© 2017 Your Stat Class.  Latrica L. Williams

ALL RIGHTS RESERVED. No part of this work covered by the copyright herein may be reproduced, transmitted, stored, or used in any form or by any means, graphic, electronic, or mechanical including but not limited to photocopying, recording, scanning, digitizing, taping, Web distribution, information networks, information storage and retrieval systems, without written permission of the company or author. Use in violation of these terms and conditions is prohibited.

REPRODUCING THIS WORK BY ANY MEANS WILL RESULT IN LEGAL ACTION. VIOLATORS WILL BE PROSECUTED TO THE FULL EXTENT OF THE LAW.
Chapter 7

CONFIDENCE INTERVALS
CASE STUDY:

What percentage of Americans are concerned about global warming?

To answer the question, researchers gathered a random sample of Americans and documented the responses. After constructing a confidence interval, it was determined that 51% ± 4% of all Americans are concerned about global warming.

So, what was the methodology used to determine the results?

Well, they had to do the following:

1. Compute a point estimate.
2. Determine a margin of error.
3. Construct a confidence interval.

In the following chapter, you will learn about these concepts and their applications.

You will also learn how to:

4. Determine a minimum sample size.
Section 7.1: Z-Interval

**Student Learning Outcomes**

By the end of the section,

1. You will understand the relationship between point estimates and population parameters.
2. You will compute point estimates for population means.
3. You will understand margin of error.
4. You will construct a confidence interval for a population mean with $\sigma$ known.
Thus far, the descriptive branch of statistics has been examined. Moving forward, the inferential branch of statistics will be the focus. Inferential statistics is very important in many real-world applications. It is extremely beneficial because it allows researchers to generalize sample data to the population. Recall, all subjects of interest are rarely examined due to time, money and resources. Therefore, researchers analyze samples to make predictions and/or draw conclusions about the population based on the sample results. There are three main techniques used to make inferences in statistics, which are the following: confidence intervals, hypothesis testing and regression. Confidence intervals will be explored in this chapter. In subsequent chapters, hypothesis testing and regression will be discussed.

Before diving into concepts related to confidence intervals, let’s examine the following question:

**What is the life expectancy of Americans?**

What would be your guess? 75, 82, 69, 80, 77? To provide a consensus of these guesses, we can *guesstimate* the life expectancy to be 75 ± 5 years. 75 ± 5 years can be written as the following interval: (70, 80). This is referred to as an interval estimate.

In the real-world, researchers do not guess. They use sample data to create intervals and use the intervals to make inferences about the population. More importantly from a statistical standpoint, they create *confidence* intervals to make inferences about the population.
Common confidence levels are 90%, 95%, 98%, 99% and 99.9%. Although confidence levels can range between 0% and 100%, researchers prefer to be highly confident and rarely choose confidence levels that are less than 80%. Confidence levels are chosen based on the researchers’ preference and goal of the study.

The basic formation of a confidence interval is the following:

\[
\text{point estimate} \pm \text{error}
\]

This formation is always used for any confidence interval constructed.

Point estimates are derived from sample data. They are sample statistics. Points estimates are used to estimate their corresponding population parameter. For example, the sample mean is used to estimate the population mean (\(\mu\)), the sample standard deviation is used to estimate the population standard deviation (\(\sigma\)) and the sample proportion is used to estimate the population proportion (\(p\)).

<table>
<thead>
<tr>
<th>Point Estimate</th>
<th>Population Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\bar{x})</td>
<td>(\mu)</td>
</tr>
<tr>
<td>(s)</td>
<td>(\sigma)</td>
</tr>
<tr>
<td>(\hat{p})</td>
<td>(p)</td>
</tr>
</tbody>
</table>
Error, also known as margin of error, is the distance between the point estimate and the population parameter. Since researchers analyze samples to make inferences about the population, there will always be error because not every subject in the population is being analyzed. The formula for margin of error will vary based on the sampling distribution analyzed.

In this section, the sampling distribution of the sample mean will be the focal point. In conjunction with the Central Limit Theorem, the margin of error used to construct a confidence interval for the population mean with \( \sigma \) known is the following:

\[
E = z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right)
\]

By combining the point estimate with the margin of error, the following confidence interval is created:

**DEFINITION**

Confidence Interval for \( \mu \) with \( \sigma \) known:

\[
\bar{x} \pm z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right)
\]

This confidence interval will be referred to as the z-interval. The z-interval can only be constructed when the following condition is met:

- The population standard deviation is known.

Let’s explore an example.

In a recent effort to generate more revenue for city renovations, the city council decides to add a toll to a frequented intersection.
After a year, they randomly select 36 days and determine the mean revenue to be $126. Assume the population standard deviation is $16. Construct a 95% confidence interval for the population mean revenue the toll at the intersection will generate each day.

The point estimate (in this case the sample mean) will be used to make an inference about the population parameter (the population mean). To construct a confidence interval, identify the sample mean, the population standard deviation, the sample size and the $z$ critical value.

Note: $z_{\alpha/2}$ is often referred to as the $z$ critical value.

The $z$ critical value is based on the desired level of confidence. For this example, the confidence level is 95%. As seen below and based on the information discussed in Section 6.3: Normal Distributions,
the z critical value for a 95% level of confidence is ±2. Recall, the Empirical Rule provides approximations. Since researchers prefer to be precise, the Standard Normal Table should be used. Using the Standard Normal Table, the z critical value for a 95% level of confidence is ±1.96.

Since more than likely the common confidence levels will be used, below is a table of common confidence levels and the z critical values.

<table>
<thead>
<tr>
<th>Confidence Levels</th>
<th>( z_{\alpha/2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.9%</td>
<td>3.29</td>
</tr>
<tr>
<td>99%</td>
<td>2.575</td>
</tr>
<tr>
<td>98%</td>
<td>2.33</td>
</tr>
<tr>
<td>95%</td>
<td>1.96</td>
</tr>
<tr>
<td>90%</td>
<td>1.645</td>
</tr>
</tbody>
</table>

Using this information, the confidence interval results in the following:

\[
\bar{x} \pm z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right)
\]

\[
126 \pm 1.96 \left( \frac{16}{\sqrt{36}} \right)
\]

\[
126 \pm 1.96(2.667)
\]

\[
126 \pm 5
\]

\[
(121, 131)
\]
Note: When constructing confidence intervals, ensure to round off at the last step.

Thus, we are 95% confident that the population mean revenue the toll at the intersection will generate each day is between $121 and $131. *Note: $121 is considered the lower bound confidence limit and $131 is considered the upper bound confidence limit.*

On average, the city can conclude that this is the amount the toll at the intersection will generate each day. With that being said, there may be days where the amount is higher or lower than the average.

Be careful when making interpretations about a confidence interval. For example, being 95% confident means that 95% of the confidence intervals constructed will contain the true population parameter. It DOES NOT imply that there is a 95% chance the interval contains the population parameter. As a matter of fact, researchers don’t know if the population parameter is contained in the interval or not since every subject in the population was not analyzed. Researchers can only say with a certain level of confidence that the true population parameter lies in the interval. This concept also applies to other levels of confidence.

There are 4 key points to keep in mind when constructing confidence intervals.
I. **The width of the interval increases as the confidence level increases.** For example, suppose researchers wanted to construct a 99% confidence interval. It would result in the following:

\[
\bar{x} \pm z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right)
\]

126 ± 2.575 \left( \frac{16}{\sqrt{36}} \right)

126 ± 2.575(2.667)

126 ± 7

(119, 133)

The interval width widened, which makes sense because 99% of the area under the curve is much larger than 95% of the area under the curve.

II. **In contrast, the width of the interval becomes narrower (tighter) as the confidence level decreases.** For example, a 90% confidence interval would result in the following: (122, 130).

III. **Margin of error will decrease as n, the sample size, increases.** As mentioned earlier, point estimates are used to make inferences. By increasing the sample size, the point estimate will be close to the population parameter since almost all subjects in the population are being analyzed.
Thus, there will be less error between the point estimate and population parameter.

**IV.** *To be highly confident and have a narrower (tighter) interval, increase n.* The goal is to create a narrow interval! If the interval is narrower, there will be less error and the interval estimate will be a more precise reflection of the population parameter. For example, an interval of (12, 15) is more precise than an interval of (10, 20). As previously stated, decreasing the confidence level will result in a narrower interval. With that being said, a 99% confidence level is much better than a 90% confidence level. Therefore, to counter the effect, researchers must increase sample size. By increasing the sample size, researchers are able to zone in on the true value of the population parameter. This is an effect of the law of large numbers.

A Note about Assumptions: For all confidence intervals constructed (including those covered in upcoming sections and chapters), assume the population is normal and samples are randomly selected.
Section 7.2:

T-Interval

**Student Learning Outcomes**

By the end of the section,

1. You will understand the t-distribution.
2. You will construct a confidence interval for a population mean with $\sigma$ unknown.
In the previous section, a confidence interval for a population mean, \( \mu \), given \( \sigma \) was known was constructed. However, most real-world applications result in \( \sigma \) being unknown. When \( \sigma \) is unknown, researchers use the sample standard deviation, \( s \), as an estimate. Since \( \sigma \) is no longer being used, the \( z \)-interval should not be used. As an alternative, the \( t \)-distribution must be used.

**Definition**

\( t \)-distribution: a continuous probability distribution for a random variable, \( x \). The \( t \)-distribution has the following properties:

1. The mean, median and mode are equal to 0.
2. The curve is bell-shaped and symmetric about the mean.
3. The total area under the curve sums to 1.
4. The curve approaches but never touches the \( x \)-axis.
5. There are a family of curves called degrees of freedom (df). Degrees of freedom (df) = \( n – 1 \). The distribution changes as the sample size changes. Thus, there are different curves for the different degrees of freedom, that is, for the different sample sizes.
6. As the degrees of freedom increase, the \( t \)-distribution becomes very close to the standard normal distribution.

*Note:* To use the \( t \)-distribution, the distribution of the population must be normal.

**The \( t \)-distribution can be seen by viewing the table in the Appendix.**

While viewing the table, the first row shows the different levels of confidence, the second row shows the area in one tail and the third row shows the area in two tails. *(Area in one tail and area in two tails will be discussed in further detail in the next chapter.)* The first column shows the degrees of freedom (df).
Note: By comparison, the standard normal distribution is one curve and the t-distribution is a family of curves. Since the t-distribution becomes close to the standard normal distribution as the degrees of freedom approach infinity, the t-distribution can be used for all sample sizes.

Now that the t-distribution has been introduced, let’s examine the confidence interval for \( \mu \) with \( \sigma \) unknown.

### Definition

**Confidence Interval for \( \mu \) with \( \sigma \) unknown:**

\[
\bar{x} \pm t_{\alpha/2} \left( \frac{s}{\sqrt{n}} \right)
\]

*with degrees of freedom (df) = \( n - 1 \)*

This confidence interval will be referred to as the t-interval. The **t-interval** can only be constructed when one of the following conditions is met:

- The population is normal.
- The sample size is greater than or equal to 30.

Let’s explore an example.

A dean at a local university wanted to determine the average starting salary for recent graduates receiving an MBA from their School of Business. He randomly sampled 20 graduates and recorded their starting salaries. From the sample of 20 graduates, the mean was $95,093 and the standard deviation was $1724. Assume the sample was taken from a normal population. Construct a 99% confidence interval for the
population mean starting salary of all graduates receiving an MBA from the School of Business.

Once again, the point estimate (i.e. sample mean) will be used to make an inference about the population mean, \( \mu \). To construct the confidence interval, plug in the value for the sample mean as well as the values for the sample standard deviation, sample size and t critical value.

\textit{Note: } \( t_{\alpha/2} \) is often referred to as the t critical value.

The t critical value can be determined by referring to the t-distribution in the Appendix. Go to the row labeled “Confidence Level” and find the appropriate level of confidence being examined. Once the level of confidence is selected, use the “df” column to locate the degrees of freedom being examined. The number at the cross section of the column and row is the desired t critical value.

For this example, the level of confidence is 99%. The degrees of freedom = \( n - 1 = 20 - 1 = 19 \). So, the t critical value is 2.86093. Therefore, the confidence interval is the following:

\[
\bar{x} \pm t_{\alpha/2} \left( \frac{s}{\sqrt{n}} \right)
\]

\[
95093 \pm 2.86093 \left( \frac{1724}{\sqrt{20}} \right)
\]

\[
95093 \pm 2.86093(385.498)
\]

\[
95093 \pm 1103
\]

\[
(93990, 96196)
\]
Thus, we are 99% confident that the population mean starting salary of all graduates receiving an MBA from the School of Business is between $93,990 and $96,196. *Note:* $93,990 is considered the lower bound confidence limit and $96,196 is considered the upper bound confidence limit.
Section 7.3:

Proportion Z-Interval

Student Learning Outcomes

By the end of the section,

1. You will understand the properties of the sampling distribution of sample proportion.
2. You will compute point estimates for population proportions.
3. You will construct a confidence interval for a population proportion.
In the previous two sections, confidence intervals were constructed for the population mean. However, not all real-world situations can be analyzed using means. Population means and sample means are used to summarize quantitative data. In situations where the data is qualitative, the sample mean cannot be used. Thus, proportions are computed.

The term proportion is not an entirely new concept. A proportion is a percentage that can be defined as “a part of a whole”. It is computed by taking the number of successes divided by the sample size.

Inferences about population proportions can be made by using sample proportions as point estimates.

**Definition**

Sample Proportion:

\[
\hat{p} = \frac{x}{n}
\]

where \(x\) = number of success and \(n\) = sample size

*Note:* \(\hat{p}\) is read as “p-hat”

To make inferences about the population proportion, let’s first examine the sampling distribution of the sample proportion. The sampling distribution of the sample proportion can be examined in the same fashion as when the sampling distribution of the sample mean was examined. Refer to Section 6.4: Sampling Distributions of \(\bar{x}\) for a review. Therefore, the mean and standard deviation of the sampling distribution of the sample proportion is the following:
Chapter 7: Confidence Intervals

**Note**: The distribution can be approximated by a normal distribution if \( np \geq 5 \) and \( nq \geq 5 \), where \( q = 1 - p \).

Based on this information, the confidence interval for a population proportion can be constructed.

**Definition**

The Mean of the Sampling Distribution of the Sample Proportion:

\[
\mu_{\hat{p}} = p
\]

The Standard Deviation of the Sampling Distribution of the Sample Proportion (Standard Error of the Proportion):

\[
\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}
\]

**Definition**

Confidence Interval for \( p \):

\[
\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}
\]

This interval will be referred to as the proportion z-interval. The **proportion z-interval** can only be constructed when both of the following conditions are met:

- \( np \geq 5 \)
- \( nq \geq 5 \)

**Note**: \( \hat{q} \) is read as “q-hat”. In addition, \( \hat{q} = 1 - \hat{p} \)
Let’s explore an example.

In 2004, ACT Inc. reported that in a sample of 1650 randomly selected college freshman, 1221 of them returned to college the next year. Construct a 95% confidence interval for the population proportion of freshman who will return to college the next year.

\[
\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}
\]

\[
0.74 \pm 1.96 \sqrt{\frac{(0.74)(0.26)}{1650}}
\]

\[
0.74 \pm 1.96 \sqrt{\frac{(0.1924)}{1650}}
\]

\[
0.74 \pm 1.96 \sqrt{(0.0001166)}
\]

\[
0.74 \pm 1.96(0.0108)
\]

\[
0.74 \pm 0.021
\]

(0.719, 0.761)

Thus, we are 95% confident that the population proportion of freshman who will return to college the next year is between 0.719 and 0.761.

In many applications, the result is typically not stated as a confidence interval. Instead, for this example, the result would be stated as “74% with a margin of error of 2.1%”. It can also be stated as “74% plus or minus 2.1%”. The latter is more commonly used in polls.
Section 7.4:

Sample Size

Student Learning Outcomes

By the end of the section,

1. You will utilize the sample size formula for estimating a population mean.
2. You will utilize the sample size formula for estimating a population proportion.
One of the four key points mentioned in Section 7.1: Z-Interval stated that researchers can increase the sample size in order to be highly confident and have a narrower confidence interval. Increasing sample size reduces error. Thus, it raises the question, “How large a sample size is needed?” Due to time, money and resources, researchers may not have the luxury of sampling 20,000 subjects. Therefore, there are sample size formulas that can help researchers determine a minimum sample size needed to estimate a population parameter based on certain criteria. The criteria include a desired level of confidence and margin of error. These values are determined by the researcher. A researcher can select any level of confidence and any margin of error they deem appropriate based on the goal of the study.

Sample size formulas can be used to estimate population means and population proportions.

To determine the minimum sample size needed for estimating a population mean, the following formula is used:

**Sample Size Formula for Estimating µ:**

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

*Note:* In cases where the sample size yields a decimal, ALWAYS round UP to the next WHOLE number.

Let’s explore an example.

A newly married couple is looking to purchase their first home. How many homes should they sample to estimate the population mean selling price of a home to within $1000 with 95% confidence?
confidence? Assume the population standard deviation from a previous study was shown to be $3520.

Using this information, plug the values into the sample size formula for estimating a population mean.

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

\[
n = \left( \frac{(1.96)(3520)}{1000} \right)^2
\]

\[
n = \left( \frac{6899.2}{1000} \right)^2
\]

\[
n = (6.8992)^2
\]

\[
n = 47.6
\]

\[
n \approx 48
\]

The newly married couple should sample 48 homes. By using the sample of 48 homes, they will be able to estimate the population mean selling price of a home to within $1000 with 95% confidence.

To determine the minimum sample size needed to estimate a population proportion, the following formula is used:
Note: In cases where the sample size yields a decimal, ALWAYS round UP to the next WHOLE number.

Let’s explore an example.

During routine screenings, a doctor notices that 32% of her adult patients show higher than normal levels of blood pressure—a possible warning sign for heart disease. Upon hearing this, a team of medical researchers decide to conduct a study. They are hoping to estimate the population proportion of all adults having high blood pressure to within 4% with 95% confidence. How many randomly selected adults must they test?

Using this information, plug the values into the sample size formula for estimating a population proportion.

Note: Since the sample proportion from a previous study is known, the value can be used in the formula. If prior information is unknown, use 0.50 as a conservative estimate for the sample proportion. The conservative estimate will result in a slightly larger sample size.
\[ n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2 \]

\[ n = (0.32)(0.68)\left(\frac{1.96}{0.04}\right)^2 \]

\[ n = (0.32)(0.68)(49)^2 \]

\[ n = (0.32)(0.68)(2401) \]

\[ n = 522.5 \]

\[ n \approx 523 \]

The team of medical researchers should sample 523 adults. By using the sample of 523 adults, they will be able to estimate the population proportion of adults having high blood pressure to within 4% with 95% confidence.