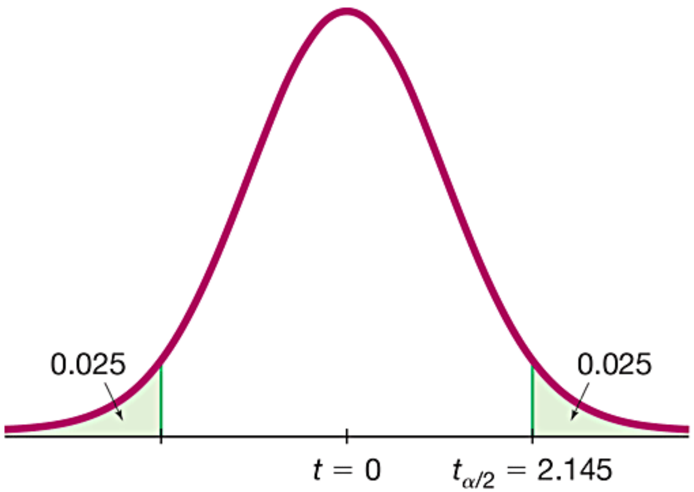


Pearson

Example: Finding a Critical Value t alpha by 2 (3 of 4)

Solution

Using Technology Technology can be used to find that for 14 degrees of freedom and an area of 0.025 in each tail, the critical value is $t_{\frac{\alpha}{2}} = t_{0.025} = 2.145$.



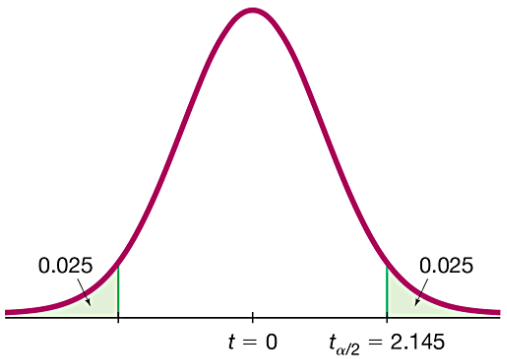
Example: Finding a Critical Value t alpha by 2 (4 of 4)

Solution

Using Table A-3 To find the critical value using Table A-3, use the column with 0.05 for the "Area in Two Tails" (or use the same column with 0.025 for the "Area in One Tail").

The number of degrees of freedom is df = $n - 1 = 14$.

We get $t_{\frac{\alpha}{2}} = t_{0.025} = 2.145$.



Pearson

Finding a Point Estimate and Margin of Error *E* from a Confidence Interval

Point estimate of μ:

$$\bar{x} = \frac{(\text{upper confidence limit}) + (\text{lower confidence limit})}{2}$$

Margin of error:

$$E = \frac{(\text{upper confidence limit}) - (\text{lower confidence limit})}{2}$$

Finding the Sample Size Required to Estimate a Population Mean: Objective

Determine the sample size n required to estimate the value of a population mean μ.

Finding the Sample Size Required to Estimate a Population Mean: Notation

μ = population mean

σ = population standard deviation

$\bar{x}$ = sample mean

E = desired margin of error

$z_{\frac{\alpha}{2}}$ = z score separating an area of $\frac{\alpha}{2}$ in the right tail of the standard normal distribution

Finding the Sample Size Required to Estimate a Population Mean: Requirement

The sample must be a simple random sample.

Finding the Sample Size Required to Estimate a Population Mean: Sample Size

The required sample size is found by

$$n = \left[\frac{z_{\frac{\alpha}{2}}\sigma}{E} \right]^2$$

Finding the Sample Size Required to Estimate a Population Mean: Round-Off Rule

If the computed sample size *n* is not a whole number, round the value of *n* up to the next **larger** whole number.

Dealing with Unknown σ When Finding Sample Size

1. Use the range rule of thumb to estimate the standard deviation as follows: $\sigma \approx \text{range}/4$, where the range is determined from sample data.
2. Start the sample collection process without knowing σ and, using the first several values, calculate the sample standard deviation s and use it in place of σ. The estimated value of σ can then be improved as more sample data are obtained, and the sample size can be refined accordingly.
3. Estimate the value of σ by using the results of some other earlier study.

Example: IQ Scores of Statistics Students (1 of 3)

Assume that we want to estimate the mean IQ score for the population of statistics students. How many statistics students must be randomly selected for IQ tests if we want 95% confidence that the sample mean is within 3 IQ points of the population mean?

Example: IQ Scores of Statistics Students (2 of 3)

Solution

For a 95% confidence interval, we have $\alpha = 0.05$, so $z_{\frac{\alpha}{2}} = 1.96$. Because we want the sample mean to be within 3 IQ points of μ, the margin of error is $E = 3$. Also, we can assume that $\sigma = 15$ (see the discussion that immediately precedes this example). We get

$$n = \left[\frac{z_{\frac{\alpha}{2}}\sigma}{E}\right]^2 = \left[\frac{1.96 \cdot 15}{3}\right]^2 = 96.04 = 97. \text{ (rounded } \textbf{up}\text{)}$$

Example: IQ Scores of Statistics Students (3 of 3)

Interpretation

Among the thousands of statistics students, we need to obtain a simple random sample of at least 97 of their IQ scores.

With a simple random sample of only 97 statistics students, we will be 95% confident that the sample mean $\bar{x}$ is within 3 IQ points of the true population mean μ.

Estimating a Population Mean When σ Is Known

If we somehow do know the value of σ, the confidence interval is constructed using the standard normal distribution instead of the Student *t* distribution, so the same procedure can be used with this margin of error:

$$\text{Margin of Error: } E = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \quad \text{(used with known } \sigma\text{).}$$

Example: Confidence Interval Estimate of μ with Known σ (1 of 6)

Use the 15 birth weights of girls given below, for which n = 15 and $\bar{x}$ = 30.9 hg. Construct a 95% confidence interval estimate of the mean birth weight of all girls by assuming that σ is known to be 2.9 hg.

33 28 33 37 31 32 31
28 34 28 33 26 30 31 28

Example: Confidence Interval Estimate of μ with Known σ (2 of 6)

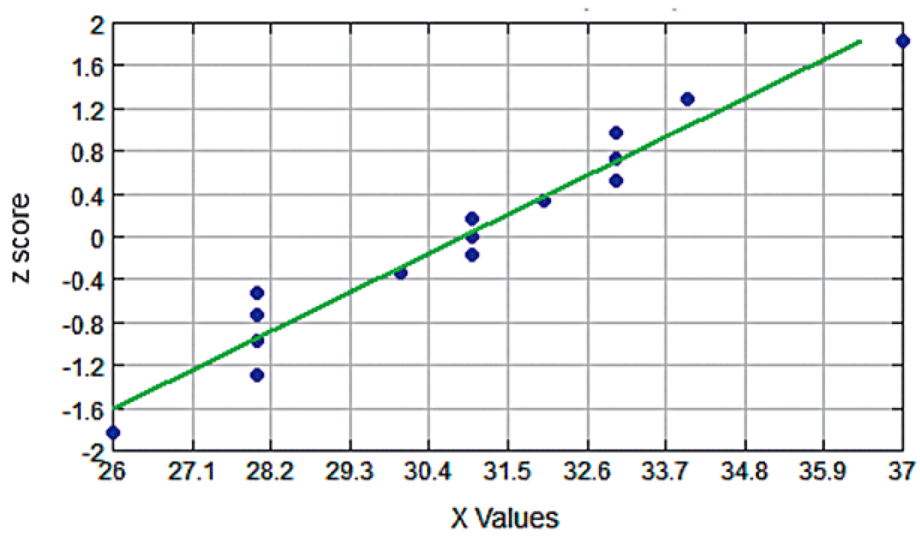
Solution

Requirement Check: (1) The sample is a simple random sample. (2) Because the sample size is n = 15, the requirement that "the population is normally distributed or the sample size is greater than 30" can be satisfied only if the sample data appear to be from a normally distributed population, so we need to investigate normality.

Example: Confidence Interval Estimate of μ with Known σ (3 of 6)

Solution

Sample data appear to be from a normally distributed population.



Example: Confidence Interval Estimate of μ with Known σ (4 of 6)

Solution

With a 95% confidence level, we have $\alpha = 0.5$, and we get $z_{\frac{\alpha}{2}} = 1.96$, $\sigma = 2.9$ hg, and $n = 15$, we find the value of the margin of error E:

$$E = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$= 1.96 \cdot \frac{2.9}{\sqrt{15}} = 1.46760$$

Example: Confidence Interval Estimate of μ with Known σ (5 of 6)

Solution

With $\bar{x}$ = 30.9 hg and E = 1.46760, we find the 95% confidence interval as follows:

$$\bar{x} - E < \mu < \bar{x} + E$$

$$30.9 - 1.46760 < \mu < 30.9 + 1.46760$$

$$29.4 \text{ hg} < \mu < 32.4 \text{ hg}$$

(rounded to one decimal place)

Example: Confidence Interval Estimate of μ with Known σ (6 of 6)

Solution

Remember, this example illustrates the situation in which the population standard deviation σ is known, which is rare. The more realistic situation with σ unknown is considered in Part 1 of this section.

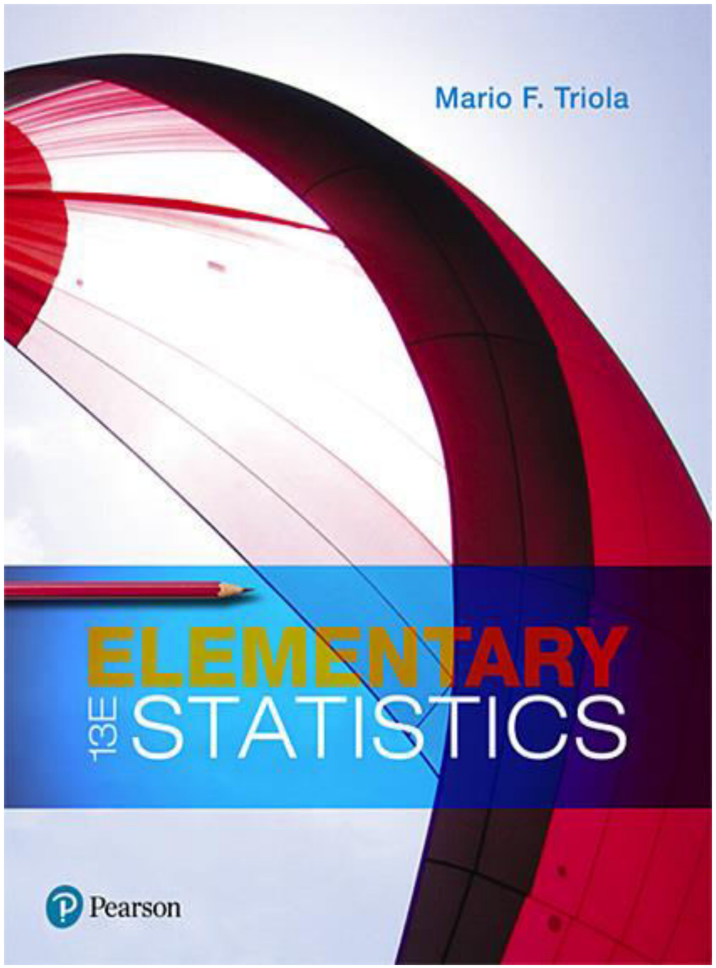
Choosing an Appropriate Distribution

Choosing between Student *t* and *z* (Normal) Distributions

Conditions	Method
σ not known and normally distributed population **or** σ not known and $n > 30$	Use student *t* distribution
σ known and normally distributed population **or** σ known and $n > 30$ (In reality, σ is rarely known.)	Use normal (z) distribution.
Population is not normally distributed and $n \leq 30$.	Use the bootstrapping method (Section 7-4) or a nonparametric method.

Elementary Statistics

Thirteenth Edition



Chapter 7

Estimating Parameters and Determining Sample Sizes

Estimating Parameters and Determining Sample Sizes

7-1 Estimating a Population Proportion

7-2 Estimating a Population Mean

7-3 Estimating a Population Standard Deviation or Variance

7-4 Bootstrapping: Using Technology for Estimates

Key Concept

This section presents methods for using a sample standard deviation s (or a sample variance s^2) to estimate the value of the corresponding population standard deviation σ (or population variance σ^2). Here are the main concepts:

- **Point Estimate:** The sample variance s^2 is the best **point estimate** of the population variance σ^2.
- **Confidence Interval:** When constructing a **confidence interval** estimate of a population standard deviation, we construct the confidence interval using the χ^2 **distribution.**

Chi-Square Distribution (1 of 6)

Key points about the χ^2 (chi-square or chi-squared) distribution:

- In a normally distributed population with variance σ^2, if we randomly select independent samples of size n and, for each sample, compute the sample variance s^2, the sample statistic $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ has a sampling distribution called the **chi-square distribution**.

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

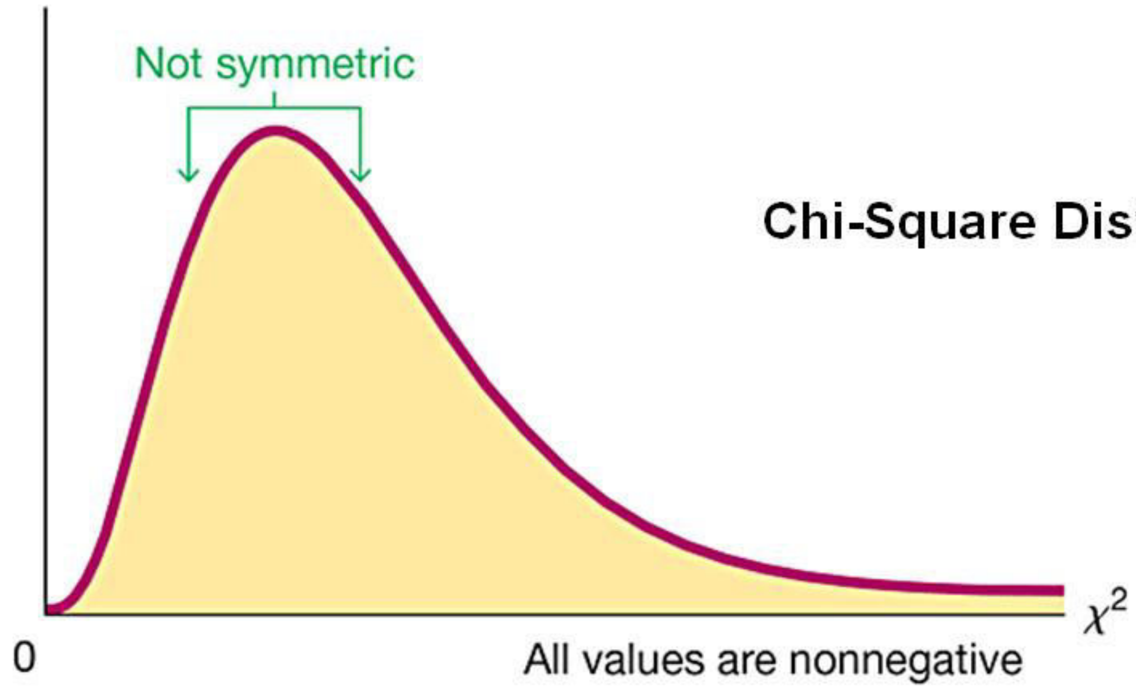
Chi-Square Distribution (2 of 6)

- **Critical Values of χ^2** We denote a right-tailed critical value by χ_R^2 and we denote a left-tailed critical value by χ_L^2. Those critical values can be found by using technology or Table A-4, and they require that we first determine a value for the number of **degrees of freedom**.
- **Degrees of Freedom**
  For the methods of this section, the number of degrees of freedom is the sample size minus 1.

$$\text{Degrees of freedom: df} = n - 1$$

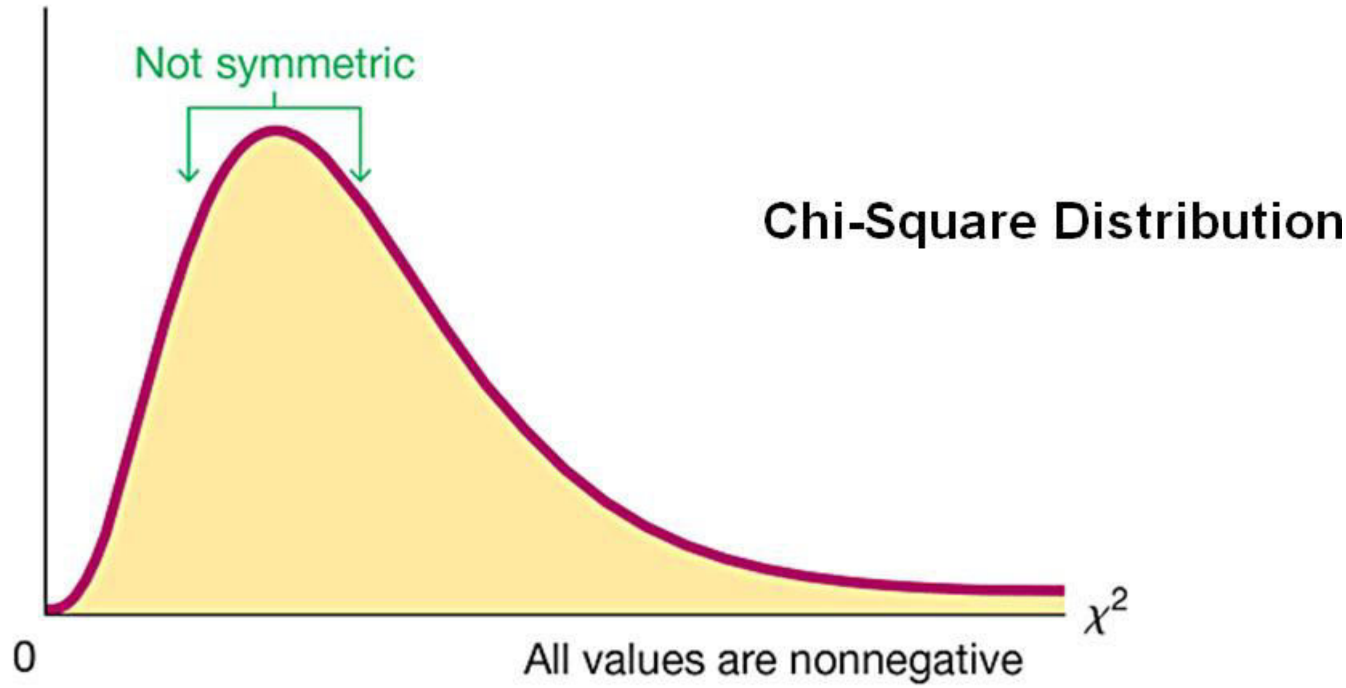
Chi-Square Distribution (3 of 6)

- The chi-square distribution is skewed to the right, unlike the normal and Student *t* distributions.



Chi-Square Distribution (4 of 6)

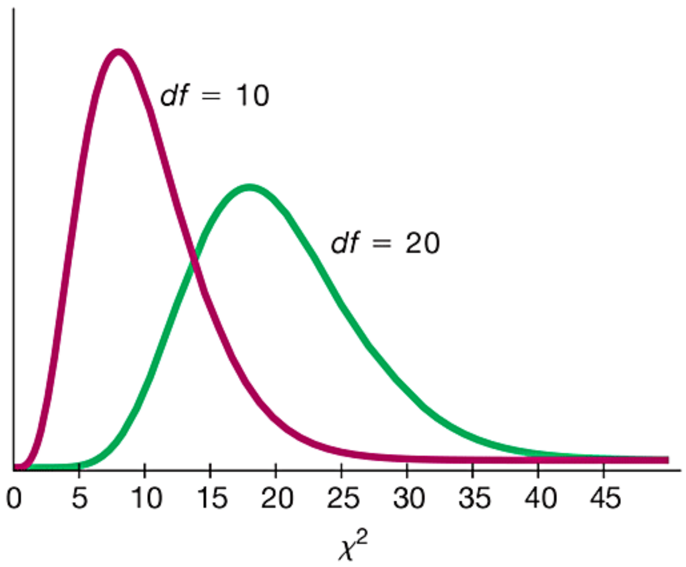
- The values of chi-square can be zero or positive, but they cannot be negative.



Chi-Square Distribution (5 of 6)

- The chi-square distribution is different for each number of degrees of freedom. As the number of degrees of freedom increases, the chi-square distribution approaches a normal distribution.

Chi-Square Distribution for df = 10 and df = 20



Chi-Square Distribution (6 of 6)

Because the chi-square distribution is not symmetric, a confidence interval estimate of σ^2 does not fit a format of $s^2 - E < \sigma^2 < s^2 + E$, so we must do separate calculations for the upper and lower confidence interval limits. If using Table A-4 for finding critical values, note the following design feature of that table:

In Table A-4, each critical value of χ^2 in the body of the table corresponds to an area given in the top row of the table, and each area in that top row is a **cumulative area to the right** of the critical value.

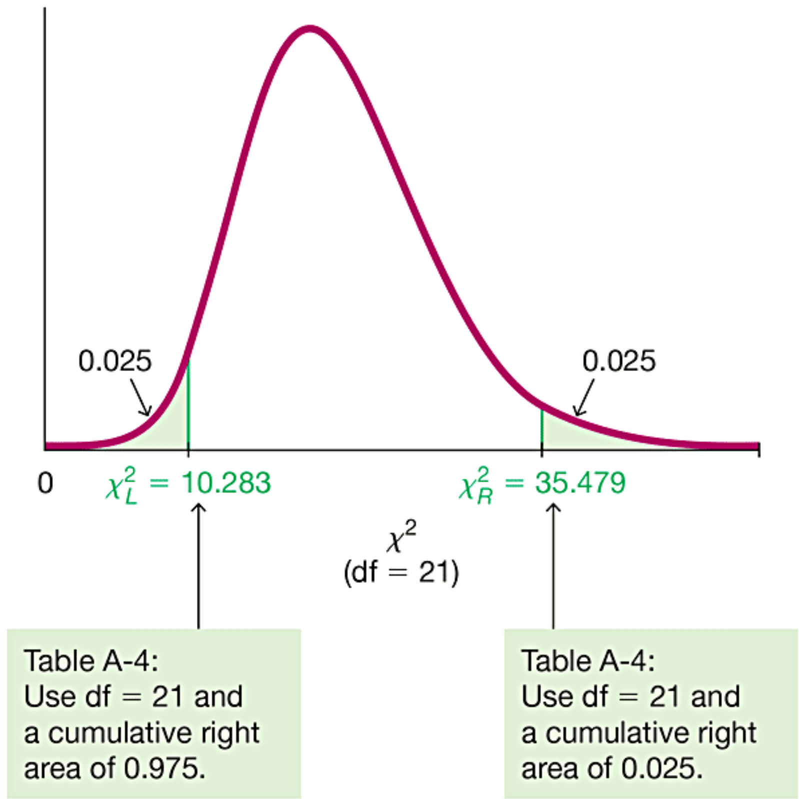
Example: Finding Critical Value of χ^2 (1 of 4)

A simple random sample of 22 IQ scores is obtained. Construction of a confidence interval for the population standard deviation s requires the left and right critical values of χ^2 corresponding to a confidence level of 95% and a sample size of $n = 22$. Find χ^2_L (the critical value of X^2 separating an area of 0.025 in the left tail), and find χ^2_R (the critical value of X^2 separating an area of 0.025 in the right tail).

Example: Finding Critical Value of χ^2 (2 of 4)

Solution

With a sample size of $n = 22$, the number of degrees of freedom is $\text{df} = n - 1 = 21$.

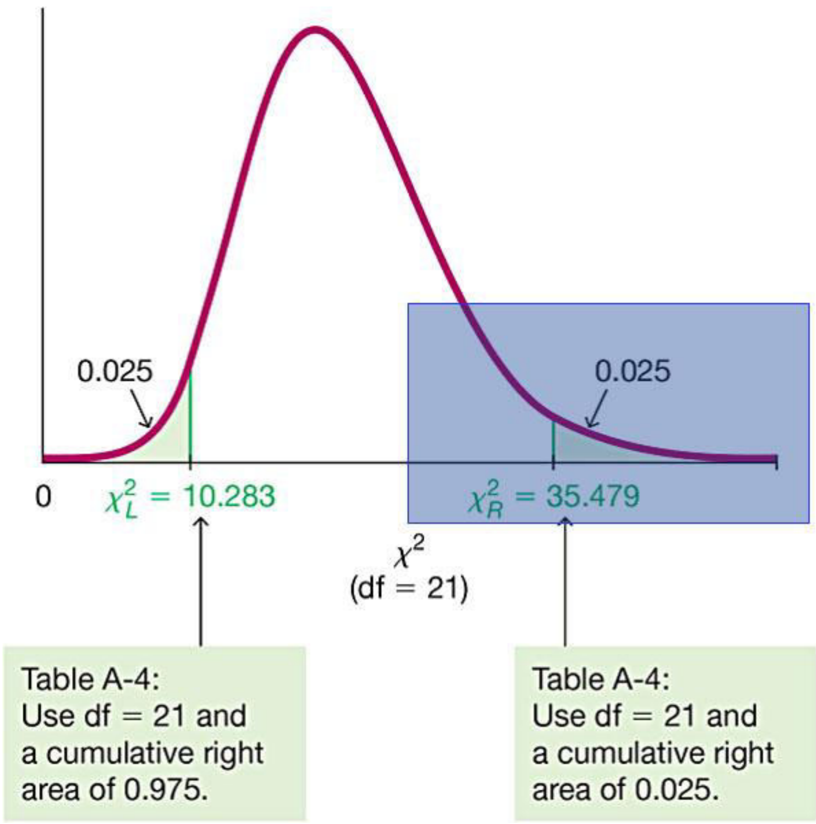


Example: Finding Critical Value of χ^2 (3 of 4)

Solution

$\chi^2_R = 35.479$

The critical value to the right is obtained from Table A-4 in a straightforward manner by locating 21 in the degrees-of-freedom column at the left and 0.025 across the top row.



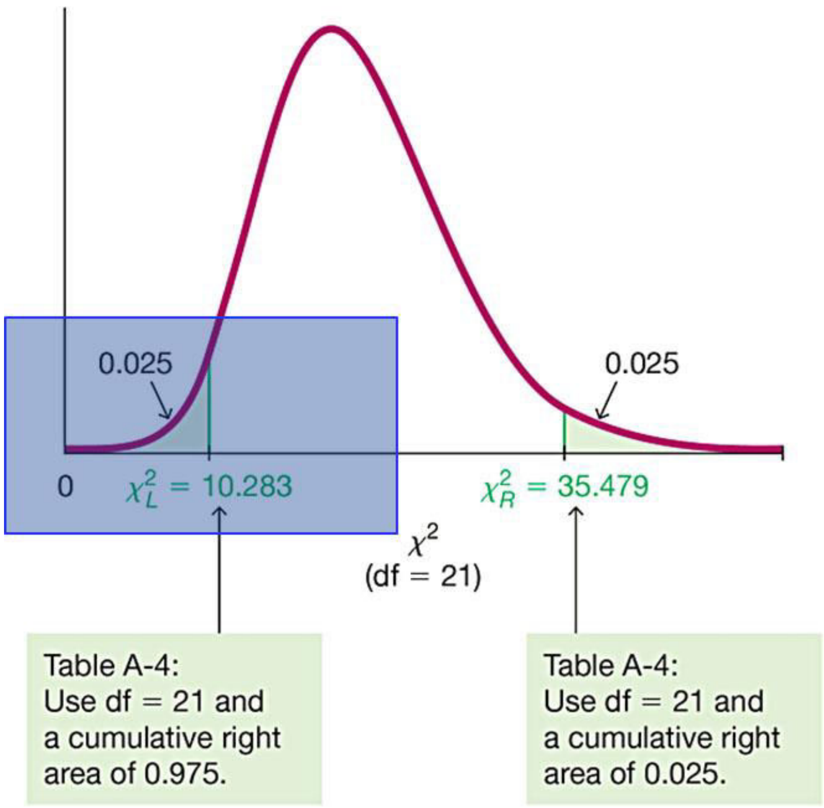
Pearson

Example: Finding Critical Value of χ^2 (4 of 4)

Solution

$\chi^2_L = 10.283$

But, we must locate 0.975 (or 1 − 0.025) across the top row because the values in the top row are always **areas to the right** of the critical value. The total area to the right of is 0.975.



Confidence Interval for Estimating a Population Standard Deviation or Variance: Objective

Construct a confidence interval estimate of a population standard deviation or variance.

Confidence Interval for Estimating a Population Standard Deviation or Variance: Notation

σ = population standard deviation

σ^2 = population variance

s = sample standard deviation

s^2 = sample variance

n = number of sample values

E = margin of error χ^2

χ^2_L = left-tailed critical value of χ^2

χ^2_R = left-tailed critical value of χ^2

Confidence Interval for Estimating a Population Standard Deviation or Variance: Requirements

1. The sample is a simple random sample.
2. The population must have normally distributed values. The requirement of a normal distribution is much stricter here than in earlier sections, so large departures from normal distributions can result in large errors.

Confidence Interval for Estimating a Population Standard Deviation or Variance: Confidence Interval for the Population Variance σ^2

$$\frac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2}$$

Confidence Interval for Estimating a Population Standard Deviation or Variance: Confidence Interval for the Population Standard Deviation *σ*

$$\sqrt{\frac{(n-1)s^2}{\chi_R^2}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi_L^2}}$$

Confidence Interval for Estimating a Population Standard Deviation or Variance: Round-Off Rule

1. **Original Data:** When using the **original set of data** values, round the confidence interval limits to one more decimal place than is used for the original data.
2. **Summary Statistics:** When using the **summary statistics** (n, s), round the confidence interval limits to the same number of decimal places used for the sample standard deviation.

Procedure for Constructing a Confidence Interval for σ or σ^2 (1 of 2)

1. Verify that the two requirements are satisfied.
2. Using $n - 1$ degrees of freedom, find the critical values χ^2_R and χ^2_L that correspond to the desired confidence level.
3. To get a confidence interval estimate of σ^2, use the following:

$$\frac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2}$$

Procedure for Constructing a Confidence Interval for σ or σ^2 (2 of 2)

4. To get a confidence interval estimate of σ, take the square root of each component of the above confidence interval.

5. Round the confidence interval limits using the round-off rule.

Using Confidence Intervals for Comparisons or Hypothesis Tests

Comparisons Confidence intervals can be used **informally** to compare the variation in different data sets, but **the overlapping of confidence intervals should not be used for making formal and final conclusions about equality of variances or standard deviations.**

Example: Confidence Interval for Estimating σ of IQ Scores (1 of 6)

Data Set 7 "IQ and Lead" in Appendix B lists IQ scores for subjects in three different lead exposure groups. The 22 full IQ scores for the group with medium exposure to lead (Group 2) have a standard deviation of 14.29263. Consider the sample to be a simple random sample and construct a 95% confidence interval estimate of σ, the standard deviation of the population from which the sample was obtained.

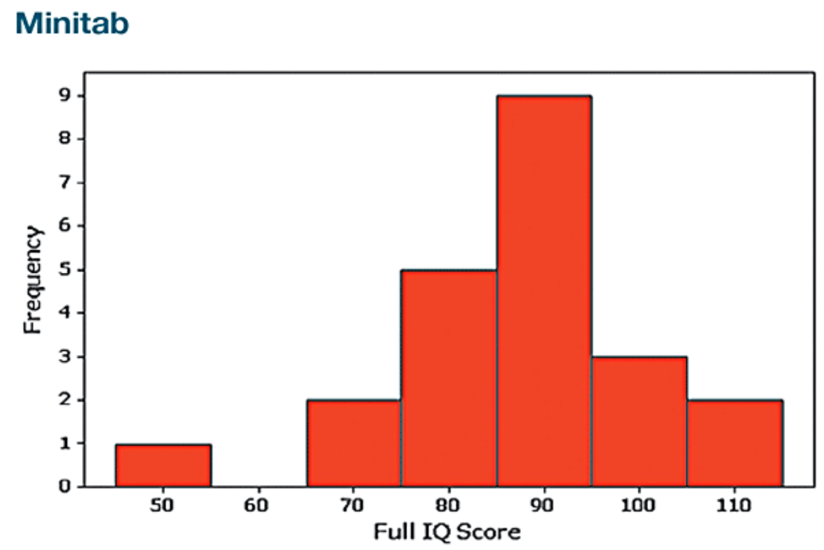
Example: Confidence Interval for Estimating σ of IQ Scores (2 of 6)

Solution

Requirement Check

Step 1: Check requirements. (1) The sample can be treated as a simple random sample.

(2) The accompanying histogram has a shape very close to the bell shape of a normal distribution.



Example: Confidence Interval for Estimating σ of IQ Scores (3 of 6)

Solution

Step 2: If using Table A-4, we use the sample size of $n = 22$ to find degrees of freedom: df = $n - 1 = 21$. Refer to the row corresponding to 21 degrees of freedom, and refer to the columns with areas of 0.975 and 0.025. (For a 95% CI, we divide $\alpha = 0.05$ equally between the two tails of the chi-square distribution, and we refer to the values of 0.975 and 0.025 across the top row.)

The critical values are $\chi^2_L = 10.283$ and $\chi^2_R = 35.479$.

Example: Confidence Interval for Estimating σ of IQ Scores (4 of 6)

Solution

Step 3: Using the critical values of 10.283 and 35.479, the sample standard deviation of s = 14.29263 and the sample size of n = 22, we construct the 95% confidence interval by evaluating the following:

$$\frac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2}$$

$$\frac{(22-1)(14.29263)^2}{35.479} < \sigma^2 < \frac{(22-1)(14.29263)^2}{10.283}$$

Example: Confidence Interval for Estimating σ of IQ Scores (5 of 6)

Solution

Step 4: Evaluating the expression results in

$$120.9 < \sigma^2 < 417.2.$$

Finding the square root of each part (before rounding), then rounding to one decimal place, yields this 95% confidence interval estimate of the population standard deviation: $11.0 < \sigma < 20.4$.

Example: Confidence Interval for Estimating σ of IQ Scores (6 of 6)

Interpretation

Based on this result, we have 95% confidence that the limits of 11.0 and 20.4 contain the true value of s. The confidence interval can also be expressed as (11.0, 20.4), **but it cannot be expressed in a format of** $s \pm E$.

Determining Sample Sizes (1 of 4)

The procedures for finding the sample size necessary to estimate σ are much more complex than the procedures given earlier for means and proportions. For normally distributed populations, the table on the following slide, or the following formula can be used:

$$n = \frac{1}{2}\left(\frac{z_{\frac{\alpha}{2}}}{d}\right)^2.$$

Determining Sample Sizes (2 of 4)

Finding Sample Size

σ

To be 95% confident that *s* is within …	Of the value of σ, the sample size *n* should be at least
1%	19,205
5%	768
10%	192
20%	48
30%	21
40%	12
50%	8

Determining Sample Sizes (3 of 4)

Finding Sample Size

σ

To be 99% confident that s is within …	Of the value of σ, the sample size n should be at least
1%	33,218
5%	1,338
10%	336
20%	85
30%	38
40%	22
50%	14

Determining Sample Sizes (4 of 4)

Statdisk also provides sample sizes. With Statdisk, select **Analysis, Sample Size Determination,** and then **Estimate Standard Deviation.** Excel, StatCrunch, and the TI-83/84 Plus calculator do not provide such sample sizes.

Example: Finding Sample Size for Estimating σ (1 of 2)

We want to estimate the standard deviation σ of all IQ scores of people with exposure to lead. We want to be 99% confident that our estimate is within 5% of the true value of σ. How large should the sample be? Assume that the population is normally distributed.

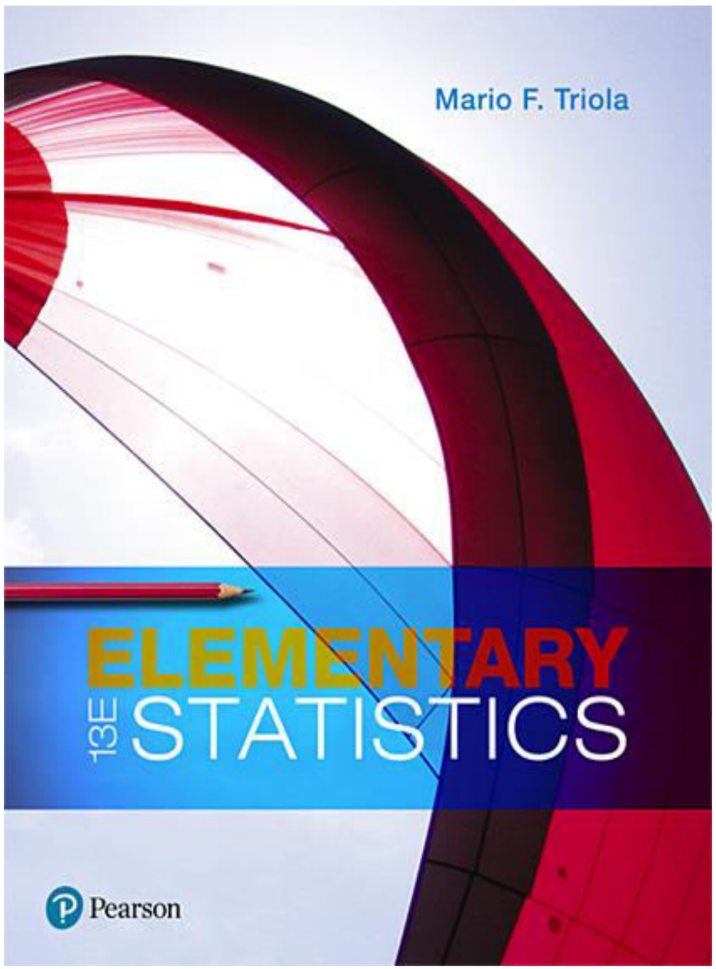
Example: Finding Sample Size for Estimating σ (2 of 2)

Solution

From the table given for finding sample size, we can see that 99% confidence and an error of 5% for *s* correspond to a sample of size 1336. We should obtain a simple random sample of 1336 IQ scores from the population of subjects exposed to lead.

Elementary Statistics

Thirteenth Edition



Chapter 7

Estimating Parameters and Determining Sample Sizes

Estimating Parameters and Determining Sample Sizes

7-1 Estimating a Population Proportion

7-2 Estimating a Population Mean

7-3 Estimating a Population Standard Deviation or Variance

7-4 Bootstrapping: Using Technology for Estimates

Key Concept

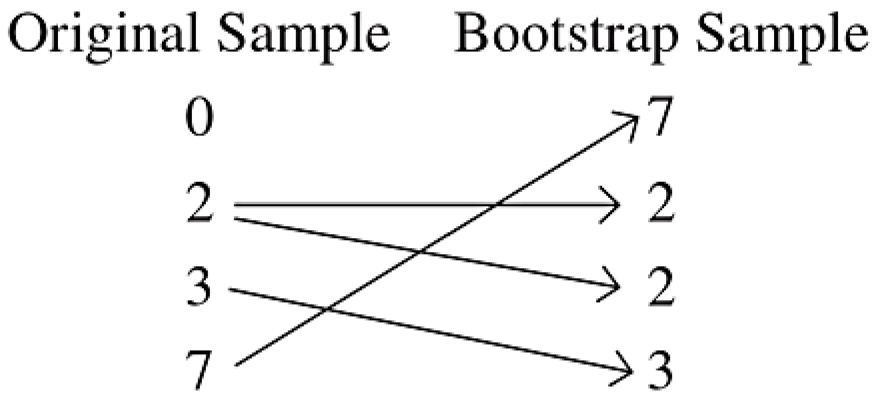
The preceding sections presented methods for estimating population proportions, means, and standard deviations (or variances). All of those methods have certain requirements that limit the situations in which they can be used. When some of the requirements are not satisfied, we can often use the bootstrap method to estimate a parameter with a confidence interval. The bootstrap method typically requires the use of software.

Bootstrap Sample

- Bootstrap Sample
  - Given a simple random sample of size *n*, a **bootstrap sample** is another random sample of *n* values obtained **with replacement** from the original sample.

Example: Bootstrap Sample of Incomes (1 of 2)

When the author collected annual incomes of current statistics students, he obtained these results (in thousands of dollars): 0, 2, 3, 7.



Example: Bootstrap Sample of Incomes (2 of 2)

The sample of {7, 2, 2, 3} is one bootstrap sample obtained from the original sample. Other bootstrap samples may be different.

Incomes tend to have distributions that are skewed instead of being normal, so we should not use the methods of Section 7-2 with a small sample of incomes. This is a situation in which the bootstrap method comes to the rescue.

Bootstrap Procedure for a Confidence Interval Estimate of a Parameter (1 of 2)

1. Given a simple random sample of size *n*, obtain many (such as 1000 or more) bootstrap samples of the same size *n*.
2. For the parameter to be estimated, find the corresponding statistic for each of the bootstrap samples. (Example: For a confidence estimate of μ, find the **sample mean** $\bar{x}$ from each bootstrap sample.)
3. Sort the list of sample statistics from low to high.

Bootstrap Procedure for a Confidence Interval Estimate of a Parameter (2 of 2)

4. Using the sorted list of the statistics, create the confidence interval by finding corresponding percentile values. Procedures for finding percentiles are given in Section 3-3. (Example: Using a list of sorted sample means, the 90% confidence interval limits are P_5 and P_{95}. The 90% confidence interval estimate of μ is $P_5 < \mu < P_{95}$.)

Proportions

When working with proportions, it is very helpful to represent the data from the two categories by using 0's and 1's.

Example: Eye Color Survey: Bootstrap CI for Proportion (1 of 11)

In a survey, four randomly selected subjects were asked if they have brown eyes, and here are the results: 0, 0, 1, 0 (where 0 = no and 1 = yes). Use the bootstrap resampling procedure to construct a 90% confidence interval estimate of the population proportion *p*, the proportion of people with brown eyes in the population.

Example: Eye Color Survey: Bootstrap CI for Proportion (2 of 11)

Solution

Requirement Check

The sample is a simple random sample. (There is no requirement of at least 5 successes and at least 5 failures or $np \geq 5$ and $nq \geq 5$. There is no requirement that the sample must be from a normally distributed population.)

Example: Eye Color Survey: Bootstrap CI for Proportion (3 of 11)

Solution

Step 1:

In the table on the next slide, we created 20 bootstrap samples from the original sample of 0, 0, 1, 0.

Example: Eye Color Survey: Bootstrap CI for Proportion (4 of 11)

Solution

Bootstrap Samples for *p*.

Bootstrap Sample				$\hat{p}$	Sorted $\hat{p}$	
1	0	0	1	0.50	0.00	$P_5 = 0.00$
1	0	1	0	0.50	0.00	
0	1	1	1	0.75	0.00	
0	0	0	0	0.00	0.00	
0	1	0	0	0.25	0.25	
1	0	0	0	0.25	0.25	
0	1	0	1	0.50	0.25	
1	0	0	0	0.25	0.25	
0	0	0	0	0.00	0.25	
0	0	1	1	0.50	0.25	**90% Confidence Interval:**
0	0	0	1	0.25	0.25	$0.00 < p < 0.75$
0	0	1	0	0.25	0.25	
1	1	1	0	0.75	0.50	
0	0	0	0	0.00	0.50	
0	0	0	0	0.00	0.50	
0	1	1	0	0.50	0.50	
0	0	1	0	0.25	0.50	
1	0	0	0	0.25	0.75	
1	1	1	0	0.75	0.75	$P_{95} = 0.75$
0	0	0	1	0.25	0.75	

Example: Eye Color Survey: Bootstrap CI for Proportion (5 of 11)

Solution

Step 2:

Because we want a confidence interval estimate of the population proportion *p*, we want the sample proportion $\hat{p}$ for each of the 20 bootstrap samples, and those sample proportions are shown in the column to the right of the bootstrap samples on the next slide.

Example: Eye Color Survey: Bootstrap CI for Proportion (6 of 11)

Solution

Bootstrap Samples for *p*.

Bootstrap Sample				$\hat{p}$	Sorted $\hat{p}$
1	0	0	1	0.50	0.00
1	0	1	0	0.50	0.00
0	1	1	1	0.75	0.00
0	0	0	0	0.00	0.00
0	1	0	0	0.25	0.25
1	0	0	0	0.25	0.25
0	1	0	1	0.50	0.25
1	0	0	0	0.25	0.25
0	0	0	0	0.00	0.25
0	0	1	1	0.50	0.25
0	0	0	1	0.25	0.25
0	0	1	0	0.25	0.25
1	1	1	0	0.75	0.50
0	0	0	0	0.00	0.50
0	0	0	0	0.00	0.50
0	1	1	0	0.50	0.50
0	0	1	0	0.25	0.50
1	0	0	0	0.25	0.75
1	1	1	0	0.75	0.75
0	0	0	1	0.25	0.75

$P_5 = 0.00$

90% Confidence Interval:

$0.00 < p < 0.75$

$P_{95} = 0.75$

Example: Eye Color Survey: Bootstrap CI for Proportion (7 of 11)

Solution

Step 3:

The column of data shown farthest to the right is a list of the 20 sample proportions arranged in order (“sorted”) from lowest to highest.

Example: Eye Color Survey: Bootstrap CI for Proportion (8 of 11)

Solution

Bootstrap Samples for *p*.

Bootstrap Sample				$\hat{p}$	Sorted $\hat{p}$	
1	0	0	1	0.50	0.00	$P_5 = 0.00$
1	0	1	0	0.50	0.00	
0	1	1	1	0.75	0.00	
0	0	0	0	0.00	0.00	
0	1	0	0	0.25	0.25	
1	0	0	0	0.25	0.25	
0	1	0	1	0.50	0.25	
1	0	0	0	0.25	0.25	
0	0	0	0	0.00	0.25	
0	0	1	1	0.50	0.25	**90% Confidence Interval:**
0	0	0	1	0.25	0.25	$0.00 < p < 0.75$
0	0	1	0	0.25	0.25	
1	1	1	0	0.75	0.50	
0	0	0	0	0.00	0.50	
0	0	0	0	0.00	0.50	
0	1	1	0	0.50	0.50	
0	0	1	0	0.25	0.50	
1	0	0	0	0.25	0.75	
1	1	1	0	0.75	0.75	$P_{95} = 0.75$
0	0	0	1	0.25	0.75	

Example: Eye Color Survey: Bootstrap CI for Proportion

Solution

Step 4:

Because we want a confidence level of 90%, we want to find the percentiles P_5 and P_{95}. Recall that P_5 separates the lowest 5% of values, and P_{95} separates the top 5% of values.

Using the methods from Section 3-3 for finding percentiles, we use the *sorted* list of bootstrap sample proportions to find that $P_5 = 0.00$ and $P_{95} = 0.75$. The 90% confidence interval estimate of the population proportion is $0.00 < p < 0.75$.

Example: Eye Color Survey: Bootstrap CI for Proportion (10 of 11)

Solution

Bootstrap Samples for *p*.

Bootstrap Sample				$\hat{p}$	Sorted $\hat{p}$
1	0	0	1	0.50	0.00
1	0	1	0	0.50	0.00
0	1	1	1	0.75	0.00
0	0	0	0	0.00	0.00
0	1	0	0	0.25	0.25
1	0	0	0	0.25	0.25
0	1	0	1	0.50	0.25
1	0	0	0	0.25	0.25
0	0	0	0	0.00	0.25
0	0	1	1	0.50	0.25
0	0	0	1	0.25	0.25
0	0	1	0	0.25	0.25
1	1	1	0	0.75	0.50
0	0	0	0	0.00	0.50
0	0	0	0	0.00	0.50
0	1	1	0	0.50	0.50
0	0	1	0	0.25	0.50
1	0	0	0	0.25	0.75
1	1	1	0	0.75	0.75
0	0	0	1	0.25	0.75

$\longrightarrow P_5 = 0.00$

90% Confidence Interval:

$0.00 < p < 0.75$

$\longrightarrow P_{95} = 0.75$

Pearson

Example: Eye Color Survey: Bootstrap CI for Proportion (11 of 11)

Interpretation

The confidence interval of $0.00 < p < 0.75$ is quite wide. After all, every confidence interval for every proportion must fall between 0 and 1, so the 90% confidence interval of $0.00 < p < 0.75$ doesn’t seem to be helpful, but it is based on only four sample values.

Means

Earlier in this chapter we noted that when constructing a confidence interval estimate of a population mean, there is a requirement that the sample is from a normally distributed population or the sample size is greater than 30. The bootstrap method can be used when this requirement is not satisfied.

Example: Incomes: Bootstrap CI for Mean (1 of 10)

When the author collected a simple random sample of annual incomes of his statistics students, he obtained these results (in thousands of dollars): 0, 2, 3, 7. Use the bootstrap resampling procedure to construct a 90% confidence interval estimate of the mean annual income of the population of all of the author's statistics students.

Example: Incomes: Bootstrap CI for Mean (2 of 10)

Solution

Requirement Check

The sample is a simple random sample and there is no requirement that the sample must be from a normally distributed population. Because distributions of incomes are typically skewed instead of normal, we should not use the methods of Section 7-2 for finding the confidence interval, but the bootstrap method can be used.

Example: Incomes: Bootstrap CI for Mean (3 of 10)

Solution

Step 1:

In the table on the next slide, we created 20 bootstrap samples (with replacement!) from the original sample of 0, 2, 3, 7. (Here we use only 20 bootstrap samples so we have a manageable example that doesn’t occupy many pages of text, but we usually want at least 1000 bootstrap samples.)

Example: Incomes: Bootstrap CI for Mean (4 of 10)

Solution

Bootstrap Samples for μ.

Bootstrap Sample				$\bar{x}$	Sorted $\bar{x}$	
3	3	0	2	2.00	1.75	$P_5 = 1.75$
0	3	2	2	1.75	1.75	
7	0	2	7	4.00	1.75	
3	2	7	3	3.75	2.00	
0	0	7	2	2.25	2.00	
7	0	0	3	2.50	2.25	
3	0	3	2	2.00	2.50	
3	7	3	7	5.00	2.50	
0	3	2	2	1.75	2.50	
0	3	7	0	2.50	2.75	**90% Confidence Interval:**
0	7	2	2	2.75	3.00	$1.75 < \mu < 4.875$
7	2	2	3	3.50	3.25	
7	2	3	7	4.75	3.25	
2	7	2	7	4.50	3.50	
0	7	2	3	3.00	3.75	
7	3	7	2	4.75	4.00	
3	7	0	3	3.25	4.50	
0	0	3	7	2.50	4.75	
3	3	7	0	3.25	4.75	
2	0	2	3	1.75	5.00	$P_{95} = 4.875$

Example: Incomes: Bootstrap CI for Mean (5 of 10)

Solution

Step 2:

Because we want a confidence interval estimate of the population mean μ, we want the sample mean $\bar{x}$ for each of the 20 bootstrap samples, and those sample means are shown in the column to the right of the bootstrap samples.

Example: Incomes: Bootstrap CI for Mean (6 of 10)

Solution

Bootstrap Samples for μ.

Bootstrap Sample				$\bar{x}$	Sorted $\bar{x}$	
3	3	0	2	2.00	1.75	
0	3	2	2	1.75	1.75	$\rightarrow P_5 = 1.75$
7	0	2	7	4.00	1.75	
3	2	7	3	3.75	2.00	
0	0	7	2	2.25	2.00	
7	0	0	3	2.50	2.25	
3	0	3	2	2.00	2.50	
3	7	3	7	5.00	2.50	
0	3	2	2	1.75	2.50	
0	3	7	0	2.50	2.75	**90% Confidence Interval:**
0	7	2	2	2.75	3.00	$1.75 < \mu < 4.875$
7	2	2	3	3.50	3.25	
7	2	3	7	4.75	3.25	
2	7	2	7	4.50	3.50	
0	7	2	3	3.00	3.75	
7	3	7	2	4.75	4.00	
3	7	0	3	3.25	4.50	
0	0	3	7	2.50	4.75	
3	3	7	0	3.25	4.75	
2	0	2	3	1.75	5.00	$\rightarrow P_{95} = 4.875$

Pearson

Example: Incomes: Bootstrap CI for Mean (7 of 10)

Solution

Step 3:

The column of data shown farthest to the right is a list of the 20 sample means arranged in order (“sorted”) from lowest to highest.

Example: Incomes: Bootstrap CI for Mean (8 of 10)

Solution

Bootstrap Samples for μ.

Bootstrap Sample				$\bar{x}$	Sorted $\bar{x}$
3	3	0	2	2.00	1.75
0	3	2	2	1.75	1.75
7	0	2	7	4.00	1.75
3	2	7	3	3.75	2.00
0	0	7	2	2.25	2.00
7	0	0	3	2.50	2.25
3	0	3	2	2.00	2.50
3	7	3	7	5.00	2.50
0	3	2	2	1.75	2.50
0	3	7	0	2.50	2.75
0	7	2	2	2.75	3.00
7	2	2	3	3.50	3.25
7	2	3	7	4.75	3.25
2	7	2	7	4.50	3.50
0	7	2	3	3.00	3.75
7	3	7	2	4.75	4.00
3	7	0	3	3.25	4.50
0	0	3	7	2.50	4.75
3	3	7	0	3.25	4.75
2	0	2	3	1.75	5.00

$P_5 = 1.75$

90% Confidence Interval:

$1.75 < \mu < 4.875$

$P_{95} = 4.875$

Example: Incomes: Bootstrap CI for Mean (9 of 10)

Solution

Step 4:

Because we want a confidence level of 90%, we want to find the percentiles P_5 and P_{95}. Recall that P_5 separates the lowest 5% of values, and P_{95} separates the top 5% of values. Using the methods from Section 3-3 for finding percentiles, we use the **sorted** list of bootstrap sample means to find that $P_5 = 1.75$ and $P_{95} = 4.875$. The 90% confidence interval estimate of the population mean is $1.75 < \mu < 4.875$, where the values are in thousands of dollars.

Example: Incomes: Bootstrap CI for Mean (10 of 10)

Solution

Bootstrap Samples for μ.

Bootstrap Sample				$\bar{x}$	Sorted $\bar{x}$
3	3	0	2	2.00	1.75
0	3	2	2	1.75	1.75
7	0	2	7	4.00	1.75
3	2	7	3	3.75	2.00
0	0	7	2	2.25	2.00
7	0	0	3	2.50	2.25
3	0	3	2	2.00	2.50
3	7	3	7	5.00	2.50
0	3	2	2	1.75	2.50
0	3	7	0	2.50	2.75
0	7	2	2	2.75	3.00
7	2	2	3	3.50	3.25
7	2	3	7	4.75	3.25
2	7	2	7	4.50	3.50
0	7	2	3	3.00	3.75
7	3	7	2	4.75	4.00
3	7	0	3	3.25	4.50
0	0	3	7	2.50	4.75
3	3	7	0	3.25	4.75
2	0	2	3	1.75	5.00

$P_5 = 1.75$

90% Confidence Interval:

$1.75 < \mu < 4.875$

$P_{95} = 4.875$

Pearson

Standard Deviations

In Section 7-3 we noted that when constructing confidence interval estimates of population standard deviations or variances, there is a requirement that the sample must be from a population with normally distributed values. Even if the sample is large, this normality requirement is much stricter than the normality requirement used for estimating population means. Consequently, the bootstrap method becomes more important for confidence interval estimates of σ or σ^2.

Example: Incomes: Bootstrap CI for Standard Deviation (1 of 4)

Use these same incomes (thousands of dollars) from the previous example: 0, 2, 3, 7. Use the bootstrap resampling procedure to construct a 90% confidence interval estimate of the population standard deviation σ, the standard deviation of the annual incomes of the population of the author's statistics students.

Example: Incomes: Bootstrap CI for Standard Deviation (2 of 4)

Solution

Requirement Check

The same requirement check used in the previous example applies here.

Example: Incomes: Bootstrap CI for Standard Deviation (3 of 4)

Solution

The same basic procedure used in the previous example is used here. The previous example already includes 20 bootstrap samples, so here we find the **standard deviation** of each bootstrap sample, and then we sort them to get this sorted list of sample standard deviations:

1.26	1.26	1.26	1.41	1.41	2.22	2.31	2.38	2.63	2.63
2.87	2.87	2.89	2.94	2.99	3.30	3.32	3.32	3.32	3.56

Example: Incomes: Bootstrap CI for Standard Deviation (4 of 4)

Solution

The 90% confidence interval limits are found from this sorted list of standard deviations by finding P_5 and P_{95}. Using the methods from Section 3-3, we get $P_5 = 1.26$ and $P_{95} = 3.44$. The 90% confidence interval estimate of the population standard deviation s is $1.26 < \sigma < 3.44$, where the values are in thousands of dollars.