Lesson 6: Confidence Intervals

Introduction

Learning objectives for this lesson

Upon completion of this lesson, you should be able to:

- Correctly interpret the meaning of confidence intervals
- Construct confidence intervals to estimate a population proportion
- Construct confidence intervals for estimate a population mean
- Calculate correct sample sizes for a study
- Recognize whether a given situation requires a proportion or means confidence interval

Toward Statistical Inference

Two designs for producing data are sampling and experimentation, both of which should employ randomization. As we have already learned, one important aspect of randomization is to control bias. Now we will see another positive. Because chance governs our selection (think of guessing whether a flip of a fair coin will produce a head or a tail) we can make use of probability laws – the scientific study of random behavior – to draw conclusions about an entire population from which the subjects originated. This is called **statistical inference**.

We previously defined a population and a sample. Now we will consider what we use to describe their values.

> **Parameter:** a number that describes the population. It is fixed but we rarely know it. Examples include the true proportion of all American adults who support the president, or the true mean of weight of all residents of New York City.
>
> **Statistic:** a number that describes the sample. This value is known since it is produced by our sample data, but can vary from sample to sample. For example, if we calculated the mean heights of a random sample of 1000 residents of New York City this mean most likely would vary from the mean calculated from another random sample of 1000 residents of New York City.

Examples

1. A survey is carried out at a university to estimate the proportion of undergraduate students who drive to campus to attend classes. One thousand students are randomly selected and asked whether they drive or not to campus to attend classes. The **population** is all of the undergraduates at that university campus. The **sample** is the group of 1000 undergraduate students surveyed. The **parameter** is the true proportion of all undergraduate students at that university campus who drive to campus to attend classes. The **statistic** is the proportion of the 1000 sampled undergraduates who drive to campus to attend classes.

2. A study is conducted to estimate the true mean yearly income of all adult residents of the state of California. The study randomly selects 2000 adult residents of California. The **population** consists of all adult residents of California. The **sample** is the group of 2000 California adult residents in the study. The **parameter** is the true mean yearly income of all adult residents of California. The **statistic** is the mean of the 2000 sampled adult California residents.

Ultimately we will measure statistics and use them to draw conclusions about unknown parameters. This is statistical inference.

APPLET

A "Begin" button will appear below when the applet is finished loading. This may take a minute or two depending on the speed of your internet connection and computer. Please be patient.

This applet simulates sampling from a population with a mean of 50 and a standard deviation of 10. For each sample, the 95% and 99% confidence intervals on the mean are computed based on the sample mean and sample standard deviation.

The intervals for the various samples are displayed by horizontal lines as shown below. The first two lines represent samples for which the 95% confidence interval contains the population mean of 50. The 95% confidence interval is orange and the 99% confidence interval is blue. In the third line, the 95% confidence interval does not contain the population mean; it is shown in red. In the seventh and last line shown below, the 99% interval does not contain the population mean; it is shown in white.

Constructing confidence intervals to estimate a population proportion

NOTE: the following interval calculations for the proportion confidence interval is dependent on the following assumptions being satisfied: np ≥ 10 and n(1-p) ≥ 10. If p is unknown then use the sample proportion.

The goal is to estimate p = proportion with a particular trait or opinion in a population.

- Sample statistic = $\hat{p}$ (read "p-hat") = proportion of observed sample with the trait or opinion we're studying.
- Standard error of $\hat{p} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where *n* = sample size.
- Multiplier comes from this table

Confidence Level	**Multiplier**
.90 (90%)	1.645 or 1.65
.95 (95%)	1.96, usually rounded to 2
.98 (98%)	2.33

.99 (99%)	2.58

The value of the **multiplier increases** as the **confidence level increases**. This leads to **wider** intervals for **higher confidence** levels. We are **more confident** of catching the **population** value when we use a **wider** interval.

Example

In the year 2001 Youth Risk Behavior survey done by the U.S. Centers for Disease Control, 747 out of *n* = 1168 female 12*th* graders said the always use a seatbelt when driving.
Goal: Estimate proportion always using seatbelt when driving in the population of all U.S. 12*th* grade female drivers. **Check assumption**: (1168)*(0.64) = 747 and (1168)*(0.36) = 421 both of which are at least 10.

Sample statistic is = $\hat{p}$ = 747 / 1168 = .64

Standard error = $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.64(1-.64)}{1168}} = .014$

A 95% confidence interval estimate is .64 ± 2 (.014), which is .612 to .668

With 95% confidence, we estimate that between .612 (61.2%) and .668 (66.8%) of all 12*th* grade female drivers always wear their seatbelt when driving.

Example Continued: For the seatbelt wearing example, a 99% confidence interval for the population proportion is

.64 ± 2.58 (.014), which is .64 ± .036, or .604 to .676.

With 99% confidence, we estimate that between .604 (60.4%) and .676 (67.6%) of all 12*th* grade female drivers always wear their seatbelt when driving.

Notice that the 99% confidence interval is slightly wider than the 95% confidence interval. IN the same situation, the greater the confidence level, the wider the interval.

Notice also, that the only the value of the multiplier differed in the calculations of the 95% and 98% intervals.

Using Confidence Intervals to Compare Groups

A somewhat informal method for comparing two or more populations is to compare confidence intervals for the value of a parameter. If the confidence intervals do not overlap, it is reasonable to conclude that the parameter value differs for the two populations.

Example

In the Youth Risk Behavior survey, 677 out of *n* = 1356 12*th* grade males said they always wear a seatbelt. To begin, we'll calculate a 95% confidence interval estimate of the population proportion. **Check assumption:** (1356)*(0.499) = 677 and (1356)*(0.501) = 679 both of which are at least 10.

Sample statistic is $\hat{p}$ = 677 / 1356 = .499

Standard error = $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.499(1-.499)}{1356}} = .0136$

A 95% confidence interval estimate, calculated as Sample statistic ± multiplier × Standard Error is

. 499 ± 2 (.0137), or .472 to .526.

With 95% confidence, we estimate that between .472 (47.2%) and .526 (52.6%) of all 12*th* male drivers always wear their seatbelt when driving.

Comparison and Conclusion: For females, the 95% confidence interval estimate of the percent always wearing a seatbelt was found to be 61.2% to 66.8%, an obviously different interval than for males. It’s reasonable to conclude that 12*th* grade males and females differ with regard to frequency of wearing a seatbelt when driving.

Using Confidence Intervals to "test" how parameter value compares to a specified value

Values in a confidence interval are "acceptable" possibilities for the true population value. Values not in the confidence interval are not acceptable (reasonable) possibilities for the population value.

Example

The 95% confidence interval estimate of percent of 12*th* grade females who always wear a seatbelt is 61.2% to 66.8%. Any percent in this interval is an acceptable guess at the population value.

This has the consequence that it’s safe to say that a majority (more than 50%) of this population always wears their seatbelt (because all values 50% and below can be rejected as possibilities.)

If somebody claimed that 75% of all 12*th* grade females always used a seatbelt, we should reject that assertion. The value 75% is not within our confidence interval.

Finding sample size for estimating a population proportion

When one begins a study to estimate a population parameter they typically have an idea as how confident they want to be in their results and within what degree of accuracy. This means they get started with a set level of confidence and margin of error. We can use these pieces to determine a minimum sample size needed to produce these results by using algebra to solve for n in our margin of error:

$$n = \frac{z^2\hat{p}(1-\hat{p})}{M^2}$$

where M is the margin of error.

Conservative estimate: If we have no preconceived idea of the sample proportion (e.g. previous presidential attitude surveys) then a conservative (i.e. guaranteeing the largest sample size calculation) is to use 0.5 for the sample proportion. For example, if we wanted to calculate a 95% confidence interval with a margin of error equal to 0.04, then a conservative sample size estimate would be:

$$n = \frac{(1.96^2)(0.5)(1-0.5)}{(0.04)^2} = 600.25$$

And since this is the *minimum* sample size and we cannot get 0.25 of a subject, we **round up**. This results in a sample size of 601.

Estimate when proportion value is hypothesized: If we have an idea of a proportion value, then we simply plug that value into the equation. Note that using 0.5 will always produce the largest sample size and this is why it is called a conservative estimate.

Constructing confidence intervals to estimate a population mean

Previously we considered confidence intervals for 1-proportion and our multiplier in our interval used a z-value. But what if our variable of interest is a quantitative variable (e.g. GPA, Age, Height) and we want to estimate the population mean? In such a situation proportion confidence intervals are not appropriate since our interest is in a **mean** amount and not a proportion.

Therefore we apply similar techniques but now we are interested in estimating the population mean, μ, by using the sample statistic $\bar{x}$ and the multiplier is a t-value. These t-values come from a t-distribution which is similar to the standard normal distribution from which the z-values came. The similarities are that the distribution is symmetrical and centered on 0. The difference is that when using a t-table we need to consider a new feature: **degrees of freedom** (**df**). This degree of freedom will be based on the sample size, n.

You may wonder why we are switching from "z" to "t" in our discussion. The reason for this switch is fairly straightforward: previously, we assumed that our random variable came from a normal distribution with a known population standard deviation, σ. However, typically we do not know this parameter and therefore must estimate it. This is done by using the standard deviation of the sample which is expressed as "**S**". Since we need to make this estimate we lose our reference to the variable being from a normal distribution, but instead apply a new distribution called "Student's t", or simply the t-distribution. As stated above, this t-distribution is quite similar to the normal distribution, and in fact as the sample size increases the t-distribution melds into the normal distribution. As a result, we will now use "s" instead of "σ" to represent the standard deviation.

This *t* approach is valid under the following two conditions:

1. The population from which the sample observations come is bell-shaped. Then for any sample size taken from such a population the application of the *t* method is valid.
2. The population from which the sample observations come is not bell-shaped, but a "large enough" size sample is taken and the distribution of this sample (think histogram!) does not display a large amount of skewness or outliers. In such an instance, the application of the *t* method is valid. Typcially "large enough" is a sample size of at least 30 --- This should sound familiar from the Central Limit Theorem.

History Lesson: You might be curious as to how such a strange name as Student's t came about. The t-distribution was discovered in the early 1900's by William Gossett. Gosset was a statistician employed by the Guinness brewing company which had stipulated that he not publish under his own name. He therefore wrote under the pen name "Student."

As we will see the interval calculations are identical just some notation differs. The reason for the similarity is that when we have paired data we can simply consider the differences to represent one set of data. So what is paired data?

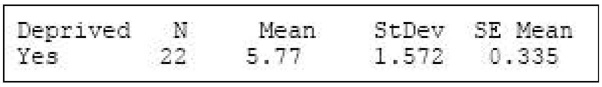
Estimating a Population Mean μ

- The sample statistic is the sample mean $\overline{x}$
- The standard error of the mean is $\frac{s}{\sqrt{n}}$ where s is the standard deviation of individual data values.
- The multiplier, denoted by t*, is found using the t-table in the appendix of the book. It's a simple table. There are columns for .90, .95,.98, and .99 confidence. Use the row for df = n − 1.

Thus the **formula for a confidence interval for the mean** is $\overline{x} \pm t^{*} \frac{s}{\sqrt{n}}$

For large n, say over 30, using t* = 2 gives an approximate 95% confidence interval.

Example 1: In a class survey, students are asked if they are sleep deprived or not and also are asked how much they sleep per night. Summary statistics for the n = 22 students who said they are sleep deprived are:



```
Deprived   N     Mean    StDev  SE Mean
Yes       22     5.77    1.572    0.335
```

- Thus n = 22, $\overline{x}$ = 5.77, s = 1.572, and standard error of the mean = $\frac{1.572}{\sqrt{22}}$ = 0.335
- A confidence interval for the mean amount of sleep per night is 5.77 ± t* (0.335) for the population that feels sleep deprived.
- Go to the t-table in the appendix of the book or T-table and use the df = 22 – 1 = 21 row. For 95% confidence the value of t* = 2.08.
- A 95% confidence interval for μ is 5.77 ± (2.08) (0.335), which is 5.77 ± 0.70, or 5.07 to 6.7
- Interpretation: With 95% confidence we estimate the **population mean** to be between 5.07 and 6.47 hours per night.

Example 1 Continued:

- For a 99% confidence interval we would look under .99 in the df = 21 in the T-table. This gives t* = 2.83.
- The 99% confidence interval is 5.77 ± (2.83) (0.335), which is 5.77 ± 0.95, or 4.82 to 6.72 hours per night.

Notice that the 99% confidence interval is wider than the 95% confidence interval. In the same situation, a higher confidence level gives a wider interval.

Finding sample size for estimating a population mean

Calculating sample size for estimating a population mean is similar to that for estimating a population proportion: we solve for n in our margin for error. However, since the t-distribution is not as "neat" as the standard normal distribution, the process can be iterative. This means that we would solve, reset, solve, reset, etc. until we reached a conclusion. Yet, we can avoid this iterative process if we employ an approximate method based on t-distribution approaching the standard normal distribution as the sample size increases. This approximate method invokes the following formula:

$$n = \frac{z^2 S^2}{M^2}$$

where S is a sample standard deviation possibly based on prior studies or knowledge.

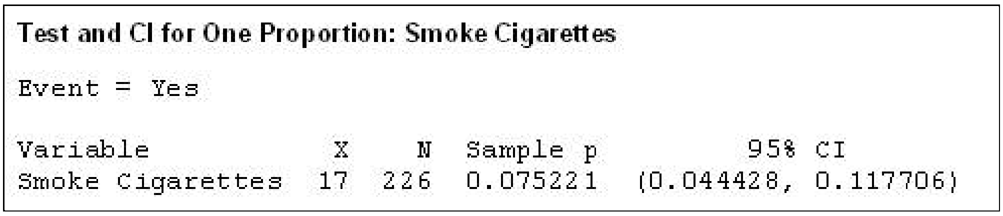
Using Minitab To Calculate Confidence Intervals

Consider again the Class Survey data set (Class_Survey.MTW) that consists of student responses to survey given last semester in a Stat200 course. If we consider this to be a random selection of students from the population of undergraduate students at the university, then we can use this data to estimate population parameters.

Estimating Population Proportion – Raw Data

1. Opening the Class Survey data set.
2. From the menu bar select Stat > Basic Statistics > 1 Proportion
3. In the text box Samples in Columns enter the variable Smoke Cigarettes
4. Click Options and edit the level of confidence (default value is 95%)
5. Click OK

The following is the 95% confidence interval for the true proportion of students who smoke cigarettes at the university.



```
Test and CI for One Proportion: Smoke Cigarettes

Event = Yes

Variable            X    N  Sample p          95% CI
Smoke Cigarettes   17  226  0.075221  (0.044428, 0.117706)
```

Estimating Population Proportion – Summarized Data

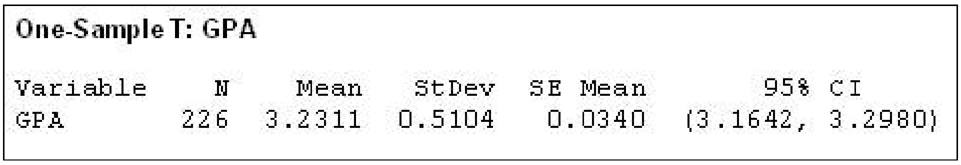
Summarized data simply means that you have the sample size and the total number of interest. For example, the summarized data from the above output would be 17 for the students who said "Yes" to smoking and the 226 that participated in the study. To use Minitab, complete the above steps except click the Summarize Data radio button and enter 226 for the number of trials and 17 for the number of events.

Special Note: Minitab calculates intervals based on the alphabetical order of the responses. If the answers are Yes and No, then the interval will be for the Event = Yes; if Male and Female the event of interest for Minitab will be Male. This is where the summarized data feature can help. If we wanted to get the proportion for those that said No for smoking, we could find this number by using the Stat > Tables > Tally Individual Variables.

Estimating Population Mean

Keeping with the Class Survey data set (Class_Survey.MTW), say we were interested in estimating the true undergraduate GPA at the university.

1. Opening the Class Survey data set.
2. From the menu bar select Stat > Basic Statistics > 1 Sample t
3. In the text box Samples in Columns enter the variable GPA
4. Click Options and edit the level of confidence (default value is 95%)
5. Click OK



```
One-Sample T: GPA

Variable    N    Mean   StDev  SE Mean       95% CI
GPA       226  3.2311  0.5104   0.0340  (3.1642, 3.2980)
```

Summarized data options for confidence intervals for a mean operate similarly to those described above for proportions.