9.1 Estimating a Population Proportion

Definitions:

**Confidence Interval** – an interval of numbers for an unknown parameter.

*Example*: An interval for the population proportion \( p \), which is a parameter, could be indicated as \((x_L, x_U)\), where \( x_L \) is the lower bound and \( x_U \) is the upper bound; that is, with a certain level of confidence, we estimate that \( x_L < p < x_U \).

**Significance Level** – the probability \( \alpha \) of obtaining a sample with a confidence interval that does not contain the parameter.

*Example*: Suppose for \( \alpha = 0.05 \), we obtain confidence intervals for \( p \) for a number of samples of some size \( n \). Then we would expect that of all the confidence intervals, only 5% would fail to contain the true value of \( p \).

**Confidence Level** – the expected proportion of intervals that will contain the parameter if a large enough number of samples is obtained, which is just the complement of the significance level, that is, \((1 - \alpha)100\%\).

*Example*: A 95% confidence level \( (\alpha = 0.05) \) implies that if 100 different confidence intervals are constructed, each based on a different sample from the population, then we will expect 95 of the intervals to include the parameter and 5 to not include the parameter.

**Point Estimate** – the value of a statistic that estimates the value of a parameter, and is the center of the confidence interval.

*Example*: The point estimate for a population proportion \( p \) is the sample proportion \( \hat{p} \). This is the center of the confidence interval. That is, for a confidence interval \((x_L, x_U)\), we have that

\[
\hat{p} - x_L = x_U - \hat{p} \quad \Rightarrow \quad \hat{p} = \frac{x_L + x_U}{2}
\]
Chapter IX  Estimating the Value of a Parameter  9-2

**Critical Value** – the \( z \)-score \( z_{\alpha/2} \) that separates the middle portion of the normal distribution from the right and left tails for a confidence level of \((1 - \alpha)100\%\).

*Example*: Find the critical values corresponding to confidence levels of 90\%, 95\%, and 99\%.

*Solution*: Using the `invNorm` function or Table V, we construct the following table.

<table>
<thead>
<tr>
<th>((1 - \alpha)100%)</th>
<th>(\alpha/2)</th>
<th>(z_{\alpha/2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>0.05</td>
<td>1.645</td>
</tr>
<tr>
<td>95%</td>
<td>0.025</td>
<td>1.960</td>
</tr>
<tr>
<td>99%</td>
<td>0.005</td>
<td>2.575</td>
</tr>
</tbody>
</table>

**Standardized Statistic** – a statistic is standardized by calculating its \( z \)-score.

*Example*: A sample gives a confidence interval \((x_L, x_U)\), centered about \( \hat{p} \), for the population proportion \( p \). We standardize the lower and upper boundaries of the confidence interval so that they coincide with the critical values corresponding to the left and right tails.

\[
 z_{\alpha/2} = \frac{x_U - \hat{p}}{\sigma_{\hat{p}}} \quad \Rightarrow \quad x_U = \hat{p} + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}
\]

\[
 -z_{\alpha/2} = \frac{x_L - \hat{p}}{\sigma_{\hat{p}}} \quad \Rightarrow \quad x_L = \hat{p} - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}
\]

**Margin of Error** – a measure \( E \) of the accuracy of a point estimate in estimating a parameter using a confidence interval. A confidence interval is of the form

\[
 \text{confidence interval} = \text{point estimate} \pm \text{margin of error}
\]

*Example*: Suppose a confidence interval for the population proportion \( p \) is of the form \((x_L, x_U)\). Then

\[
 E = \frac{x_U - x_L}{2}
\]
\[ x_L = \hat{p} - E, \quad x_U = \hat{p} + E \]

\[ (x_L, x_U) = (\hat{p} - E, \hat{p} + E) \]

**Confidence Interval for \( p \) – a \((1 - \alpha)100\%\) confidence interval can be constructed under the following conditions:**

1. Simple random sample of size \( n \).

2. The sampling distribution of \( \hat{p} \) is normal, that is,

\[ n\hat{p}(1 - \hat{p}) \geq 10 \]

3. Each sample must be independent, that is,

\[ n \leq 0.05N \]

where \( N \) is the population size.

Since \( p \) is unknown, we will use \( \hat{p} \) to estimate the standard deviation and obtain the error term

\[ E = z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \]

Hence the confidence interval is

\[ \left( \hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right) \]

Note that the margin of error \( E \) depends on three things:

1. The critical value \( z_{\alpha/2} \). A desire for a higher level of confidence necessitates a higher confidence interval. Therefore, increasing \( z_{\alpha/2} \) increases the error.

2. The product \( \hat{p}(1 - \hat{p}) \), which has a maximum value of 0.25 when \( \hat{p} = 0.5 \). An equally likely chance for success or failure increases the prediction error.

3. The sample size \( n \). Due to the law of large numbers, as we sample more and more of the population, the difference between the statistic \( \hat{p} \) and the parameter \( p \) decreases. Therefore, as \( n \) increases, the error term decreases.
**Example:** Construct a 90% confidence interval of the population proportion for a number of successes \( x = 30 \) and sample size \( n = 150 \).

**Solution:** Using the TI 83/84, we select STAT | TESTS | 1-PropZInt... We enter data as shown below left, and obtain the interval shown below right.

```
1-PropZInt
\[ \hat{x} = 30 \]
\[ n = 150 \]
\[ \text{C-Level: 90} \]

1-PropZInt
\[ \hat{p} = 0.2 \]
\[ n = 150 \]

Calculate
```

Note that the calculator returns the point estimate

\[ \hat{p} = \frac{30}{150} = 0.2 \]

This may be used to find the error term in one of three ways:

\[ E = 0.2 - 0.14628 \]
\[ = 0.25372 - 0.2 \]
\[ = \frac{0.25372 - 0.14628}{2} \]
\[ = 0.05372 \]

**Sample Size for Margin of Error** – the sample size needed to be within a certain margin of error may be calculated by solving the margin of error expression for \( n \). This gives

\[ n \geq \left( \frac{z_{\alpha}}{E} \right)^2 \hat{p}(1 - \hat{p}) \]

To determine the sample size, we need to know the confidence level, the error, and the sample proportion. But we don’t know the sample proportion until we actually obtain the sample. And we can’t obtain the sample until we know the sample size, which is what we are trying to determine. Therefore, we could do one of two things:

(1) Use an estimate of \( \hat{p} \) from a previous study.

(2) Use a value of \( \hat{p} \) that maximizes \( \hat{p}(1 - \hat{p}) \). This would be the value \( \hat{p} = 0.5 \), which gives \( \hat{p}(1 - \hat{p}) = 0.25 \).
Example: A researcher wishes to estimate the proportion of adults who have high-speed Internet access. What size sample should be obtained if she wishes the estimate to be within 0.03 with 99% confidence if

(a) she uses a 2007 estimate of 0.69 obtained from Nielson NetRatings?
(b) she does not use any prior estimates?

Solution: (a) First, we determine that $1 - \alpha = 0.99 \Rightarrow \alpha = 0.01 \Rightarrow \alpha/2 = 0.005$. Then $z_{0.005} = 2.575$. So

\[
\begin{align*}
  n & \geq \left( \frac{2.575}{0.03} \right)^2 0.69(0.31) = 1575.9 \\
\end{align*}
\]

We always round up to the nearest integer. Thus $n = 1576$.

(b) Now we use $\hat{p}(1 - \hat{p}) = 0.25$ and obtain

\[
\begin{align*}
  n & \geq \left( \frac{2.575}{0.03} \right)^2 0.25 = 1841.8 \\
\end{align*}
\]

Thus $n = 1842$.

Exercises:

1. Concepts and vocabulary (1 – 6; 46 – 51).
2. Determine the critical value corresponding to a confidence level (7 – 10).
3. Determine the point estimate and margin of error from a confidence interval (11 – 14).
4. Find a point estimate and construct a confidence interval for the population proportion (15 – 20).
5. Construct and interpret a confidence interval for the population proportion (21, 25, 29, 41).
6. Determine the sample size necessary for estimating a population proportion within a margin of error (33, 37, 39).
9.2 Estimating a Population Mean

Definitions:

Point Estimate for \( \mu \) – the sample mean \( \bar{x} \). So the confidence interval for the population mean has the form

\[
(\bar{x} - E, \bar{x} + E)
\]

To find the error term, we could start by standardizing \( \bar{x} \) as was done for \( \hat{p} \) to get

\[
z = \frac{\bar{x} - \mu}{\sigma} = \frac{\bar{x} - \mu}{s/\sqrt{n}}
\]

And then for a confidence level \((1 - \alpha)100\%\), rewrite the expression to obtain the error

\[
E = \bar{x} - \mu = \frac{z_{\alpha} \sigma}{\sqrt{n}}
\]

But the problem with this is that we do not know \( \sigma \) and must approximate it with \( s \) much like we approximated \( \hat{p} \) with \( p \). When this is done, instead of using the \( z \)-distribution, we use Student’s \( t \)-distribution.

Student’s \( t \)-Distribution – a sampling distribution of the \( t \) values of simple random samples of size \( n \), taken from a population that is normally distributed with mean \( \mu \), with the following properties:

1. Shape – bell-shaped, but shorter and wider than \( z \) curve (standard normal); this is because the \( t \)-distribution depends on \( s \), which is itself a random variable, unlike \( \sigma \).

2. Center – centered about \( t = 0 \).

3. Spread – depends on the degrees of freedom, \( n - 1 \); thus, each sample size has its own \( t \)-distribution; as the sample size increases, by the law of large numbers, the \( t \) curve approaches the \( z \) curve and has standard deviation 1.
Example: For a sample size of 26, find the $t$-value such that the area in the right tail is 0.10.

Solution: Table VI gives $t$-values for an area to the right. Entering the table with area = 0.1 and degrees of freedom, $df = 25$, we read that

$$t_{0.10} = 1.316$$

This value may be calculated using the TI 84 by selecting DISTR | invT(0.9,25). Note that unlike the table, the calculator uses the area to the left.

Example: For $n = 19$, find the $t$-value such that the area to the left is 0.01.

Solution: Now we wish to find $t_{0.99}$ for $df = 18$. Again entering Table VI, we look at column 0.01 and row 18 to retrieve the value $t_{0.01} = 2.552$. Therefore, by the symmetry of the $t$-distribution, we have

$$t_{0.99} = -2.552$$

Using the TI 84, we calculate $\text{invT}(0.01,18) = -2.552$.

Example: Find the critical $t$-value that corresponds to 90% confidence for a sample size of 21.

Solution: For $1 - \alpha = 0.9$, we have $\alpha/2 = 0.05$ in each tail. So entering Table VI and looking at row $df = 20$ and column 0.05, we read the critical value

$$t_{\alpha/2} = 1.725$$

Studentized Statistic – a statistic based on Student’s $t$-distribution, where for a simple random sample of size $n$ with mean $\bar{x}$ and standard deviation $s$,

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

Rearranging the Studentized statistic gives the error for a confidence level of $(1 - \alpha)100\%$ as
$$E = \bar{x} - \mu = \frac{t_{\alpha/2}s}{\sqrt{n}}$$

Note that the margin of error $E$ depends on three things:

(1) the critical value $t_{\alpha/2}$. A desire for a higher level of confidence necessitates a higher confidence interval. Therefore, as $1 - \alpha$ increases, so too does $t_{\alpha/2}$, and hence the error term increases.

(2) the sample standard deviation $s$. The more spread there is in the sample, and by approximation, in the population, the more spread there will be in the confidence interval. Therefore, as $s$ increases, the error term increases.

(3) the sample size $n$. Due to the law of large numbers, as we sample more and more of the population, the difference between the statistic $\bar{x}$ and the parameter $\mu$ decreases. Therefore, as $n$ increases, the error term decreases.

Confidence Interval for $\mu$ – using the point estimator and error above, we may construct a $(1 - \alpha)100\%$ confidence interval when we meet three criteria:

(1) The data must be obtained using simple random sampling; that is, every individual in the population must have an equally likely chance of being sampled.

(2) The sample size is small relative to the population; that is, $n \leq 0.05N$.

(3) The population must be normally distributed or the sample size must be large; that is, $n \geq 30$. If the sample size is small, the population is checked for normality using either a Minitab normality plot, as discussed in §7.3, or the NORMTEST program on the TI 83/84.

(4) The sample contains no outliers. This is because unlike the median, the sample mean is not resistant. Outliers can severely affect the sample mean and give errant confidence intervals. One should first construct a box plot to check for outliers.
However, the procedure for constructing a $t$-interval is robust and gives good estimates even when there are small departures from the three criteria. The interval is determined by

Point Estimator: $\bar{x}$

Error: $t_\frac{\alpha}{2} \frac{s}{\sqrt{n}}$

Interval: $\left( \bar{x} - t_\frac{\alpha}{2} \frac{s}{\sqrt{n}}, \bar{x} + t_\frac{\alpha}{2} \frac{s}{\sqrt{n}} \right)$

**Example:** An arborist is interested in determining the mean diameter of mature white oak trees. A random sample yields the data set \{64.0, 33.4, 45.8, 56.0, 51.5, 29.2, 63.7\} in centimeters. Construct a 95% confidence interval for the mean diameter of a mature white oak tree. Interpret this interval.

**Solution:** Since this is a small data set, we must check that the population is normal and there are no outliers. Using the TI 83/84, we enter the data into L1. We display the results of the normality test and the box plot below.

It can be seen that $r = 0.968 \geq 0.898 = r_{cr}$, and there are no outliers. Therefore, we proceed. Using the function STAT | TESTS | TInterval... we enter the data shown below, highlight Calculate and hit ENTER. As can be seen, we obtain a confidence interval of (36.327, 61.845). This is interpreted to mean that we are 95% confident that the mean diameter of mature white oak trees is somewhere between 36.3 and 61.8 centimeters.
Example: A Gallup poll of 547 adult Americans employed full or part time conducted August 13-16, 2007, asked, “How much total time in minutes do you spend commuting to and from work in a typical day?” Survey results indicate that $\bar{x} = 45.6$ minutes and $s = 31.4$ minutes. Construct and interpret a 90% confidence interval for the mean commute time of adult Americans employed full or part time. Is it possible that the mean commute time is less than 40 minutes? Is it likely?

Solution: Since we have $n = 547 \geq 30$, we proceed to construct a $t$-interval. Using the TI 83/84, we enter the data and obtain the results shown below.

Thus, we are 90% confident that the mean commute time of adult American full or part-time workers is somewhere between 43.4 and 47.8 minutes. It is possible that the mean commute time is 40 minutes, since the $t$-distribution goes from $-\infty$ to $\infty$. We may have obtained a sample that gave a point estimate far in the right tail such that the confidence interval did not capture the mean. But since for a 90% confidence level, only 5% is in the right tail, it would be an unusual event and is not likely.

Example: Work 9.2.32, 43, 45, 50(a-c)

Exercises:

2. Determine $t$-values (7 – 8).
3. Determine if a $t$-interval can be constructed (9 – 14).
4. Determine the point estimate and margin of error from a confidence interval (15 – 18).
5. Construct a confidence interval for a population mean given statistics (19 – 22).

6. Understand and interpret a confidence interval (23 – 30).

7. Construct a confidence interval for a population mean given data (35, 36, 38).

8. Read a Minitab output for a confidence interval (43).

9. Determine the sample size needed to construct a confidence interval for a population mean (45, 47).

10. Determine the effect of an outlier on a confidence interval (50).
9.3 Putting It Together: Which Procedure Do I Use?

Shown below is a flowchart for selecting the correct confidence interval. It applies to a simple random sample.

![Flowchart](image)

Exercises:

1. Construct a confidence interval for \( p \) \((1, 2, 8, 11)\).
2. Construct a confidence interval for \( \mu \) \((4, 7, 9, 10, 12)\).
3. Decide if a \( t \)-distribution can be constructed \((13 – 16)\).