Section 2D – Introduction to Confidence Intervals

Vocabulary

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data. For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Point Estimate: When someone takes a sample statistic and then claims that it is the population parameter.

Margin of Error: Total distance that a sample statistic might be from the population parameter. For normal sampling distributions and a 95% confidence interval, the margin of error is approximately twice as large as the standard error.

Standard Error: The standard deviation of a sampling distribution. The distance that typical sample statistics are from the center of the sampling distribution. Since the center of the sampling distributions is usually close to the population parameter, the standard error tells us how far typical sample statistics are from the population parameter.

Confidence Interval: Two numbers that we think a population parameter is in between.

95% Confident: 95% of confidence intervals contain the population value and 5% of confidence intervals do not contain the population value.

90% Confident: 90% of confidence intervals contain the population value and 10% of confidence intervals do not contain the population value.

99% Confident: 99% of confidence intervals contain the population value and 1% of confidence intervals do not contain the population value.

What is the population percentage of people worldwide that have congestive heart failure (CHF)? What is the population mean average salary of every working adult in Japan? Estimating population parameters is very important if we are to understand the world around us.

Estimating Population Parameters

There are two ways for finding a population parameter, an unbiased census or the center of a sampling distribution from thousands of large random samples. If you collect data from everyone in your population, and have not incorporated bias into the data, then you have collected an unbiased census. In that case, you know the entire population. Unbiased census data can be used to find population parameters like the population mean (μ), the population standard deviation (σ), or the population proportion (π). Simply calculate the mean, proportion, or standard deviation of the census and you know your population parameter.

We also learned that if you collect many large random samples from a population, you could create a sampling distribution. The center of the sampling distribution is usually a very good estimate of the population parameter.

This is not what happens usually in the real world. Populations may have millions of people, making it virtually impossible to take a census (unless you are the government). Most data scientists simply cannot collect a census from large populations. Random samples are usually very difficult to collect and can be expensive. Therefore, it is rare to see someone collect many random samples from the same population. Certainly not thousands of random samples. Therefore, we often cannot create a sampling distribution from the population either.

A person analyzing data usually has one large random sample. The question is can we estimate a population parameter with one large random sample?

Remember the principle of sampling variability.

Sampling Variability: Random sample statistics will usually be different from each other and different from the population parameter.

Every time we take a random sample, we will get something different. The sample statistic you calculate from random sample data will usually be off from the population parameter. Remember there will always be a margin of error.

Key: If all you have is one random sample, you will not be able to find the population parameter. The sample statistic you calculate will be off from the real population parameter.

If we have one random sample, can we estimate the population parameter at least? Yes, but we should be careful how we label it.

Point Estimates

Point Estimate: Some people take the random sample statistic and then just tell us in their article or report that the sample statistic is the population parameter.

Most of the time, when someone in an article gives us a population parameter, it usually is not the actual population parameter. It is a point estimate. They took some sample data, calculated the sample mean, and then tell us that the sample mean is the population mean. As you can imagine this creates a lot of confusion. Many people read articles and think the author knows the exact population mean or the exact population percentage, when in fact the number the author is quoting came from a sample. It is important to be aware of this. A good scientific report will usually make this distinction.

Good Point Estimate: "We tested a random sample of people for high cholesterol and found that 31.7% of the sample had high cholesterol. So we estimate that the population percentage of people worldwide with high cholesterol is about 31.7% with a 1.2% margin of error."

Bad Point Estimate: "The population percentage of people worldwide that have high cholesterol is 31.7%."

The second example shows what most articles say. It can be very confusing for most people since they believe that the author knows the population percentage for everyone worldwide. They do not realize it was just a sample percentage. We know from our study of sampling distributions and sampling variability that this sample percentage can be far off from the real population percentage.

Confidence Intervals

A sample statistic will usually be off from the population parameter. In other words, the sample statistic has a margin of error.

Margin of Error: The distance that a sample statistic might be from the population parameter.

It is relatively easy to calculate margin of error if already know the population parameter. Remember we rarely know the population parameter in the real world. It can be very difficult to estimate margin of error when you do not know the population parameter. Many mathematicians and statisticians put a lot of thought into finding formulas that would estimate the margin of error. We will go over some of these famous margin of error formulas throughout the chapter.

If you know the margin of error and the sampling distribution was relatively normal or symmetric, then you can use the margin of error to create a confidence interval.

Confidence Interval: Two numbers that we think a population parameter is in between.

When all you have is one random sample, you will not be able to find the population parameter exactly, but you can find two numbers that we think the population parameter may be in between. This is called a "confidence interval". For example, we may know what the population percentage is, but we think it is between 10.2% and 13.6%.

Here is a common formula for calculating a confidence interval.

Sample Statistic ± Margin of Error

Example 1: Suppose we look at a random sample of gas mileage (miles per gallon) for various cars. We want to estimate the population mean average mpg for all cars in the world. The sample mean $(\bar{x})$ was 24.761 mpg but remember this does not mean that the population mean is 24.761. Using a formula, we were able to calculate the margin of error for this sample to be 2.152 mpg. So what would the confidence interval be?

Sample Statistic ± Margin of Error

$\bar{x}$ ± Margin of Error

24.761 mpg ± 2.152 mpg

Lower Limit: 24.761 – 2.152 = 22.609 mpg

Upper Limit: 24.761 + 2.152 = 26.913 mpg

Therefore, a sample mean average gas mileage of 24.761 mpg tells us that the population mean average gas mileage for cars could be in between 22.609 mpg and 26.913 mpg.

Confidence Intervals can be written in three ways: interval notation, inequality notation, or just give the sample statistic and margin of error. In this example, here are the three ways the confidence interval may be written.

Interval Notation: (22.609 mpg , 26.913 mpg)
Most computer programs write their confidence intervals in interval notation. This does not mean (x , y) like in algebra. It means the population parameter could be any of the millions of numbers in between 22.609 mpg and 26.913 mpg.

Inequality Notation: 22.609 mpg < μ < 26.913 mpg
Remember this interval was trying to find two numbers that the population mean (μ) *is in between. That is exactly what this says.*

Sample Statistic and Margin of Error: 24.761 mpg (± 2.152 mpg error)
Many scientific journals or articles write it this way. They write the sample statistic as their point estimate with the margin of error.

Example 2: In the article earlier, we were looking for the percentage of people worldwide with high cholesterol. What would the confidence interval be for this problem?

"We tested a random sample of people for high cholesterol and found that 31.7% of the sample had high cholesterol. There was a 1.2% margin of error."

When calculating confidence intervals from a percentage, we usually convert the sample percentage into a sample proportion $(\hat{p})$. We should also convert the margin of error into a proportion.

31.7% = 0.317

1.2% = 0.012

Sample Statistic ± Margin of Error

$\hat{p}$ ± Margin of Error

0.317 ± 0.012

Lower Limit: 0.317 – 0.012 = 0.305

Upper Limit: 0.317 + 0.012 = 0.329

We can convert this proportion back into percentages if we wish. Notice that we can write the confidence interval in three ways again. Remember a population proportion can be written with the letter "p" or "π".

Interval Notation: (0.305 , 0.329) or (30.5% , 32.9%)
Inequality Notation: $0.305 < \pi < 0.329$ or $30.5\% < \pi < 32.9\%$
Sample Statistic and Margin of Error: 31.7% (± 1.2% error)

You should be comfortable converting percentages into proportions and proportions into percentages. Notice that when calculating the upper and lower limits we could have added and subtracted the percentages and got the same answer.

Lower Limit: 31.7% – 1.2% = 30.5%

Upper Limit: 31.7% + 1.2% = 32.9%

So a sample percentage of 31.7% does not tell us that the population percentage. It tells us that the population percentage could be in between 30.5% and 32.9%.

Important Note: Never add or subtract a proportion and a percentage. Yes, they are equivalent, but they are not the same. Either add and subtract the proportions, or add and subtract the percentages.

Never do this!! 11.9 ± 0.017

In the last two examples, how confident are we about these results?

Confidence Levels

When calculating confidence intervals, it is important to know what "confidence level" was used. A confidence level is not an abstract feeling about how confident you are. It is tied to the mathematical calculation of the margin of error. The most common confidence levels are 90%, 95% and 99%, with 95% being by far the most common. Whenever you ask a computer to calculate a confidence interval you must choose what level you want to use. Usually it is 95%.

Think of it this way. The more confident you are, the larger the margin of error and the wider you make the confidence interval. That way you are more likely to have the actual population parameter in between the two numbers. The less confident you are the smaller the margin of error and the narrower the confidence interval. I like to think of the confidence level as a catcher's mitt in baseball. If I want to be 90% confident that I catch the ball (catch the population parameter), I will use a regular sized catcher's mitt. If I want to be 95% confident I catch the ball, I will use a jumbo-sized catcher's mitt. If I want to be 99% confident that I catch the ball, I will use a huge soccer net.

90% confidence level → Small margin of error → Narrow confidence interval *(Regular sized Mitt)*

95% confidence level → Larger margin of error → Wider confidence interval *(Jumbo sized Mitt)*

99% confidence level → Extremely Large margin of error → Very wide confidence interval *(Soccer Net)*

Example: Earlier we looked at creating a confidence interval to estimate two numbers that we think the population mean average gas mileage (mpg) is in between. The following printout shows the calculation for 90%, 95% and 99% confidence levels. Notice that as the confidence level increases, the margin of errors are increasing the numbers in the confidence intervals are getting farther apart.

Confidence Interval - One population mean: confidence level = 0.9
Input: C1 MPG
σ unknown

Var	N	Mean	Stdev	Margin of Error	90.0%CI
C1 MPG	38.0	24.761	6.547	1.792	(22.9686, 26.5524)

Confidence Interval - One population mean: confidence level = 0.95
Input: C1 MPG
σ unknown

Var	N	Mean	Stdev	Margin of Error	95.0%CI
C1 MPG	38.0	24.761	6.547	2.152	(22.6085, 26.9126)

Confidence Interval - One population mean: confidence level = 0.99
Input: C1 MPG
σ unknown

Var	N	Mean	Stdev	Margin of Error	99.0%CI
C1 MPG	38.0	24.761	6.547	2.884	(21.8765, 27.6446)

Confidence Interval Sentence

Computers can calculate confidence intervals. The job of a data analyst, data scientist or statistician is to explain. In other words, the sentences are very important. Whenever we write a sentence to explain a confidence interval, we should always state the confidence level that was used. For one-population confidence intervals, we should also give the two numbers that the population parameter is in between.

One Population Confidence Interval Sentence:

> *"We are (90%, 95% or 99%) confident that the population parameter is in between # and #".*

Here is the sentence for the 90% confidence interval estimate of the population mean average mpg. Remember in quantitative data, you can round the answers to one more decimal point than is present in the original data. (In this case, since the original sample data mpg values were rounded to the tenths place, we can round the confidence intervals to the hundredths place.) If you do not know the accuracy of your data, it is better not to round the numbers.

We are 90% confident that population mean average gas mileage for all cars is between 22.97 mpg and 26.55 mpg.

Here is the sentence for the 95% confidence interval estimate of the population mean average mpg. We rounded the Statcato answers to the hundredths place.

We are 95% confident that population mean average gas mileage for all cars is between 22.61 mpg and 27.64 mpg.

Here is the sentence for the 99% (rounded) confidence interval estimate of the population mean average mpg.

We are 99% confident that population mean average gas mileage for all cars is between 21.88 mpg and 26.91 mpg.

Here is the sentence for the genetic trait population percentage confidence interval. We will assume it was a 95% confidence level.

We are 95% confident that population percentage of all people worldwide that have high cholesterol is between 30.5% and 32.9%.

Understanding Confidence Levels

Here are the definitions of confidence. Notice that these definitions are not talking about a "feeling" about confidence. They are also not talking about being sure that the population parameter is in between the exact two numbers in a confidence interval. So what do these mean?

95% Confident: 95% of confidence intervals contain the population value and 5% of confidence intervals do not contain the population value.

90% Confident: 90% of confidence intervals contain the population value and 10% of confidence intervals do not contain the population value.

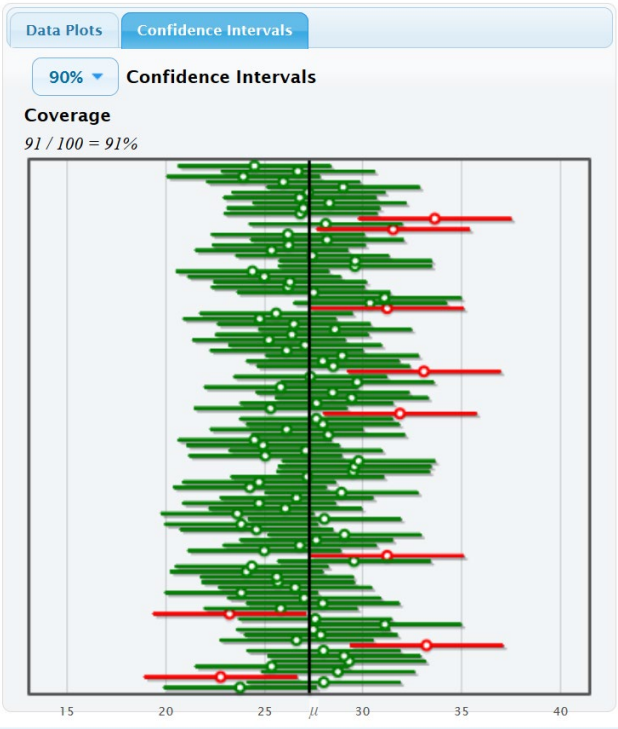
99% Confident: 99% of confidence intervals contain the population value and 1% of confidence intervals do not contain the population value.

These definitions are talking about many samples, many confidence intervals. In essence, a sampling distribution.

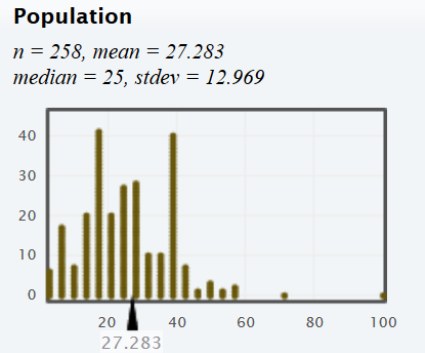
Example 1: 90% confidence level sampling distribution

Work Hours per Week for working COC Statistics Students (Fall 2015 semester)

In a previous section, we created a sampling distribution of sample means for the work hours of statistics students. We did not use students that said they work "zero" hours. To understand confidence levels, we are going to take it a step further. Instead of just taking many random samples and calculating many sample means, we are going to use StatKey to calculate many confidence intervals. All of the confidence intervals will have a 90% confidence level. We used a sample size of 30 this time.



Let us see if we understand what we are looking at. The dark line indicates the population mean of 27.283 hours of work per week. If the population mean is in between the two numbers in the confidence level, then the confidence interval is green. This indicates that the confidence interval contains the population parameter. If the population mean is not in between the two numbers in the confidence level, then the confidence interval is red. This indicates that the confidence interval does not contain the population parameter.



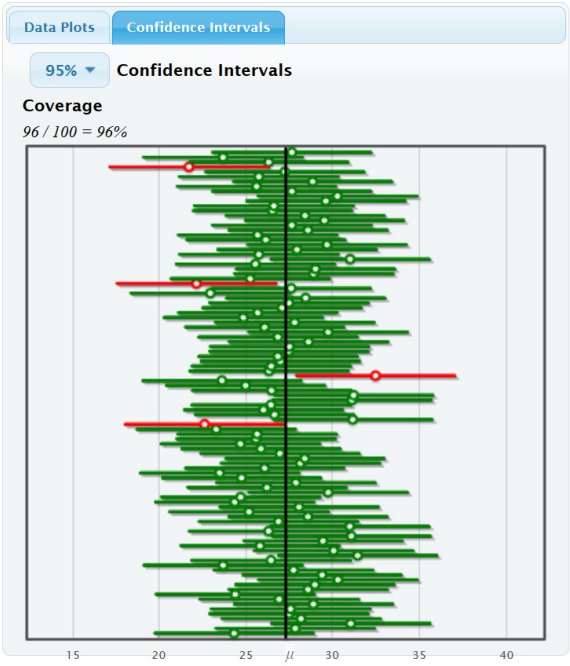
Notice when we use a 90% confidence level, about 90% of them were green (contained the population parameter) and about 10% of them were red (did not contain the population parameter). In other words, not all confidence intervals contain the population parameter! This is what the definition of 90% confidence is talking about. If we take many random samples, and create many confidence intervals, about 90% of the confidence intervals will have the population parameter in between the two numbers and 10% of them will not.

Notice that the green and red lines describing the confidence interval have a lot of variability. This is sampling variability at work. Random samples will always be different. That means that the confidence interval numbers will also be different for every random sample.

Also, notice that the number of green confidence intervals was not exactly 90%. In fact, it was 91% for the first hundred samples. 90% is a limit. This means that because of sampling variability, the exact percentage of green confidence intervals will fluctuate. As the number of samples increase, the number usually gets closer and closer to 90%.

Example 2: 95% confidence level sampling distribution

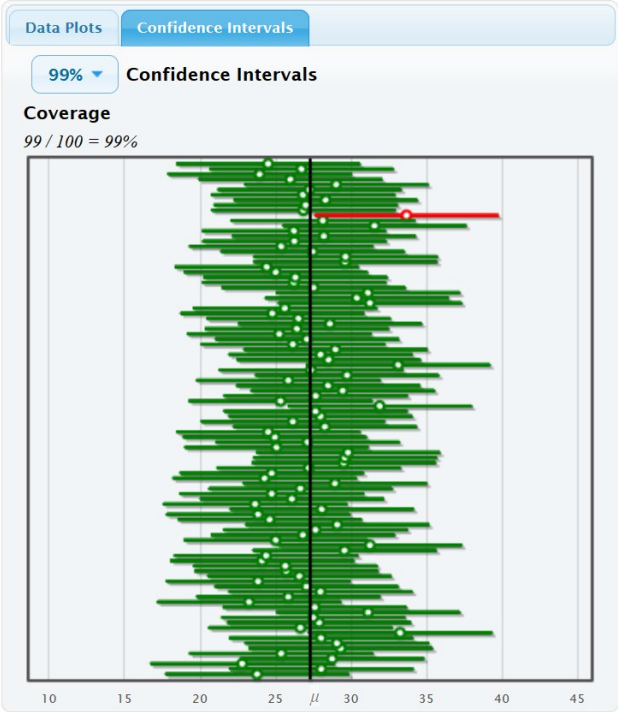
Work Hours per Week for working COC Statistics Students (Fall 2015 semester)



Now we will set the confidence levels to 95%. We calculated many confidence intervals and all of them have a 95% confidence level. Notice that the percentage of green confidence intervals that contain the population mean average is now approaching 95%. It is actually 96% for these first 100 samples, but as the number of samples increase, the percentage will get closer to 95%. Also, notice that the percentage of red confidence intervals that do not contain the population mean average is now approaching 5%. It is actually 4% for these first 100 samples, but as the number of samples increase, the percentage will get closer to 5%. This again is what the definition of 95% confidence is talking about. If we create many 95% confidence intervals, about 95% of them will be green and contain the population parameter, and about 5% of them will be red and not contain the population parameter.

Example 3: 99% confidence level sampling distribution

Work Hours per Week for working COC Statistics Students (Fall 2015 semester)



If we set the confidence levels to 99%, we see that the percentage of green confidence intervals that contain the population mean average is now approaching 99% and the percentage of red confidence intervals that do not contain the population mean average is now approaching 1%. This again is what the definition of 99% confidence is talking about. If we create many 99% confidence intervals, about 99% of them will be green and contain the population parameter, and about 1% of them will be red and not contain the population parameter.

Finding the sample statistic and margin of error from a confidence interval

Occasionally you may have an article or scientific report that gives a confidence interval to estimate a population mean or a population proportion, yet neglects to tell you the margin of error or the sample statistic. If you have a bootstrap distribution that looks relatively normal, you will know the confidence interval, but may not know the margin of error. Some computer programs will tell you the upper and lower limit of the confidence interval but not tell you the margin of error. In these situations, there is a way to figure out the sample statistic and the margin of error. Remember, these formulas are only used when you know the upper and lower limit of the confidence interval and you have a normal sampling distribution.

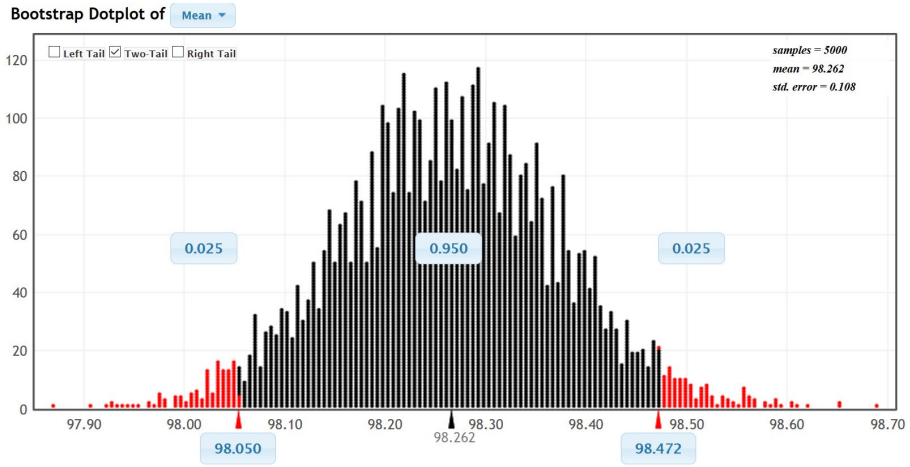
Confidence Interval Back-Solving Formula for Sample Statistic: $Sample\ Statistic = \frac{(Upper\ Limit + Lower\ Limit)}{2}$

Confidence Interval Back-Solving Formula for Margin of Error: $Margin\ of\ Error = \frac{(Upper\ Limit - Lower\ Limit)}{2}$

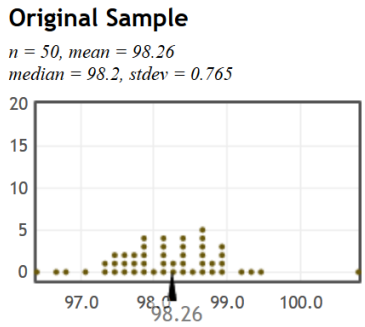
Example 1: Bootstrap Confidence Interval for Population Mean Average Body Temperature

Bootstrapping is a technique for calculating confidence intervals. The following bootstrap confidence interval was calculated from a random sample of 50 adult body temperatures in degrees Fahrenheit. The upper and lower limits for the confidence interval are given at the bottom right and left of the bootstrap distribution. The confidence level is given in the middle of the bootstrap distribution. So the 95% confidence interval is $(98.050\ °\text{F},\ 98.472\ °\text{F})$.

Confidence Interval Sentence: We are 95% confident that the population mean average body temperature of human adults is between 98.050°F and 98.472°F.



StatKey did tell us that the sample mean was 98.26°F, but notice that we do not know the margin of error. This is a perfect time to use the back-solving formula for margin of error.



$$Margin\ of\ Error = \frac{(Upper\ Limit - Lower\ Limit)}{2} = \frac{(98.472 - 98.050)}{2} = \frac{(0.422)}{2} = 0.211\ °\text{F}$$

Let us check the sample statistic formula and see how close it is to the actual sample mean.

$$\text{Sample Statistic (mean)} = \frac{(Upper\ Limit + Lower\ Limit)}{2} = \frac{(98.472 + 98.050)}{2} = \frac{(196.522)}{2} = 98.261\ °\text{F}$$

Notice the sample statistic is very close to the actual sample mean of 98.26 °F.

Example 2: An article claims that the population percentage of young adults ages 18-25 years in the U.S. that have depression is in between 9.59% and 12.27%. This is a confidence interval. We will assume they used a 95% confidence level. What was the sample proportion and the margin of error? Again, this would be a good time to use the back-solving formulas. Remember to either use the proportions or the percentages but do not add or subtract a proportion and a percentage.

$Margin\ of\ Error = \frac{(Upper\ Limit - Lower\ Limit)}{2} = \frac{(0.1227 - 0.0959)}{2} = \frac{(0.0268)}{2} = 0.0134$ (or 1.34%)

$Margin\ of\ Error = \frac{(Upper\ Limit - Lower\ Limit)}{2} = \frac{(12.27\% - 9.59\%)}{2} = \frac{(2.68\%)}{2} = 1.34\%$ (or 0.0134 as a proportion)

Sample Statistic (proportion) $= \frac{(Upper\ Limit + Lower\ Limit)}{2} = \frac{(0.1227 + 0.0959)}{2} = \frac{(0.2186)}{2} = 0.1093$ (or 10.93%)

Sample Statistic (percentage) $= \frac{(Upper\ Limit + Lower\ Limit)}{2} = \frac{(12.27\% + 9.59\%)}{2} = \frac{(21.86\%)}{2} = 10.93\%$ (or 0.1093)

Summary of Confidence Intervals

- Be aware of point estimates. When a person claims to know the exact population parameter, they probably just calculated a sample statistic and are telling you it is the population parameter. We only know the population parameter if we have collected a census or if we have collected many, many random samples and look for the center of the sampling distribution. We can never know the exact population parameter from a large population if all we have is one random sample. If we have one random sample, all we can do is estimate the population parameter with a confidence interval.

- A confidence interval is two numbers that we think a population parameter is in between.

- Remember, a confidence interval should never be calculated from a census. If you already know the population parameter, there is no need to estimate it with a confidence interval. Confidence interval are calculated when we only have random sample data and need to estimate the population parameter.

- It is important to know what confidence levels were used. 90%, 95% and 99% are all sometimes used, though 95% is the most common. Remember, these levels do not refer to a feeling of confidence about one confidence interval. They are part of the confidence interval calculation and refer to the process of calculating thousands of confidence intervals.

- Here are definitions of 90%, 95%, and 99% confidence. These definitions imply that not all confidence intervals contain the population parameter. Sometimes the population parameter will not be in between the two numbers in the confidence interval.

  90% Confident: 90% of confidence intervals contain the population value and 10% of confidence intervals do not contain the population value.

  95% Confident: 95% of confidence intervals contain the population value and 5% of confidence intervals do not contain the population value.

  90% Confident: 99% of confidence intervals contain the population value and 1% of confidence intervals do not contain the population value.

- The margin of error is how far we think the sample statistic is from the population parameter. A common formula that is sometimes used to calculate a confidence interval is the sample statistic ± margin of error.

- Be able to explain the confidence interval. Here is a common sentence used for one-population confidence intervals: We are (*90%, 95% or 99%*) confident that the population parameter (*mean, proportion, median, standard deviation, or variance*) is in between # and #.

- If you know the upper and lower limit of a confidence interval from a normal sampling distribution, you can use these back solving formulas to find the sample statistic and the margin of error.

$$Sample\ Statistic = \frac{(Upper\ Limit + Lower\ Limit)}{2}$$

$$Margin\ of\ Error = \frac{(Upper\ Limit - Lower\ Limit)}{2}$$
