Chapter 2: Estimating Population Parameters

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.

Sampling Distribution: Take many random samples from a population, calculate a sample statistic like a mean or percent from each sample and graph all of the sample statistics on the same graph. The center of the sampling distribution is a good estimate of the population parameter.

Sampling Variability: Random samples values and sample statistics are usually different from each other and usually different from the population parameter.

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. Can be calculated by either a bootstrap distribution or by adding and subtracting the sample statistic and the margin of error.

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.

Bootstrapping: Taking many random samples values from one original real random sample with replacement.

Bootstrap Sample: A simulated sample created by taking many random samples values from one original real random sample with replacement.

Bootstrap Statistic: A statistic calculated from a bootstrap sample.

Bootstrap Distribution: Putting many bootstrap statistics on the same graph in order to simulate the sampling variability in a population, calculate standard error, and create a confidence interval. The center of the bootstrap distribution is the original real sample statistic.

Introduction: The goal of learning Statistics or Data Science is to be able to analyze data to learn about populations in the world around us. The best way to understand a population is collect and analyze unbiased data from that population, namely a census. The trouble is we rarely have an unbiased census. It is sometimes impossible to collect data from everyone in a population. We have to rely on samples, small subgroups of the population. The next few chapters deal with the subject of using samples to understand populations. This is sometimes called "inferential statistics". We will start by trying to distinguishing between population parameters from sample statistics.

Section 2A – Statistics and Parameters

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.

Bias: When data does not represent the population.

The goal of collecting and analyzing data is to understand the world around us. To this end, our goal is understand populations. The population is all of the people or objects you plan to study. A population can be large (like all people living in Brazil) or small (like all students in a particular statistics class). It goes without saying that the larger the population the more difficult it is to understand.

The best data for representing populations is an unbiased census. A census is an attempt to collect data from everyone in a population. A census is easier if we have a small population like the people in a particular statistics class. The advantage of collecting an unbiased census is that we can calculate population values (parameters) directly with reasonable certainty. Governments may sometimes attempt to do a census and collect data on all of the people living in a particular country. It should be noted that though they attempt to get data on everyone, they rarely succeed. There will always be some people fall through the cracks and are not represented in the census. An unbiased census of a large population still represents a high percentage of the people, so is generally better than a small sample of people.

A data scientist rarely has the ability to collect a census unless the population is relatively small. People that work in statistics and data science usually rely on collecting samples. Remember a sample is a small subgroup of the population. It is usually less than 10% of the population and is often significantly less than 10%. If the sample is unbiased, we then try to analyze the sample data and make guesses as to what is happening at the population level. Therefore, a data scientist or statistician needs to be able to use sample values (statistics) to figure out approximate population value (parameters).

Statistic: A number calculated from sample data in order to understand the characteristics of the data.

Parameter: A population value. It can be calculated from an unbiased census, but is often just a guess about what someone thinks the population value might be.

It is very important to note that statistics and parameters are not the same thing. A statistic calculated from 250 people in a sample will often be very different from the actual population parameter from millions of people. The question that is important to ask is how far off is the sample statistic from the population parameter? That is sometimes called "margin of error" and is a key topic in this chapter.

Common Statistics

$\bar{x}$: ("x-bar") Sample mean average

s : Sample standard deviation (typical distance from the sample mean)

s^2: Sample variance (sample standard deviation squared)

$\hat{p}$: ("p-hat") Sample proportion (sample percentage)

n : Sample size or frequency (number of people or objects in the sample)

r : Sample correlation coefficient (measures quantitative relationships between samples)

b_1 : Sample slope (The slope of a regression line calculated from sample data.)

b_0 : Sample Y-intercept (The Y-intercept of a regression line calculated from sample data.)

Common Parameters

μ : ("mu") Population mean average

σ : ("sigma") Population standard deviation (typical distance from the population mean)

σ^2: Population variance (population standard deviation squared)

π: ("pi") Population proportion (population percentage) *(Some people use "p" for population proportion.)*

N : Population size or frequency (number of people or objects in the population)

ρ : ("rho") Population correlation coefficient (measures quantitative relationships between populations. Note this is not a "p". It is the Greek letter "rho".)

β_1 : Population slope (The slope of the population regression line. Used when studying quantitative relationships between populations.)

β_0 : Population Y-intercept (The Y-intercept of the population regression line. Used when studying quantitative relationships between populations.)

Let us look at some examples of using statistics and parameters. It is important to be able to identify if a number used is a statistic or a parameter and what letter we might use in the computer program.

Example

"We think the mean average ACT score for all high school students is about 22. The mean average ACT score for a random sample of 85 high school students was 21.493"

μ = 22 (parameter)
n = 85 (statistic)
$\bar{x}$ = 21.493 (statistic)

Example

"A random sample showed that 13.2% of adults were infected, but this indicates that the population percentage could be 17%". *(Note: Computer programs often require you to convert the percentages into decimal proportions.)*

$\hat{p}$ = 0.132 (statistic)
π = 0.17 (parameter)

Example

The standard deviation for the heights of all women is thought to be about 2.5 inches. A random sample of women heights had a standard deviation of 2.618 inches.

σ = 2.5 (parameter)
s = 2.618 (statistic)

Example

"Sample data indicated that the correlation coefficient was 0.239 and the slope was 47.3 dollars per pound. Let's compare these to the population claims that the correlation coefficient is zero and the slope is about 50 dollars per pound."

r = 0.239 (statistic)
b_1 = 47.3 (statistic)
ρ = 0 (parameter)
β_1 = 50 (parameter)

Problem Set Section 2A

1. Describe each of the following symbols. What does the symbol represent? Is the symbol describing a sample statistic or a population parameter?

N , n , π , $\hat{p}$, μ , $\bar{x}$, σ , s , σ^2, s^2, ρ , r , β_1 , b_1

(#2-25) Directions: Determine if the numbers in the following clips from magazines and newspapers are describing a population parameter or a sample statistic. In each case, give the symbol we would use for the parameter or statistic. (N , n , π , $\hat{p}$, μ , $\bar{x}$, σ , s , σ^2, s^2, ρ , r , β_1 , b_1)

2. "Our study found that of the 200 people tested in the sample, only 3% showed side effects to the medication."

3. "It has been speculated for years that the mean average height of all men is 69.2 inches, but our sample data disagrees with this. Our sample mean average was 69.5 inches."

4. "The standard deviation for all humans is about 1.8 degrees Fahrenheit. A random sample of 52 people found a standard deviation of 1.739 degrees Fahrenheit".

5. "We tested a sample of 300 incoming college freshman and found that their mean average IQ was 101.9 with a standard deviation of 14.8".

6. "The mean average human body temperature has long been thought to be 98.6 degrees Fahrenheit, but our sample of 63 randomly selected adults had a mean average was 98.08".

7. "The mean average number of units that students take per semester is about 12, but when we took a random sample of 160 college students found that the mean average was 12.37 units."

8. "A public opinion poll showed that 47.2% of voters would vote for the candidate, but when the votes or entire population were counted we found that only 41.3% voted for the candidate."

9. "According to the California Department of Finance, the Los Angeles county population as of January 2015 was approximately 10,136,559 people."

10. "We want to check and see if the population correlation coefficient could be zero. The sample correlation coefficient was 0.338."

11. “Many experts think that the population slope for weight gain in these type of bears is about 3 pounds per month, but the sample slope from 54 bears was 2.7055 pounds per month.”

12. “A random sample of 40 men found that the sample variance for systolic blood pressure was 109.474, but this indicates that the population variance could be as high as 173.”

13. “According to the 2015 U.S. census, approximately 78% of U.S. households own a computer. A random sample of 165 households found that 81.2% of them owned a computer.”

14. “We think that the population correlation coefficient is zero. The sample correlation coefficient was 0.0371.”

15. “IQ tests are supposed to have a population mean of 100 and a population standard deviation of 15 IQ points. This could be correct since our random sample data had a mean of 97.7 and the standard deviation of 15.3 IQ points.”

16. “When analyzing the relationship between the amount of mercury and the pH of Florida lakes, we found a sample slope of −0.152. We are wondering if the population slope could be zero.”

17. “We believe that the population mean average pH of Florida lakes is approximately 6.7, yet our sample data from 53 randomly selected lakes had a mean of 6.591.”

18. “While the sample variance is 37.882, we think the population variance could be as high as 50.”

19. “We believe there are approximately 59,530 people currently living in Canyon Country, CA.”

20. “A random sample of 60 adults found that 21.7% of them had this characteristic. However, we think the population percentage is probably closer to 15%.”

21. “The mean average weight of the 10 male lions was 437.2 pounds. Most people believe that the mean average weight of all male lions is closer to 420 pounds.”

22. “The correlation coefficient for the ordered pair sample data was 0.922. This seems very significant, but does this indicate that the population correlation coefficient is 1?”

23. “We analyzed the gas usage and distance for large 18-wheeler trucks and found the sample slope to be 6.23 miles per gallon. Articles online indicate that the population slope for all 18-wheeler trucks is closer to 5.9 miles per gallon.”

24. “The sample standard deviation was approximately $3.78. We want to see if the population standard deviation could be $3.50.”

25. “A random sample of 38 cars, found that the mean average displacement was 177.289 and the standard deviation was 88.877.”

Section 2B – Sampling Variability and Sampling Distributions

If you wanted to study baseball players, would you only study one baseball player? If you wanted to study bears, would you only study one bear? The answer of course is no. When studying a topic like bears or baseball players, we should look at many different bears, many different baseball players. The problem with studying samples is that we usually only collect one sample at a time. We cannot learn about the behavior and variability in samples if we only look at one sample. We need to look at hundreds or even thousands of samples.

Sampling Distributions

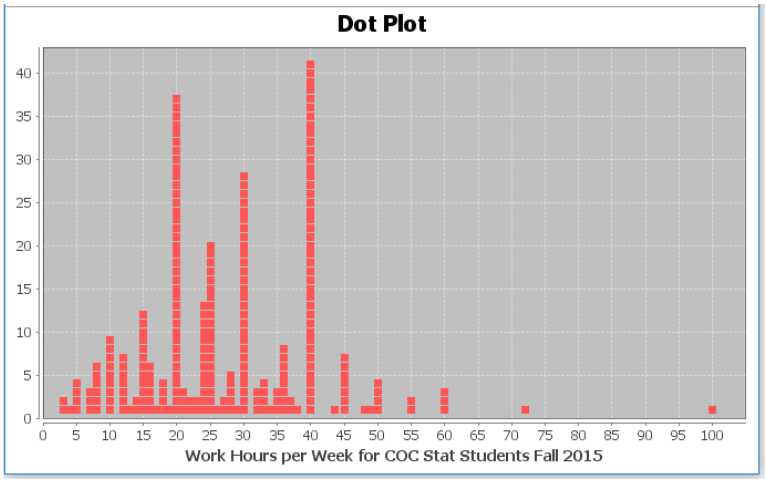
Suppose we take many, many random samples from a population. From each random sample, we calculate a statistic like the sample mean average. If we put all of those sample means on the same graph, we have created a "sampling distribution". Sampling distributions are one of the best ways to understand random samples and sampling variability.

In the real world, a data scientist has only one random sample and may have no idea what a population parameter is. In this example, we will be creating a sampling distribution by take random samples from a census. We will assume the census is unbiased. With an unbiased census, we will know what the population parameter is. That way we can compare our sample statistics to the parameter and study the variability.

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

We will start by looking at a census of the work hours of all of the working Math 140 students in the fall 2015 semester. It should be noted that we are only studying the statistics students that said they work in addition to going to school. We removed all of the students that work zero hours. We will take many random samples of size 50 from this census data and create a sampling distribution for various statistics.

Census Data (Work Hours per Week for working COC Stat Students Fall 2015)



Population Parameters

Variable	Mean	Standard Deviation
Work Hours per Week COC Stat Students	27.283	12.969

Variable	Median
Work Hours per Week COC Stat Students	25.0

We see that the census data is skewed right with a population mean average of 27.283 hours per week, a population standard deviation of 12.969 hours per week, and a population median of 25 hours per week. We will assume that the census was unbiased and these are parameters.

Population mean = 27.283 hours per week
Population standard deviation = 12.969 hours per week
Population median = 25 hours per week

We learned in chapter 1 that random samples tend to minimize sampling bias, so are better representations of the populations than other samples that are not random. Does this mean that random samples are perfect representations of the population? Let us see.

Sample 1: Here is one random sample of 50 statistics students from the work hours census data.

Descriptive Statistics

Variable	Mean	Standard Deviation
work hours random sample1	26.93	11.266

Variable	Median
work hours random sample1	24.0

Variable	Sample Size
work hours random sample1	50

We see that the sample mean was 26.93 hours per week, the sample standard deviation was 11.266 hours per week, and the sample median was 24 hours per week. Notice that all of these sample statistics are different from the population parameters.

Sample 1 mean = 26.93 hours per week
Population mean = 27.283 hours per week

Sample 1 standard deviation = 11.266 hours per week
Population standard deviation = 12.969 hours per week

Sample 1 median = 24 hours per week
Population median = 25 hours per week

Sample 2: Let us take another random sample of 50 statistics students work hours from the population.

Descriptive Statistics

Variable	Mean	Standard Deviation
work hours random sample2	29.5	12.732

Variable	Median
work hours random sample2	30.0

Variable	Sample Size
work hours random sample2	50

We see that the sample mean was 29.5 hours per week, the sample standard deviation was 12.732 hours per week, and the sample median was 30 hours per week. Notice these sample statistics are also different from the population parameters. They are also different from the last random sample.

Sample 2 mean = 29.5 hours per week
Sample 1 mean = 26.93 hours per week
Population mean = 27.283 hours per week

Sample 2 standard deviation = 12.732 hours per week
Sample 1 standard deviation = 11.266 hours per week
Population standard deviation = 12.969 hours per week

Sample 2 median = 30 hours per week
Sample 1 median = 24 hours per week
Population median = 25 hours per week

These examples show us that random sample statistics will usually be different from the population parameters. Random sample statistics will also be different from each other. Every time we take another random sample from the same population, we will get different values. This is the principle of "sampling variability" and is a major roadblock on the quest to estimating population parameters.

Sampling Variability: Random samples values and sample statistics are usually different from each other and usually different from the population parameter.

Let us continue taking random samples from the population of working statistics students in fall 2015. Every time we take a random sample, we keep getting different values and different statistics. Hardly any of the samples are close to the population parameter. In this example, we will focus on the mean. Remember the population mean average was 27.283 hours per week. No matter how many random samples we take, the sample means are usually different from the population mean of 27.283 hours per week. Every sample has a "margin of error".

Margin of Error: How far off a sample statistic can be from the population parameter.

In the first random sample, the sample mean was 26.93 hours per week. So the sample mean of 26.93 hours per week was 0.353 hours lower than the population mean of 27.283 hours per week. This is the margin of error.

In the second random sample, the sample mean was 29.5 hours per week. So the sample mean of 29.5 hours per week was 2.217 hours higher than the population mean of 27.283 hours per week. Again, that is the margin of error for that sample.

What does this tell us?

The principle of sampling variability tells us that sample statistics will usually be off from the population parameter. In other words, almost all samples have a margin of error. Sometimes random samples are closer to the population parameter like sample 1 and sometimes the random samples are farther away like sample 2.

Important Note: If you know the population parameter, then it is relatively easy to calculate the margin of error (sample statistic – population parameter). Most of the time, we are working with sample data, so have no idea what the population parameter is. In that case, it is much more difficult to figure out the potential margin of error. Formulas were developed in order to estimate what the margin of error could be.

Point Estimates

People are usually very interested to know population values. However, we rarely ever know the population parameter. In the real world, we usually only have one random sample. Sometimes, a person will simply tell you that the sample statistic is the population parameter. This is called a "point estimate" and tends to create a lot of confusion for people.

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

In an article published by a health website, the author states that the population average weight of all men in America is 196 pounds. As with most articles, this is a guess about the population average and is not the actual population average weight of men. We call this a "point estimate". Someone took a sample of men and weighed them. We do not know the sample size or if the sample was even random. They calculated the sample average and found it to be 196 pounds. Since no one really knows the population average weight of all men in the U.S., the author simply tells us the sample average is the population average.

Think about the principle of sampling variability that we just learned. We said that a sample statistic usually has a margin of error is off from the population parameter. Yet people reading the article believe that the population average weight of all men in the U.S. is exactly 196 pounds.

Population parameters may be calculated if we had an unbiased census, but remember that is rare. (Certainly, we do not have an unbiased census of the weights of all men in the U.S.) Usually, we have one random sample. When reading an article that claims to know a population parameter like a population mean or population percentage, it is important to realize that it is just a guess about the population parameter, and that guess probably came from a sample. Sample statistics can be very off from the actual population parameter.

Sampling Distributions for Sample Mean Averages

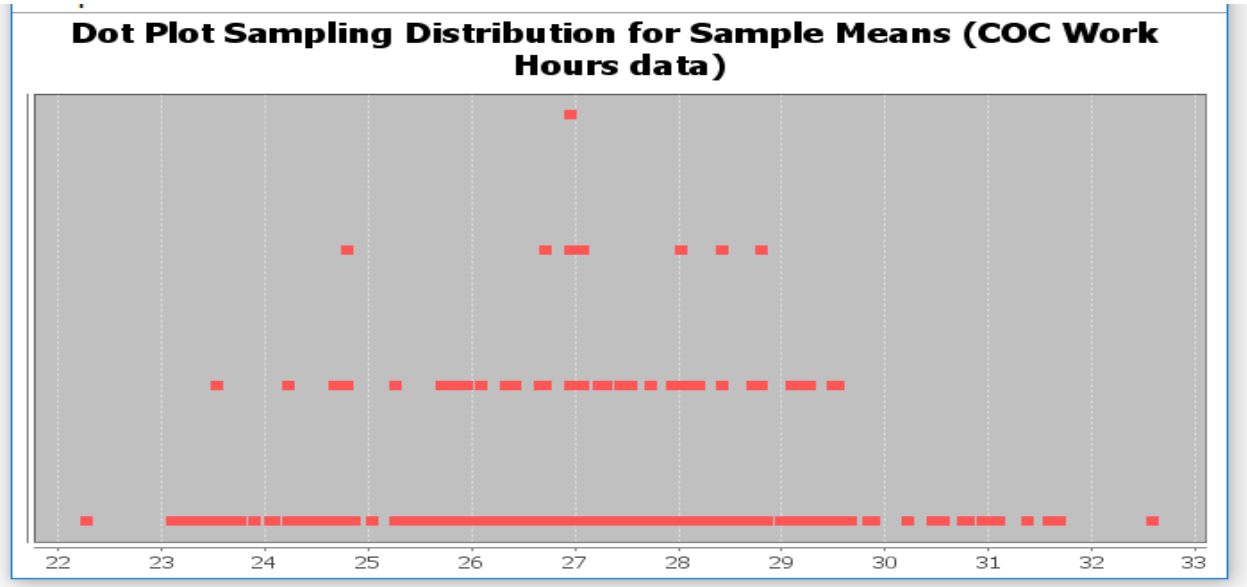
Let us go back to the example of working COC statistics students in the fall 2015 semester. We have seen that the population mean average is 27.283 hours per week, but the two random samples of 50 statistics students gave sample means that have both been off from that population mean.

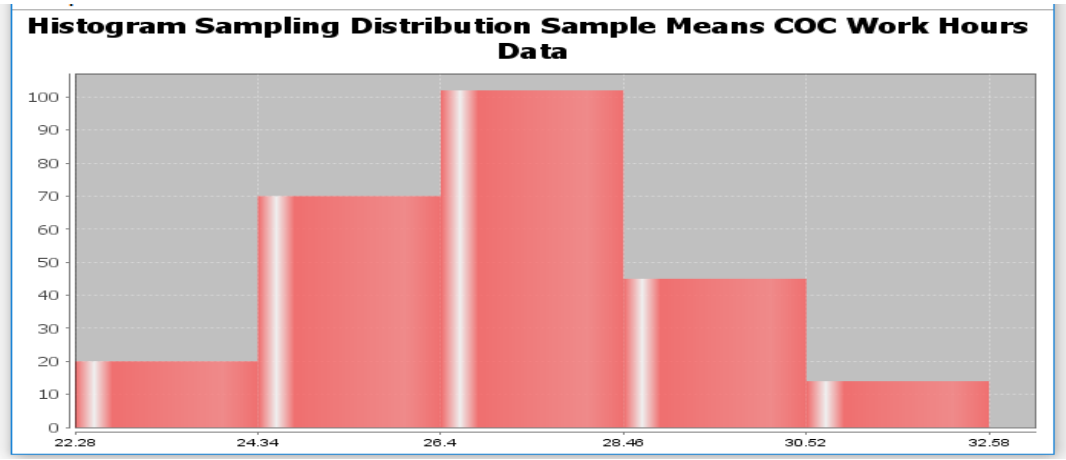
Population Parameters

Variable	Mean	Standard Deviation
Work Hours per Week COC Stat Students	27.283	12.969

Variable	Median
Work Hours per Week COC Stat Students	25.0

Let us continue to collect random samples of size 50 and calculate sample means. We collected 251 random samples and calculated 251 sample means. If we put all of the sample means on the same graph, we can create a sampling distribution.





Here is the sampling distribution we created with Statcato. Each dot in the sampling distribution represents the sample mean of a random sample. We also created a histogram of the sampling distribution to better judge the shape. Notice a few things.

Descriptive Statistics

Variable	Center (Mean) of Sampling Distribution	Standard Error
Sampling Distribution for Sample Means (COC stat students work hours)	27.127	1.916

Variable	Min	Max
Sampling Distribution for Sample Means (COC stat students work hours)	22.28	32.58

Variable	Total Number of Random Samples
C3 Sampling Distribution for Sample Means (COC stat students work hours)	251

- We took 251 random samples and calculated 251 sample means. We see sampling variability in action. The population mean is 27.283 hours per week but sample means ranged between 22.28 hours and 32.58 hours. Random sample means are usually not the same as each other and can be very different from the population mean.
- Despite the population being skewed right, the sampling distribution for these sample means is normal. This is often referred to as the "Central Limit Theorem".
- The center of the sampling distribution is 27.127 hours. This is not the mean of a sample. It is the mean average of all the sample means. Notice that the center of the sampling distribution is very close to the population mean of 27.283 hours.
- We also calculated the "standard error". This is the standard deviation of the sampling distribution (or the standard deviation of all the sample statistics) and is an important measure of sampling variability. Think of it this way. The standard error tells us how far typical sample statistics are from the center of the sampling distribution. Since the center of the sampling distribution is 27.127 hours and is pretty close to the population parameter of 27.283 hours, the standard error tells us how far typical sample statistics are from the population parameter. In this case, it tells us that typical sample means are approximately within 1.916 hours of the population mean.

Important Note: Do not confuse the standard error with the margin of error. The standard error tells us how far typical sample statistics are from the population value, but not all random samples are typical. Remember we learned from the empirical rule that typical for normal data represents only the values that are within one standard deviation from the mean (middle 68%). Usually sample values can be up to two standard deviations from the mean (middle 95%). So early statisticians thought that the margin of error should be about twice as large as the standard error. This is still a common formula for margin of error.

Margin of Error = 2 × Standard Error

Sampling Distributions for Sample Standard Deviations

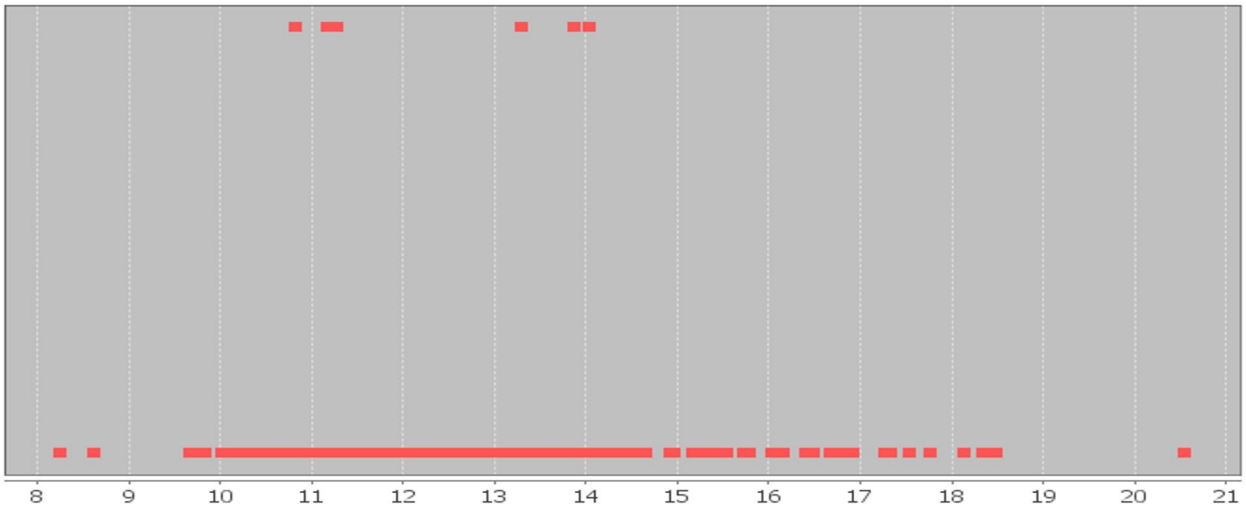
In data science, we often want to estimate many different population parameters besides the mean average. We might want to estimate the population standard deviation, the population median, or a population proportion (percentage). Using the COC work hours census data from fall 2015, we see that the population standard deviation is 12.969 hours per week. Remember, the two random sample standard deviations we have taken so far have both been off from that population standard deviation. Let us continue to collect random samples and calculate sample standard deviations. Again, we will take 251 random samples and calculate 251 random sample standard deviations. Each sample had a sample size of 50. If we put all of the sample standard deviations on the same graph, we can create a sampling distribution for sample standard deviations.

Population Parameters

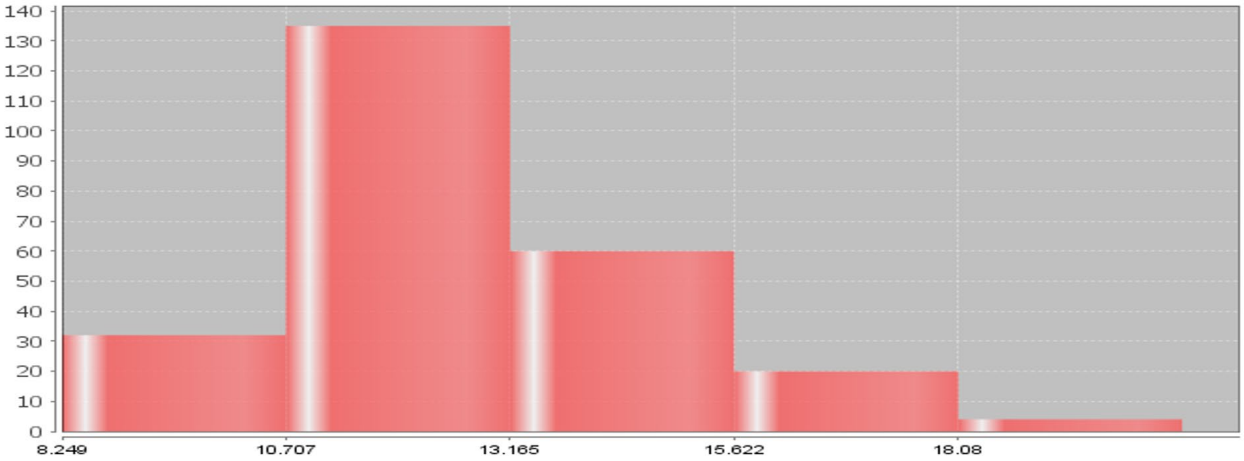
Variable	Mean	Standard Deviation
Work Hours per Week COC Stat Students	27.283	12.969

Variable	Median
Work Hours per Week COC Stat Students	25.0

Histogram of Sampling Distribution for Sample Standard Deviations COC Work Hours Data



Histogram of Sampling Distribution for Sample Standard Deviations COC Work Hours Data



Descriptive Statistics

Variable	Center (Mean) of Sampling Distribution	Standard Error
Sampling Distribution for Sample Standard Deviations (COC stat students work hours)	12.636	1.998

Variable	Min	Max
Sampling Distribution for Sample Standard Deviations (COC stat students work hours)	8.249	20.538

Variable	Total Number of Random Samples
Sampling Distribution for Sample Standard Deviations (COC stat students work hours)	251

Notice that each dot in the sampling distribution represents the sample standard deviation of a random sample of size 50. We also created a histogram of the sampling distribution to judge shape. Notice a few things.

- We took 251 random samples and calculated 251 sample standard deviations. We see sampling variability in action. The population standard deviation is 12.969 hours per week but sample standard deviations ranged between 8.249 hours all the way to 20.538 hours. Random sample standard deviations are usually not the same as each other and usually very different from the population standard deviation (σ).
- Recall that the population was skewed right. The sampling distribution for these sample standard deviations also seems to have a skew. This can be a real problem. Remember the mean (center) and standard deviation (standard error) are not very accurate when data is not normal. For this reason, when estimating a population standard deviation, we like the population to be normal.
- Notice that the center (mean) of the sampling distribution is close to the population standard deviation of 12.969 hours per week. The mean average of all the sample standard deviations was 12.636 hours per week. The median average of all the sample standard deviations was 12.229. The median is a more accurate center since this sampling distribution was skewed, but remember standard error measures the distance to the mean of the sampling distribution, not the median.
- The standard error was 1.998. Remember, the standard error tells us how far typical sample statistics are from the center (mean) of the sampling distribution. Since the center of the sampling distribution is pretty close to the population value, the standard error tells us how far typical sample statistics are from the population parameter. In this case, it tells us that typical sample standard deviations are within 1.998 hours of the population standard deviation. Again, the accuracy of the center (mean) and the spread (standard error) are in question because the sampling distribution did not look normal.

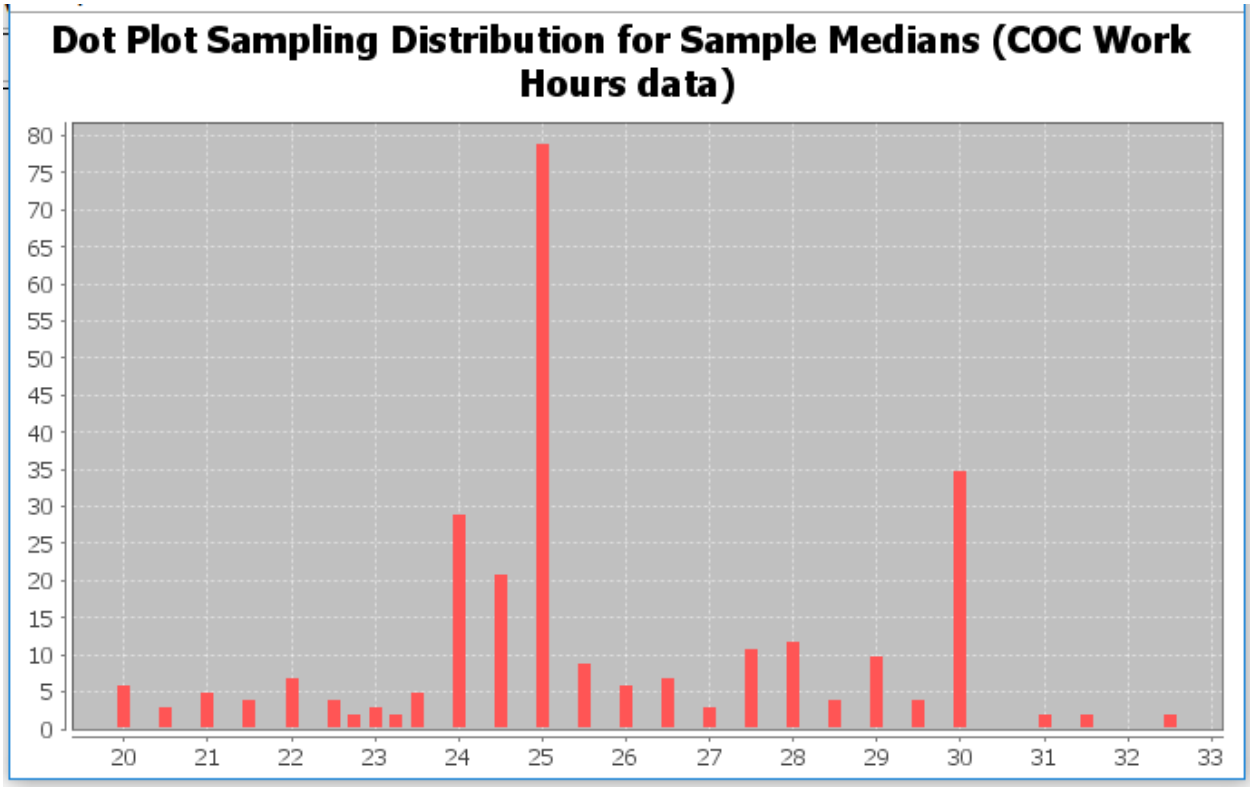
Sampling Distributions for Sample Median Averages

When data is skewed, we saw that the median average is usually more accurate than the mean, but how well do sample medians approximate population medians? Using the COC work hours census data from fall 2015, we see that the population median is 25 hours per week. Remember, the two random sample medians we have taken so far have both been off from that population median. Let us continue to collect random samples and calculate sample medians. Again, we will take 251 random samples and calculate 251 random sample medians. All of the samples had a sample size of 50. If we put all of the sample medians on the same graph, we can create a sampling distribution for sample medians.

Population Parameters

Variable	Mean	Standard Deviation
Work Hours per Week COC Stat Students	27.283	12.969

Variable	Median
Work Hours per Week COC Stat Students	25.0



Descriptive Statistics

Variable	Center (Mean) of Sampling Distribution	Standard Error
Sampling Distribution for Sample Medians (COC stat students work hours)	25.765	2.582

Variable	Min	Max
Sampling Distribution for Sample Medians (COC stat students work hours)	20.0	32.5

Variable	Total Number of Random Samples
Sampling Distribution for Sample Medians (COC stat students work hours)	251

Notice that each dot in the sampling distribution represents the sample median of a random sample. Notice a few things.

- We took 251 random samples and calculated 251 sample medians. We see sampling variability in action. The population median is 25 hours per week but sample medians ranged between 20 hours all the way to 32.5 hours. Random sample medians are usually not the same as each other and usually very different from the population median.
- Recall that the population was skewed right. The sampling distribution for these sample medians also seems to have a skew to the right. This again can be a real problem with the accuracy of the standard error.
- Again, we calculated the approximate center of the sampling distribution. This is the mean average of all of the sample medians. Notice that the center of the sampling distribution is 25.765 hours and is closer to the population median of 25 hours per week. Since this data was skewed to the right, the median of the sampling distribution will be a better measure of center. The median of the sampling distribution was 25 hours per week and in this case, was the same as the population median. Remember that the standard error measures the distance to the mean of the sampling distribution, not the median.

- We also calculated the standard error. Remember, the standard error tells us how far typical sample statistics are from the population parameter. In this case, it tells us that typical sample medians are within 2.582 hours of the population median. Again, the accuracy of the center (mean) and spread (standard error) are in question since the sampling distribution did not look normal.

Sampling Distributions for Sample Proportions (Sample Percentages)

Probably one of the most common population parameters that statisticians need to estimate is a population proportion or population percentage. There are important questions that need to be answered. What percentage of people in a country have health insurance? What percentage of people have diabetes?

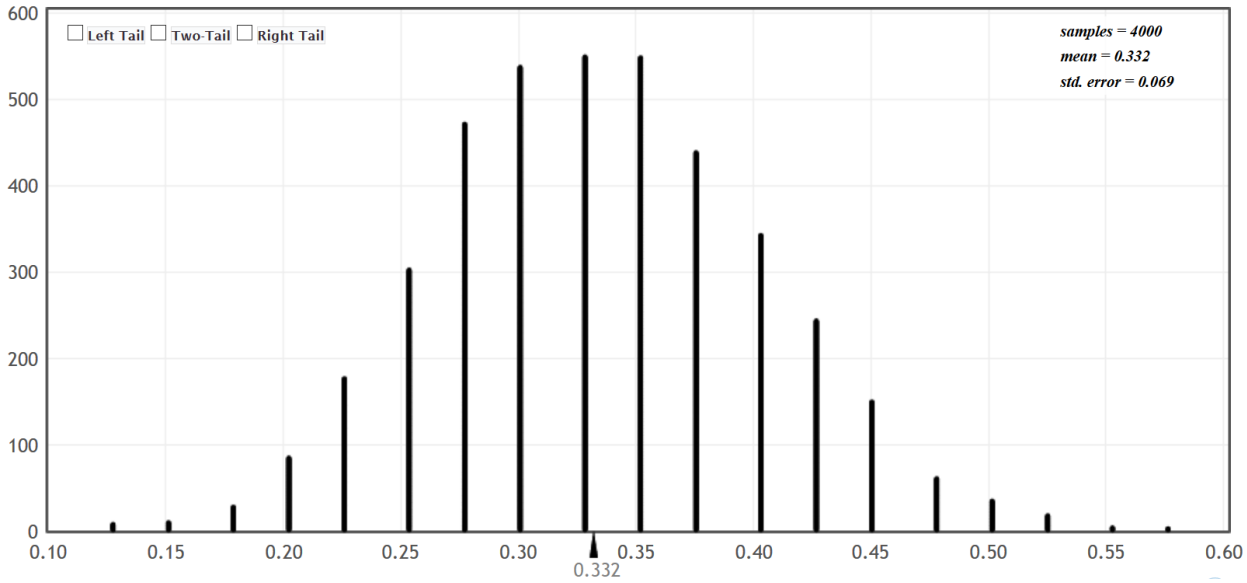
To understand sampling variability for sample percentages we will again chose an example where we have census data and therefore know the population parameter. College of the Canyons (COC) has two campuses in the Santa Clarita Valley, the Valencia campus and the Canyon Country campus. We want to know what percentage of COC statistics students attend the Canyon Country campus. In 2015, we took a census of all of the statistics students at COC and found that the population percentage that attend the Canyon Country campus was 0.332 or 33.2%. If we take random samples of 40 students at a time from that population, will the sample proportions be 0.332? Let us find out.

Here is a sampling distribution of thousands of random samples taken from the COC statistics student census. Remember the population proportion was 0.332.

Original Population

Proportion
0.332

Sampling Dotplot of Proportion



Notice that each dot in the sampling distribution represents the sample proportion of a random sample of 40 students. Notice a few things.

- We took 4000 random samples and calculated 4000 sample proportions. Again, we see sampling variability in action. The population proportion was 0.332 (33.2%) but sample proportions ranged between about 0.125 (12.5%) all the way to about 0.575 (57.5%). We see that there is a lot of sampling variability in sample proportions. Random sample proportions are usually not the same as each other and usually very different from the population proportion (π).
- Categorical data does not have a shape, but the sampling distribution for these sample proportions is normal.
- The center of the sampling distribution is calculated in the top right of the graph under "mean". This is not the mean of a sample. It is the mean average of all the sample proportions. Notice that the center of the sampling distribution is 0.332 (33.2%) and is very close to the population proportion. In fact, the center of the sampling distribution is the same as the population proportion 0.332 (33.2%).
- In the top right of the graph you will again see "standard error". Again, the standard error tells us how far typical sample statistics are from the center of the sampling distribution (population parameter). In this case it tells us that typical sample proportions are within 0.069 (6.9%) of the population proportion.

Key Notes about Sampling Distributions

1. Sampling Variability

Sampling distributions show us that random sample statistics are usually different from each other and different from the population parameter. Every time we take a random sample, we should expect to get different sample statistics and the statistics will be off from the population parameter.

2. Shape of Sampling Distributions

The shape of sampling distribution is very important. Remember the center (mean) and spread (standard error) of the sampling distribution are only accurate if the sampling distribution is normal. We saw that if the population is skewed, the sampling distribution may or may not be normal. This important topic needs further exploration.

3. Population Parameter ≈ Center of the Sampling Distributions

While one sample statistic can be far off from the population parameter, the center of a sampling distribution is usually very close to the population parameter. Let us suppose you are in a situation where you cannot collect an unbiased census. If you are able to collect multiple random samples, you can start to create a sampling distribution. Then look for the center of the distribution and you will usually have a good approximation of the population parameter. If you are using the mean of the sampling distribution as the center, we will want the sampling distribution to be normal.

Political election polls are usually dramatically off from what will happen on voting day. Yet as we get closer and closer to voting day, statisticians and data scientists seem to have a better idea of how the voting will go? If we base our population percentage of voting on one sample (one poll), we may be very far of. By the time of the vote, we have taken many polls, many samples. If we put all the sample percentages on the same graph, we have created a sampling distribution for sample proportions. Go to the center of the graph and you will have a much better idea of the population proportion, the population percentage of who will vote in what direction.

4. Standard Error and Margin of Error

Standard error is the standard deviation of the sampling distribution and tells us how far typical statistics could be from the population parameter. The accuracy of the standard error is highly reliant on the sampling distribution being normal.

Remember standard error and margin of error are not the same thing. Standard error measures typical statistics. Many sample statistics may not be typical. The margin of error considers sample statistics that are not just typical. Usually the margin of error is about twice as large as the standard error.

Optional Sampling Distribution Class Activity 1

Exploring Sampling Variability for Mean Averages with a Sampling Distribution

The goal of this activity is to explore how well random samples approximate population values. Normally we do not know population values and we must use a sample value to approximate the population value. This is called a "point estimate". For this activity, we will look at some population data from International Coffee Organization (ICO). We will be using the "Columbian Mild" price data in U.S. cents per pound. The population mean average price was 136.43 cents per pound. Again, in real data analysis, we often do not know the population value, but for this activity, it is useful for comparison purposes.

Open the "Sampling Distribution Data 1" in Excel. A total of 120 random samples have been taken from the Columbian Mild data. All the data sets have 30 coffee prices. Each person in the class will be finding the mean of a few of these data sets. Once you find your sample means, you will put a magnet up or draw a dot on the board to represent the sample mean you found. When everyone's magnets or dots are up on the board, we will have generated a "sampling distribution".

Answer the following questions:

1. The population mean was 136.43 cents. How many cents was the sample mean you calculated from the population mean of 136.43 cents? (If you calculated more than one sample mean, answer the question for all the sample means you calculated.) This is called the "Margin of Error".

2. Look at the dots or magnets on the board. Did all the sample means come out to be the same as the population mean of 136.43cents? Why do you thing this happened? Aren't random samples supposed to be good approximations of the population? What does this tell you about sampling variability?

3. Normally, we may have only one random sample. If all you knew was one of the random samples on the board, how difficult would it be to determine that the population mean is really 136.43 cents? What does this tell us about the difficulty in determining population values from one random sample?

4. Estimate the shape and center of the sampling distribution on the board. Is the center of the graph close to the population mean of 136.43? Would the center of the sampling distribution be a better approximation of the population mean than a single sample mean?

5. The standard deviation of a sampling distribution is often called the "standard error" and is an important part of inferential statistics. Estimate how far typical dots are from the center of the sampling distribution. This is the standard deviation of the sampling distribution, which is called "Standard Error".

Optional Sampling Distribution Class Activity 2

Exploring Sampling Variability for Percentages with a Sampling Distribution

The goal of this activity is to explore how well random sample percentages approximate population percentages. Normally we do not know population percentage and we must use a sample percentage to approximate the population percentage. This is called a "point estimate". For this activity, we will be flipping coins 30 times and count the number of tails. Then calculate the sample percentage of tails. Each person will do three sets of 30 and therefore get three sample percentages. Again, in real data analysis, we often do not know the population value, but for this activity, it is useful for comparison purposes. Our goal is to see how well random sample percentages approximate population percentages.

Each person in the class will be finding three sample percentages. Once you find each sample percent, you will put a magnet up or draw a dot on the board to represent the sample percent you found. When everyone's magnets or dots are up on the board, we will have generated a "sampling distribution" of sample percentages.

Answer the following questions:

1. In a perfect world and a fair coin, what should the population percentage for getting tails be? So in a sample of 30 how many times do we expect to get tails? In sampling, we often do not get what we expect. How far were the sample percentages you calculated from the population percentage?

2. Look at the dots or magnets on the board. Did all the sample percentages come out to be the same as the population percentage? Why do you thing this happened? Aren't random samples supposed to be good approximations of the population? What does this tell you about sampling variability?

3. Normally, we may have only one random sample. If all you knew was one of the sample percentage on the board, and you never knew the expected population value, how difficult would it be to determine what the population percentage really is? What does this tell us about the difficulty in determining population values from one random sample?

4. Estimate the shape and center of the sampling distribution on the board. Is the center of the graph close to the population percentage of 0.5? Would the center of the sampling distribution be a better approximation of the population percentage than a single sample percentage?

5. The standard deviation of a sampling distribution is often called the "standard error" and is an important part of inferential statistics. Estimate how far typical dots are from the center of the sampling distribution. This is the standard deviation of the sampling distribution, which is called "Standard Error".

Problem Set Section 2B

Directions: Answer the following questions about sampling distributions.

1. Describe the process of making a sampling distribution.

2. What can sampling distributions tell us about sampling variability?

3. What is a point estimate? Discuss how point estimates create confusion for people reading articles and scientific reports.

4. Discuss the shape of sampling distributions. When the population is skewed, is the sampling distribution always normal? Why is it important for a sampling distribution to be normal? In the examples in this section, which statistics had a normal sampling distribution? Which statistics had a skewed sampling distribution?

5. Explain how the standard error is calculated. What does the standard error tell us about sample statistics and the population parameter? Why is the standard error only accurate when the sampling distribution is normal?

6. What is the difference between standard error and margin of error? Is the standard error smaller or larger than the margin of error?

(#7-16) For the following problems, copy the indicated census data set from the Math 140 Survey Data at www.matt-teachout.org. We will be assuming this is an unbiased census and therefore know the population mean. Open StatKey at www.lock5stat.com. Under the "sampling distributions" menu, click on "mean". You should see "sampling distribution for the mean". Under "edit data" paste in the indicated data set. Under "chose samples of size n", put in the indicated sample size. Create a sampling distribution and then answer the following questions.

7. Use StatKey to create a sampling distribution with sample size 10 from the Age in Years census data (Math 140 Survey Data).

a) What was the shape and mean average of the population?
b) Were all the sample means the same as the population mean?
c) Were all the sample means the same as each other?
d) How many random samples did you take when you created the sampling distribution?
e) What is the shape of the sampling distribution?
f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
g) What is the standard error? Write a sentence explaining the meaning of the standard error.

8. Use StatKey to create a sampling distribution with sample size 100 from the Age in Years census data (Math 140 Survey Data)

a) What was the shape and mean average of the population?
b) Were all the sample means the same as the population mean?
c) Were all the sample means the same as each other?
d) How many random samples did you take when you created the sampling distribution?
e) What is the shape of the sampling distribution?
f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
g) What is the standard error? Write a sentence explaining the meaning of the standard error.
h) How does the standard error for sample size 10 compare to the standard error for sample size 100?
i) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 100?

9. Use StatKey to create a sampling distribution with sample size 10 from the sleep hours per night census data (Math 140 Survey Data)

a) What was the shape and mean average of the population?
b) Were all the sample means the same as the population mean?
c) Were all the sample means the same as each other?
d) How many random samples did you take when you created the sampling distribution?
e) What is the shape of the sampling distribution?
f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
g) What is the standard error? Write a sentence explaining the meaning of the standard error.

10. Use StatKey to create a sampling distribution with sample size 25 from the sleep hours per night census data (Math 140 Survey Data)

a) What was the shape and mean average of the population?
b) Were all the sample means the same as the population mean?
c) Were all the sample means the same as each other?
d) How many random samples did you take when you created the sampling distribution?
e) What is the shape of the sampling distribution?
f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
g) What is the standard error? Write a sentence explaining the meaning of the standard error.
h) How does the standard error for sample size 10 compare to the standard error for sample size 25?
i) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 25?

11. Use StatKey to create a sampling distribution with sample size 10 from the cell phone bill (in dollars per month) census data (Math 140 Survey Data).

a) What was the shape and mean average of the population?
b) Were all the sample means the same as the population mean?
c) Were all the sample means the same as each other?
d) How many random samples did you take when you created the sampling distribution?
e) What is the shape of the sampling distribution?
f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
g) What is the standard error? Write a sentence explaining the meaning of the standard error.

12. Use StatKey to create a sampling distribution with sample size 100 from the cell phone bill (in dollars per month) census data (Math 140 Survey Data).

a) What was the shape and mean average of the population?
b) Were all the sample means the same as the population mean?
c) Were all the sample means the same as each other?
d) How many random samples did you take when you created the sampling distribution?
e) What is the shape of the sampling distribution?
f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
g) What is the standard error? Write a sentence explaining the meaning of the standard error.

h) How does the standard error for sample size 10 compare to the standard error for sample size 100?
i) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 100?

13. Use StatKey to create a sampling distribution with sample size 10 from the travel time to get to school in minutes (Math 140 Survey Data).

a) What was the shape and mean average of the population?
b) Were all the sample means the same as the population mean?
c) Were all the sample means the same as each other?
d) How many random samples did you take when you created the sampling distribution?
e) What is the shape of the sampling distribution?

f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
g) What is the standard error? Write a sentence explaining the meaning of the standard error.

14. Use StatKey to create a sampling distribution with sample size 40 from the travel time to get to school in minutes (Math 140 Survey Data).

a) What was the shape and mean average of the population?
b) Were all the sample means the same as the population mean?
c) Were all the sample means the same as each other?
d) How many random samples did you take when you created the sampling distribution?
e) What is the shape of the sampling distribution?
f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
g) What is the standard error? Write a sentence explaining the meaning of the standard error.
h) How does the standard error for sample size 10 compare to the standard error for sample size 40?
i) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 40?

15. Use StatKey to create a sampling distribution with sample size 10 from the work hours per week for COC college students (Math 140 Survey Data).

a) What was the shape and mean average of the population?
b) Were all the sample means the same as the population mean?
c) Were all the sample means the same as each other?
d) How many random samples did you take when you created the sampling distribution?
e) What is the shape of the sampling distribution?
f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
g) What is the standard error? Write a sentence explaining the meaning of the standard error.

16. Use StatKey to create a sampling distribution with sample size 40 from the work hours per week for COC college students (Math 140 Survey Data).

a) What was the shape and mean average of the population?
b) Were all the sample means the same as the population mean?
c) Were all the sample means the same as each other?
d) How many random samples did you take when you created the sampling distribution?
e) What is the shape of the sampling distribution?
f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
g) What is the standard error? Write a sentence explaining the meaning of the standard error.
h) How does the standard error for sample size 10 compare to the standard error for sample size 40?
i) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 40?

(#17-26) The following population proportions come from the Math 140 Survey Data at www.matt-teachout.org. We will be assuming this is an unbiased census and therefore know the population proportion (%). Open StatKey at www.lock5stat.com. Under the "sampling distributions" menu, click on "proportion". You should see "sampling distribution for a proportion". Under "edit proportion", enter the given population proportion. Create a sampling distribution and then answer the following questions.

17. A census of COC statistics students in the fall 2015 semester indicated that the population proportion of statistics students with brown hair is 0.537. Use this population proportion to create a sampling distribution with sample size 10 with StatKey.

a) Were all the sample proportions the same as the population proportion?
b) Were all the sample proportions the same as each other?
c) How many random samples did you take when you created the sampling distribution?
d) What is the shape of the sampling distribution?
e) What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
f) What is the standard error? Write a sentence explaining the meaning of the standard error.

18. A census of COC statistics students in the fall 2015 semester indicated that the population proportion of statistics students with brown hair is 0.537. Use this population proportion to create a sampling distribution with sample size 100 with StatKey.

a) Were all the sample proportions the same as the population proportion?
b) Were all the sample proportions the same as each other?
c) How many random samples did you take when you created the sampling distribution?
d) What is the shape of the sampling distribution?
e) What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
f) What is the standard error? Write a sentence explaining the meaning of the standard error.
g) How does the standard error for sample size 10 compare to the standard error for sample size 100?
h) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 100?

19. A census of COC statistics students in the fall 2015 semester indicated that the population proportion of statistics students that smoke cigarettes is 0.091. Use this population proportion to create a sampling distribution with sample size 10 with StatKey.

a) Were all the sample proportions the same as the population proportion?
b) Were all the sample proportions the same as each other?
c) How many random samples did you take when you created the sampling distribution?
d) What is the shape of the sampling distribution?
e) What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
f) What is the standard error? Write a sentence explaining the meaning of the standard error.

20. A census of COC statistics students in the fall 2015 semester indicated that the population proportion of statistics students that smoke cigarettes is 0.091. Use this population proportion to create a sampling distribution with sample size 100 with StatKey.

a) Were all the sample proportions the same as the population proportion?
b) Were all the sample proportions the same as each other?
c) How many random samples did you take when you created the sampling distribution?
d) What is the shape of the sampling distribution?
e) What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
f) What is the standard error? Write a sentence explaining the meaning of the standard error.
g) How does the standard error for sample size 10 compare to the standard error for sample size 100?
h) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 100?

21. Approximately 60% of college students in the U.S. were able to finish their bachelor's degree in six years. Use this population proportion to create a sampling distribution with sample size 10 with StatKey.

a) Were all the sample proportions the same as the population proportion?
b) Were all the sample proportions the same as each other?
c) How many random samples did you take when you created the sampling distribution?
d) What is the shape of the sampling distribution?
e) What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
f) What is the standard error? Write a sentence explaining the meaning of the standard error.

22. Approximately 60% of college students in the U.S. were able to finish their bachelor's degree in six years. Use this population proportion to create a sampling distribution with sample size 100 with StatKey.

a) Were all the sample proportions the same as the population proportion?
b) Were all the sample proportions the same as each other?
c) How many random samples did you take when you created the sampling distribution?
d) What is the shape of the sampling distribution?
e) What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
f) What is the standard error? Write a sentence explaining the meaning of the standard error.
g) How does the standard error for sample size 10 compare to the standard error for sample size 100?
h) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 100?

23. Approximately 9.4% of all adults in the U.S. have diabetes. Use this population proportion to create a sampling distribution with sample size 10 with StatKey.

a) Were all the sample proportions the same as the population proportion?
b) Were all the sample proportions the same as each other?
c) How many random samples did you take when you created the sampling distribution?
d) What is the shape of the sampling distribution?
e) What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
f) What is the standard error? Write a sentence explaining the meaning of the standard error.

24. Approximately 9.4% of all adults in the U.S. have diabetes. Use this population proportion to create a sampling distribution with sample size 100 with StatKey.

a) Were all the sample proportions the same as the population proportion?
b) Were all the sample proportions the same as each other?
c) How many random samples did you take when you created the sampling distribution?
d) What is the shape of the sampling distribution?
e) What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
f) What is the standard error? Write a sentence explaining the meaning of the standard error.
g) How does the standard error for sample size 10 compare to the standard error for sample size 100?
h) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 100?

25. Approximately 90% of all lung cancer cases are caused by cigarette smoking. Use this population proportion to create a sampling distribution with sample size 10 with StatKey.

a) Were all the sample proportions the same as the population proportion?
b) Were all the sample proportions the same as each other?
c) How many random samples did you take when you created the sampling distribution?
d) What is the shape of the sampling distribution?
e) What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
f) What is the standard error? Write a sentence explaining the meaning of the standard error.

26. Approximately 90% of all lung cancer cases are caused by cigarette smoking. Use this population proportion to create a sampling distribution with sample size 100 with StatKey.

a) Were all the sample proportions the same as the population proportion?
b) Were all the sample proportions the same as each other?
c) How many random samples did you take when you created the sampling distribution?
d) What is the shape of the sampling distribution?
e) What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
f) What is the standard error? Write a sentence explaining the meaning of the standard error.
g) How does the standard error for sample size 10 compare to the standard error for sample size 100?
h) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 100?

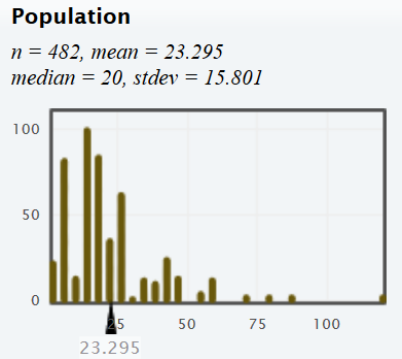
Section 2C – The Central Limit Theorem

In the last section, we saw that when estimating population parameters from samples, it is very important for a sampling distribution to be normal. The accuracy of the center of the sampling distribution (population estimate) and the spread of the sampling distribution (standard error) are tied to the sampling distribution being normal. We also saw that if the population was skewed, the sampling distribution may or may not look normal. In this section, we will discuss further the shape of sampling distributions and determine what conditions need to be met in order to get a normal sampling distribution.

Sample Means

Let us start by looking at sample means. Let us look at the census of College of the Canyons (COC) statistics students taken in the fall 2015 semester. The variable we will look at is how many minutes it takes to commute to COC.

Census Data (Commute Time in Minutes for COC Stat Students Fall 2015)



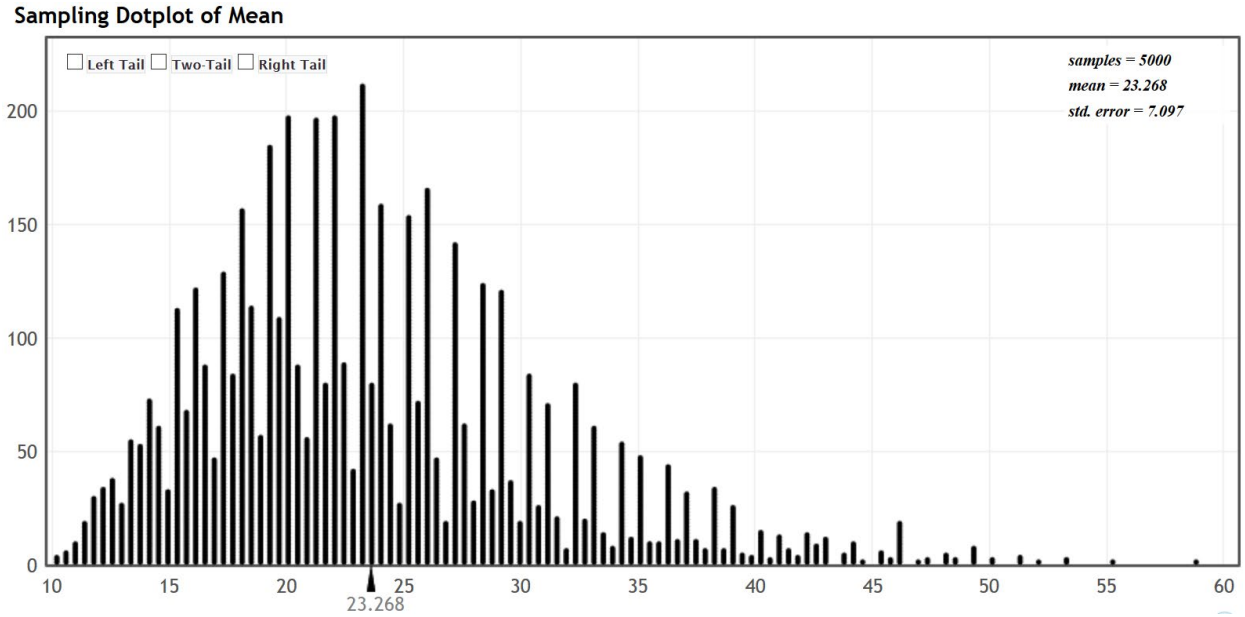
We see that the population is skewed with a population mean average commute time of 23.295 minutes. We will assume that the census is unbiased and that the population mean is really 23.295 minutes.

Key Question: If the population is skewed, what conditions need to be met in order for the sampling distribution to look normal?

Mean Example 1: Sample Size of Seven from a Skewed Population

Let us take many random samples from the census of COC stat students commute times, calculated the sample mean from each sample and then put the sample means on the same graph. This is called a sampling distribution for sample means. For this example, we used small samples with a sample size of seven. We used the sampling distribution function on StatKey to create 5000 random samples with each sample have seven commute times. We calculated 5000 sample means and put them on the same graph. Notice the sampling distribution still looks skewed. In addition, notice that the center (mean) of the sampling distribution was 23.268 minutes and the standard error is 7.097 minutes. We would not trust the accuracy of the standard error or the mean of the sampling distribution because the sampling distribution was not normal.

- Shape of Sampling Distribution: Skewed Right
- Center (mean) of the sampling distribution ≈ 23.268 minutes
- Standard error ≈ 7.097 minutes.



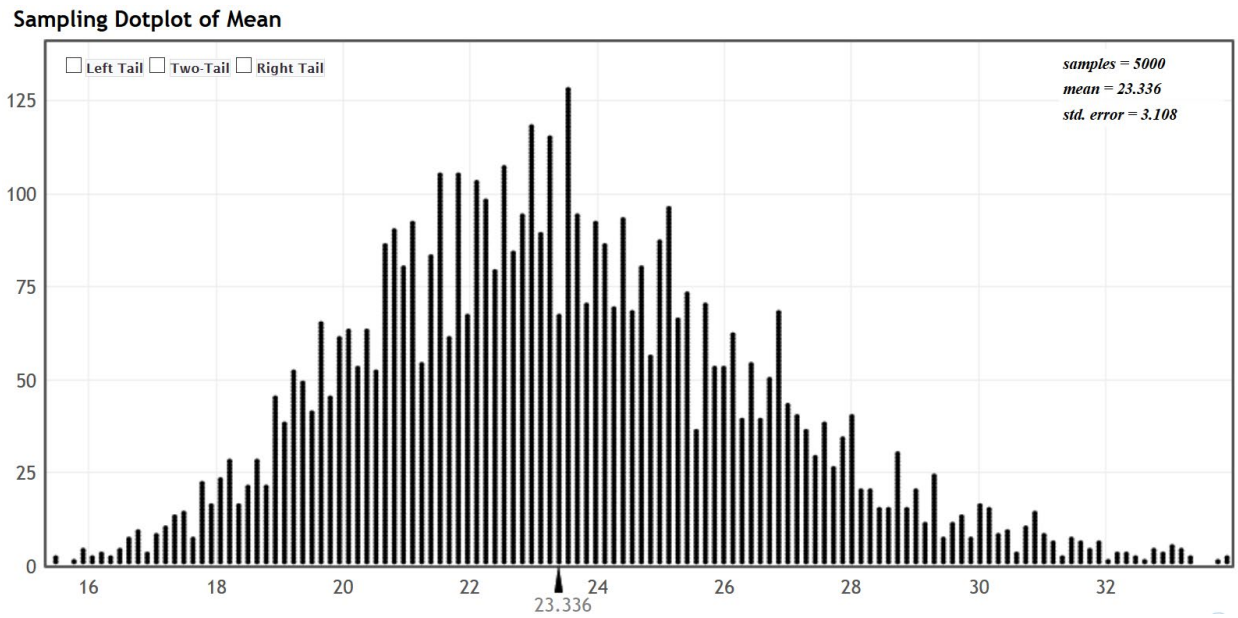
Mean Example 2: Sample Size of Twenty-Five from a Skewed Population

Let us create another sampling distribution from the census of COC stat student commute times. This time we will increase the sample size to twenty-five. Each sample will have twenty-five commute times. We used the sampling distribution function on StatKey to create 5000 random samples with each sample having a sample size of twenty-five. Notice the sampling distribution now looks nearly normal. The center (mean) of the sampling distribution was 23.336 minutes and the standard error is 3.108 minutes. We can trust the accuracy of the standard error and the mean of the sampling distribution because the sampling distribution was nearly normal.

- Shape of Sampling Distribution: Nearly Normal
- Center (mean) of the sampling distribution ≈ 23.336 minutes
- Standard error ≈ 3.108 minutes.

Notice also that the standard error for this sample size (n=25) is smaller than the standard error for the very small sample size (n=7). This is a very important principle, more random data results in less error. The larger the sample size, the smaller the standard error, and the more normal the sampling distribution looks.

More Random Data → Less Error → Sampling Distribution becomes more normal



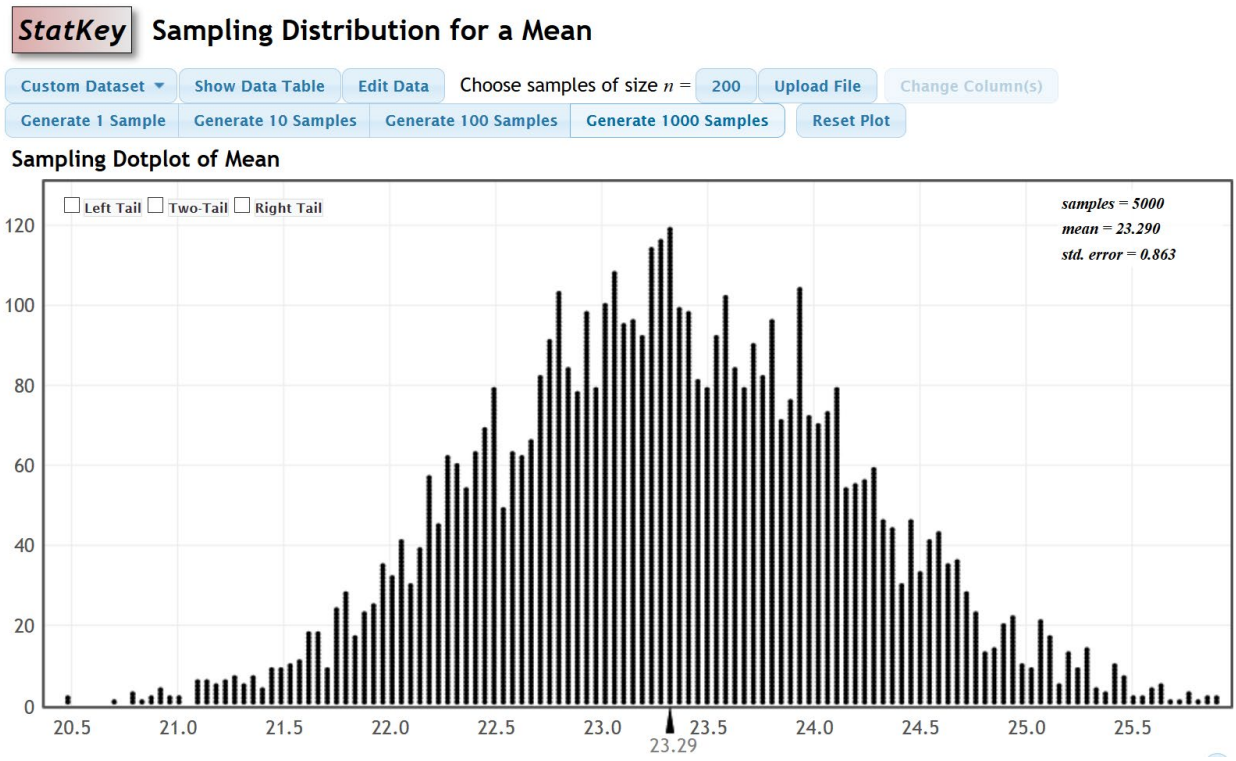
Mean Example 3: Sample Size of Two Hundred from a Skewed Population

Let us create one more sampling distribution from the COC stat students commute times data. This time we will increase the sample sizes to two hundred. Notice the sampling distribution now looks very normal. The center (mean) of the sampling distribution was 23.290 minutes and the standard error has dropped to 0.863 minutes.

- Shape of Sampling Distribution: Normal
- Center (mean) of the sampling distribution ≈ 23.290 minutes
- Standard error ≈ 0.863 minutes.

Notice also that the standard error for this sample size (n=200) is smaller than the standard error for the sample size (n=25). Remember, the larger the sample size, the smaller the standard error, and the more normal the sampling distribution looks.

More Random Data → Less Error → Sampling Distribution becomes more normal



Summary: If a population is skewed, it seems we need a larger sample size, for the sampling distribution to look normal. As the sample size increases, the standard error decreases, and the sampling distribution looks more normal. This is the idea behind the "Central Limit Theorem". A common rule when dealing with means is that if the population is skewed the sample size should be at least 30 for the sampling distribution for sample means to look normal.

Central Limit Theorem: If the sample size is sufficiently large, the sampling distribution for sample means will have a normal shape even if the population is skewed.

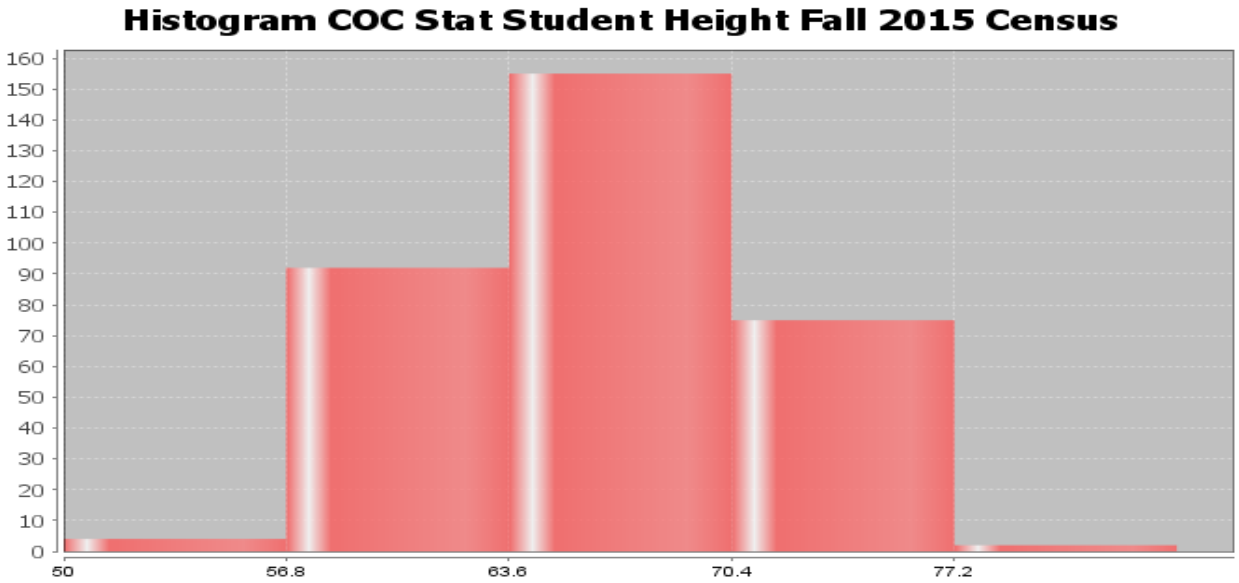
Key Question: What would happen if the population were already normal?

Mean Example 4: Sampling Distribution from a normal population.

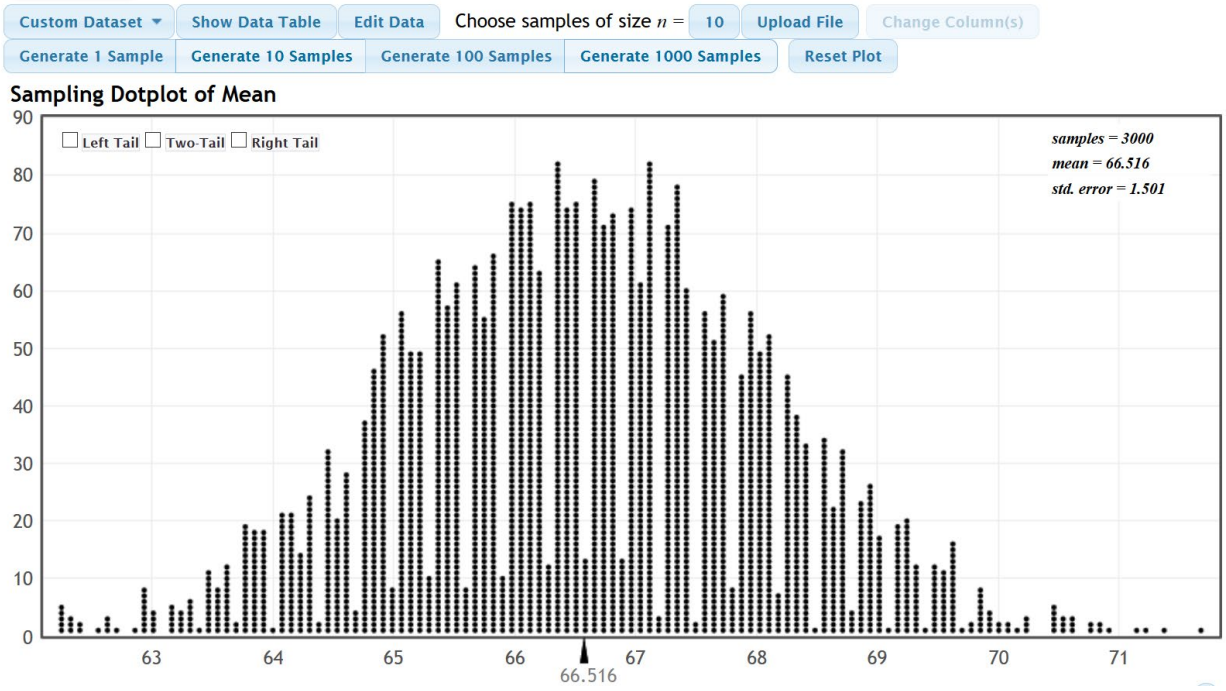
Let us now look at an example of a census with a normal shape. In the fall 2015 semester, we took a census of all of the statistics students at COC and asked them their heights in inches. We will assume this was an unbiased census. This population looked very normal with a population mean average height of 66.511 inches and a population standard deviation of 4.787 inches. For this example, we will focus on the mean.

Population Parameters

Variable	Population Mean	Population Standard Deviation
COC Stat Students Population height (in INCHES)	66.511	4.787



Now let us see what happens if we take thousands of samples from this population. We will start with small sample sizes of 10 stat students at a time.



We took 3000 random samples each of size ten and calculated 3000 sample means to create this sampling distribution. Notice a few key things.

- The sample means are different. We see sampling variability in action. The population mean was 66.511 inches but the sample means could be anywhere from about 62 inches to 72 inches. Sample statistics are different and usually very different than the population parameter.

- Even though we have a very small sample size of ten, the sampling distribution still looks normal. This means that the center (mean) of the sampling distribution and the standard error are relatively accurate even for a sample size of ten.

- The center of the sampling distribution (66.516 inches) is very close to the population mean of (66.511 inches)
- We have calculated the standard error of 1.501. For a sample size of 10, typical sample means are within 1.501 inches of the population mean. The margin of error is probably closer to 3 inches (2 x standard error).

Sample Mean Summary

Let us summarize our findings about sample means from random samples.

1. If the population is skewed, we will need a sample size of at least 30 or higher in order to insure that our sampling distribution for sample means will be nearly normal.

2. If the population is already normal, then the sampling distribution for sample means will be normal for any sample size.

Important note about sample size:

Even though the minimum requirement for sample means is a sample size of 30 or above, this does not mean we are happy with a data set of only 30. Remember less data results in more error. For random data, the bigger the sample size the better. Thirty is just the bare minimum requirement to insure that the sampling distribution for sample means will look nearly normal.

Standard Deviation Example 1: Standard Deviation and Variance

Remember that the sample variance is the square of the standard deviation. Statisticians often opt to estimate variability in sample variances instead of standard deviation. Later, we can take the square root of the variance estimates to get the standard deviation.

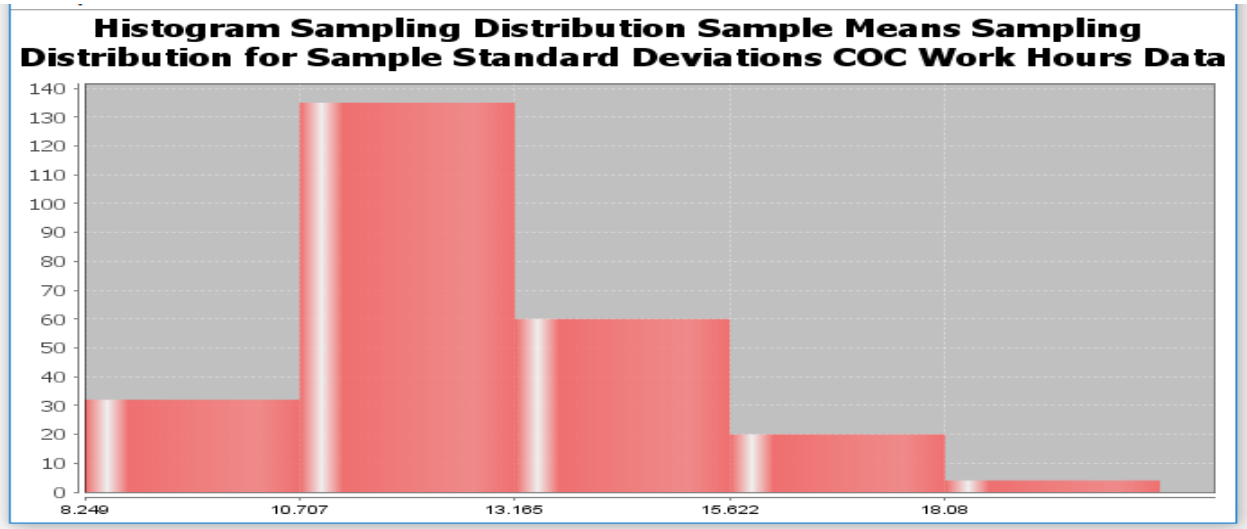
If the population was skewed, what is the shape of sampling distributions for sample standard deviations? Are there any sample size requirements for estimating sample standard deviations? What if the population was already normal?

In the last section, we looked at the COC work hours census data from fall 2015. We see that the population standard deviation is 12.969 hours per week. We created a sampling distribution of 251 random samples and calculate 251 random sample standard deviations. Each sample had a sample size of 50. If we put all of the sample standard deviations on the same graph, we can create a sampling distribution for sample standard deviations.

Population Parameters

Variable	Mean	Standard Deviation
Work Hours per Week COC Stat Students	27.283	12.969

Variable	Median
Work Hours per Week COC Stat Students	25.0



Notice that while a sample size of 50 would be large enough to ensure a sampling distribution of sample means to be normal; it does not insure a sampling distribution of sample standard deviations to be normal. If the population is skewed, the sampling distribution for sample standard deviations will tend to be skewed.

Sample Standard Deviation and Sample Variance Summary

Let us summarize our findings about sample standard deviations and sample variance from random samples.

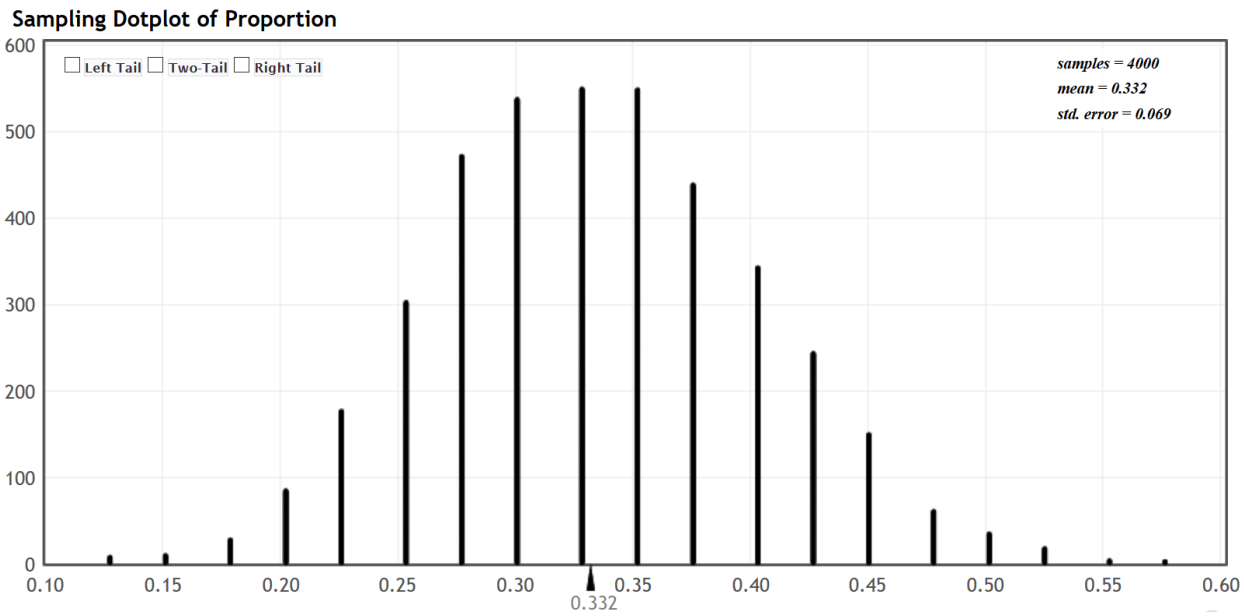
1. A sampling distributions of sample variance is usually skewed right. Later we will see that if the population is normal, the sampling distribution for sample variance will follow a skewed right Chi-Squared distribution. Requirements for traditional techniques for estimating population variance or population standard deviations usually require the population to be normal no matter what the sample size is. If the population were not normal, then we would have to resort to different technique like bootstrapping.

Sample Proportion Example 1:

In the last section, we looked at the fall 2015 census of COC stat students and found that the population percentage that attend the Canyon Country campus was 0.332 or 33.2%. Here is a sampling distribution of thousands of random samples taken from the COC statistics student census. Remember the population proportion was 0.332.

Original Population

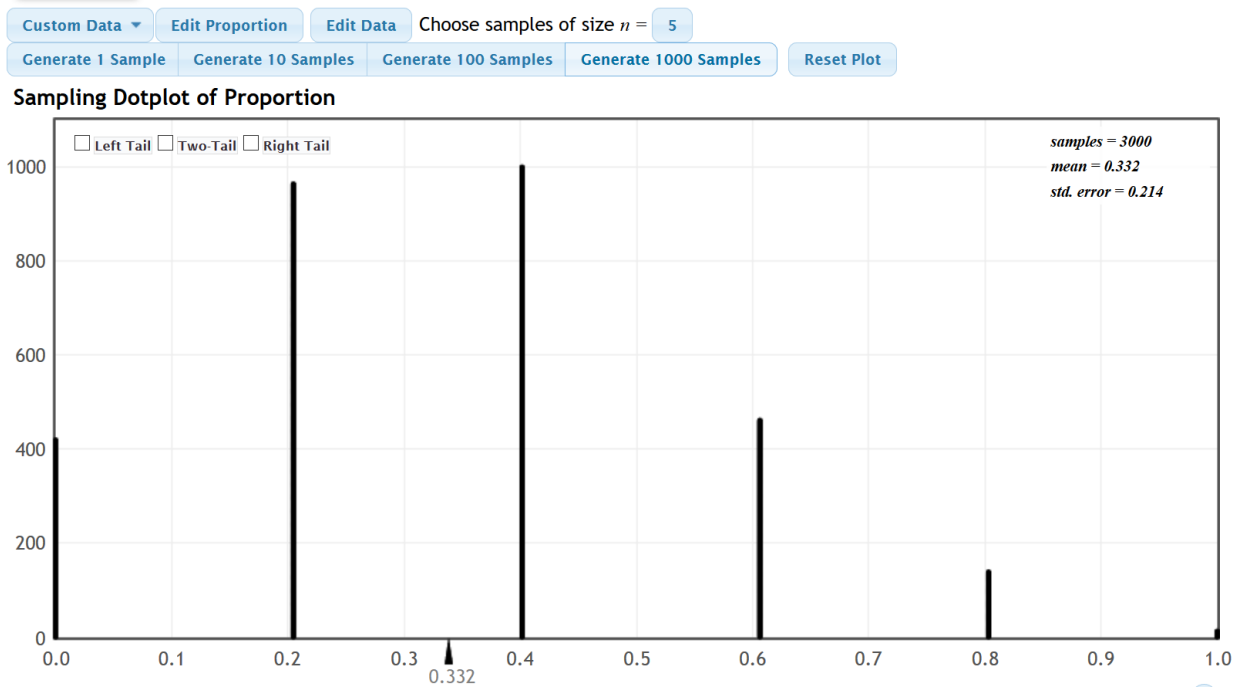
Proportion
0.332



Notice that for a sample size of 40, the sampling distribution looks normal. In addition, the center of the sampling distribution was very close to the population proportion of 0.332.

What if we decrease the sample size?

Let us look at a sampling distribution for sample proportions from the same population, but now we will decrease the sample size to five.



Notice at a sample size of only five, the sampling distribution looks skewed right. Notice that the center (mean) of all the sample proportions was still very close to 0.332, but we will not have much confidence in the standard error from a sampling distribution that is not normal.

So what sample size insures a normal sampling distribution for sample proportions?

The "At Least Ten" rule

It turns out that for random categorical data, the random sample should have at least ten successes and at least ten failures. We should have at least 10 statistics students from the Canyon Country campus and at least 10 that are not from the Canyon Country campus to insure that the sampling distribution will look normal.

Notice if we only had a random sample of five stat students, it is impossible to get at least ten from Canyon Country and at least ten not from Canyon Country.

There is no minimum sample size requirement for categorical data because the population proportion will be different in each situation.

Why did the sampling distribution for samples of size 40 work?

If we know the population proportion (π), here is a common formula for estimating the number of success and failures in random categorical sample data:

Expected number of success for sample size (n): $n(\pi)$
Expected number of failures for sample size (n): $n(1-\pi)$

For a sample size of 40, will we be likely to get ten successes and ten failures? If the population proportion for Canyon Country is 0.332, we are likely to get about 13 students from Canyon Country and 27 students not from Canyon Country.

$n(\pi)$ = 40(0.332) = 13.28

$n(1-\pi)$ = 40(1−0.332) = 40(0.668) = 26.72

Important Note: *Remember we rarely have an unbiased census, so we may have no idea what the population proportion is. All we have is random sample data. In that case, you will want your random categorical sample data to have at least ten success and at least ten failures. That does not mean twenty!*

Summary of Sampling Distributions for sample proportions (sample %)

- Categorical data does not have a shape. Yet if we compute thousands of sample proportions and put them on the same graph, the sampling distribution will have a shape.
- To insure the sampling distribution for sample proportions will be normal we want to have at least ten successes and at least ten failures in our random categorical sample data.

Key Question#1: Why is it so important for a sampling distribution to be normal?

We will discuss this in greater detail in later sections, but here are two of the main reasons.

- Remember standard error is the standard deviation of the sampling distribution and measures the typical distance from the mean (center) of the sampling distribution. Neither the standard error nor the center (mean) of the sampling distribution are very accurate unless the sampling distribution is normal.
- Before computers were invented, statisticians relied on formulas to understand sampling variability, calculate standard error and estimate population parameters. Many of these formulas are based on normal curves and are not accurate if the sampling distribution is not normal. This is why conditions or assumptions for sample means and sample proportions are often tied to making sure the sampling distribution is normal when estimating population parameters.

Key Question#2: Is there a way to estimate a population parameter and understand sampling variability when the sampling distribution is not normal?

- Yes. Computer technology may be used to understand sampling variability in the case when our sampling distribution is not likely to be normal. Techniques like bootstrapping and randomized simulation were invented to be able to understand sampling variability, calculate standard error, and estimate or check population parameters when the sampling distribution is not normal. We will discuss these techniques in later chapters.

Problem Set Section 2C

Directions: Answer the following questions about sample size requirements and the shape of sampling distributions.

1. Why is it important for a sampling distribution for sample means or sample proportions to be normal?

2. What conditions should be met to insure that a sampling distribution of sample proportions is normal?

3. State the Central Limit Theorem and explain the ideas behind it.

4. Suppose the population is not normal. If we increase the sample size, what will happen to the standard error and the shape of the sampling distribution of sample means?

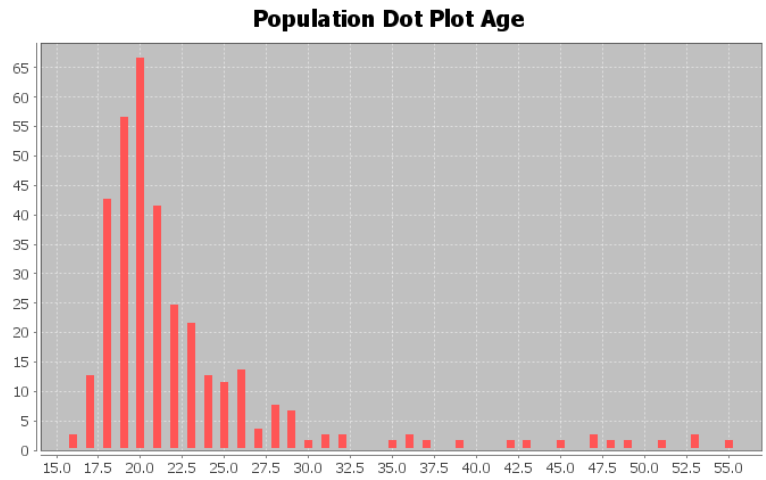
5. Suppose the population is not normal. If we decrease the sample size, what will happen to the standard error and the shape of the sampling distribution of sample means?

6. Suppose the population is not normal. What conditions should be met in order to insure that a sampling distribution of sample means is normal?

7. If the population is normal, will the sampling distribution for sample means look normal for very small sample sizes?

8. Median averages, variance and standard deviation can have very irregular looking sampling distributions. This can make traditional formula calculations difficult. Is there a way to study sampling variability and estimate population parameters when a sampling distribution is not normal or when traditional formulas are not accurate?

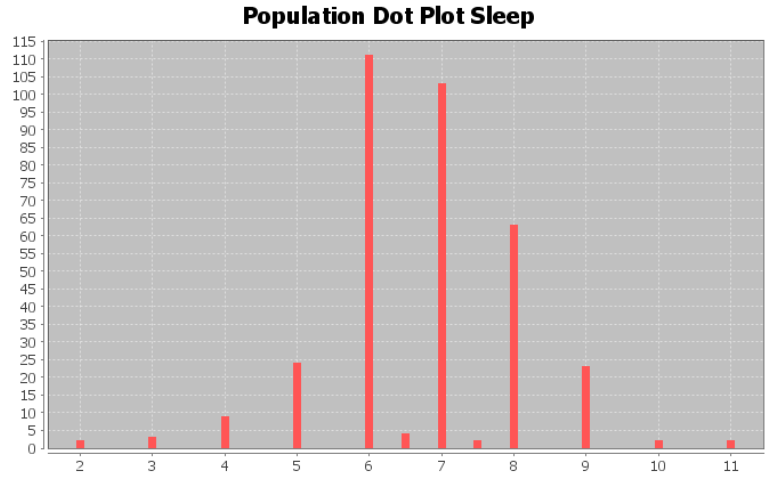
9. The following graph and population mean were created with Statcato from the "age in years" census data (Math 140 Survey Data). Assume this census represents the population of Math 140 statistics students at College of the Canyons in the fall 2015 semester.



Population	Mean
Age in years	22.172

a) What was the shape and mean average of the population?
b) If a random sample was taken from this population, what is the minimum sample size we should have in order to have a nearly normal sampling distribution for sample means?

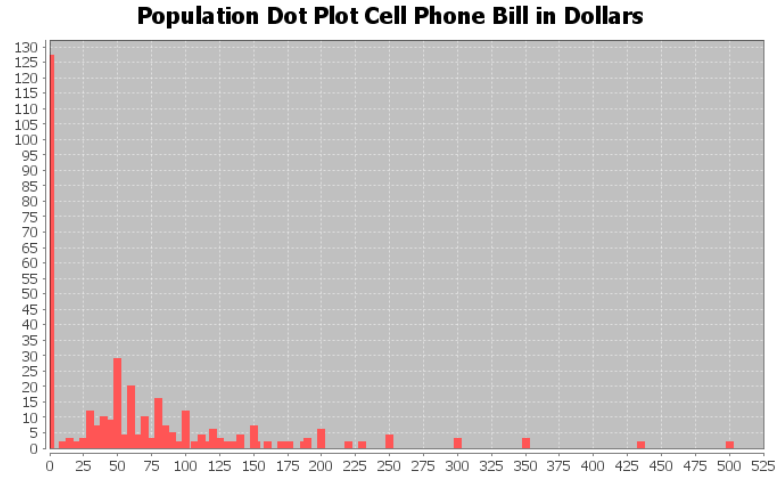
10. The following graph was created with StatKey from the "sleep hours per night" census data (Math 140 Survey Data). Assume this census represents the population of Math 140 statistics students at College of the Canyons in the fall 2015 semester.



Population	Mean
Sleep in hours per night	6.759

a) What was the shape and mean average of the population?
b) If a random sample was taken from this population, what is the minimum sample size we should have in order to have a nearly normal sampling distribution for sample means?

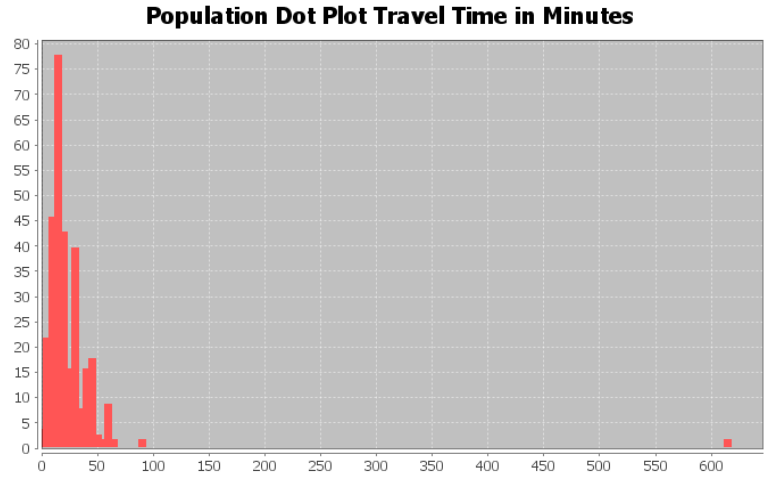
11. The following graph was created with StatKey from the cell phone bill (in dollars per month) census data (Math 140 Survey Data). Assume this census represents the population of Math 140 statistics students at College of the Canyons in the fall 2015 semester.



Population	Mean
Cell Phone Bill in Dollars per month	55.014

a) What was the shape and mean average of the population?
b) If a random sample was taken from this population, what is the minimum sample size we should have in order to have a nearly normal sampling distribution for sample means?

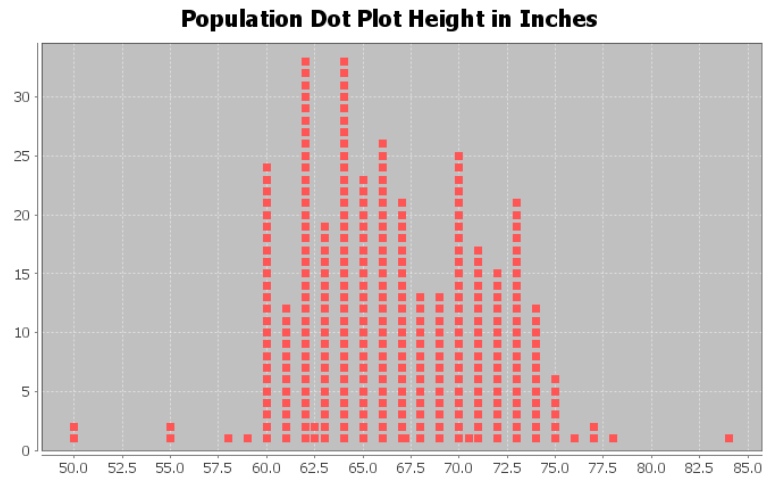
12. The following graph was created with StatKey from the travel time to school in minutes census data (Math 140 Survey Data). Assume this census represents the population of Math 140 statistics students at College of the Canyons in the fall 2015 semester.



Population	Mean
Travel Time to School in Minutes	22.742

a) What was the shape and mean average of the population?
b) If a random sample was taken from this population, what is the minimum sample size we should have in order to have a nearly normal sampling distribution for sample means?

13. The following graph was created with StatKey from the height in inches census data (Math 140 Survey Data). Assume this census represents the population of Math 140 statistics students at College of the Canyons in the fall 2015 semester.



Population	Mean
Height in Inches	66.511

a) What was the shape and mean average of the population?
b) If a random sample was taken from this population, what is the minimum sample size we should have in order to have a nearly normal sampling distribution for sample means?

14. A census of COC statistics students in the fall 2015 semester indicated that the population proportion of statistics students with brown hair is 0.537. Use this population proportion (π) to answer the following questions.

a) Use the formula n = $\frac{10}{(\pi)}$ to calculate the minimum sample size to get at least ten successes in our sample data.
b) Use the formula n = $\frac{10}{(1-\pi)}$ to calculate the minimum sample size to get at least ten failures in our sample data.
c) If our sample data has at least ten successes and at least ten failures, then we expect the sampling distribution for sample proportions to be approximately normal. What is the minimum sample size we expect to have a nearly normal sampling distribution for sample proportions?

15. A census of COC statistics students in the fall 2015 semester indicated that the population proportion of statistics students that smoke cigarettes is 0.091. Use this population proportion (π) to answer the following questions.

a) Use the formula n = $\frac{10}{(\pi)}$ to calculate the minimum sample size to get at least ten successes in our sample data.
b) Use the formula n = $\frac{10}{(1-\pi)}$ to calculate the minimum sample size to get at least ten failures in our sample data.
c) If our sample data has at least ten successes and at least ten failures, then we expect the sampling distribution for sample proportions to be approximately normal. What is the minimum sample size we expect to have a nearly normal sampling distribution for sample proportions?

16. Approximately 60% of college students in the U.S. were able to finish their bachelor's degree in six years. Use this population proportion (π) to answer the following questions.

a) Use the formula n = $\frac{10}{(\pi)}$ to calculate the minimum sample size to get at least ten successes in our sample data.
b) Use the formula n = $\frac{10}{(1-\pi)}$ to calculate the minimum sample size to get at least ten failures in our sample data.
c) If our sample data has at least ten successes and at least ten failures, then we expect the sampling distribution for sample proportions to be approximately normal. What is the minimum sample size we expect to have a nearly normal sampling distribution for sample proportions?

17. Approximately 9.4% of all adults in the U.S. have diabetes. Use this population proportion (π) to answer the following questions.

a) Use the formula n = $\frac{10}{(\pi)}$ to calculate the minimum sample size to get at least ten successes in our sample data.
b) Use the formula n = $\frac{10}{(1-\pi)}$ to calculate the minimum sample size to get at least ten failures in our sample data.
c) If our sample data has at least ten successes and at least ten failures, then we expect the sampling distribution for sample proportions to be approximately normal. What is the minimum sample size we expect to have a nearly normal sampling distribution for sample proportions?

18. Approximately 90% of all lung cancer cases are caused by cigarette smoking. Use this population proportion (π) to answer the following questions.

a) Use the formula n = $\frac{10}{(\pi)}$ to calculate the minimum sample size to get at least ten successes in our sample data.
b) Use the formula n = $\frac{10}{(1-\pi)}$ to calculate the minimum sample size to get at least ten failures in our sample data.
c) If our sample data has at least ten successes and at least ten failures, then we expect the sampling distribution for sample proportions to be approximately normal. What is the minimum sample size we expect to have a nearly normal sampling distribution for sample proportions?

19. Approximately 10% of all people are left handed. Use this population proportion (π) to answer the following questions.

a) Use the formula n = $\frac{10}{(\pi)}$ to calculate the minimum sample size to get at least ten successes in our sample data.
b) Use the formula n = $\frac{10}{(1-\pi)}$ to calculate the minimum sample size to get at least ten failures in our sample data.
c) If our sample data has at least ten successes and at least ten failures, then we expect the sampling distribution for sample proportions to be approximately normal. What is the minimum sample size we expect to have a nearly normal sampling distribution for sample proportions?

20. Approximately 2% of all vehicles sold in the U.S have a manual transmission. Use this population proportion (π) to answer the following questions.

a) Use the formula n = $\frac{10}{(\pi)}$ to calculate the minimum sample size to get at least ten successes in our sample data.
b) Use the formula n = $\frac{10}{(1-\pi)}$ to calculate the minimum sample size to get at least ten failures in our sample data.
c) If our sample data has at least ten successes and at least ten failures, then we expect the sampling distribution for sample proportions to be approximately normal. What is the minimum sample size we expect to have a nearly normal sampling distribution for sample proportions?

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 < π < 0.329 or 30.5% < π < 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 <u>not</u> 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?

<u>Confidence Levels</u>

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.

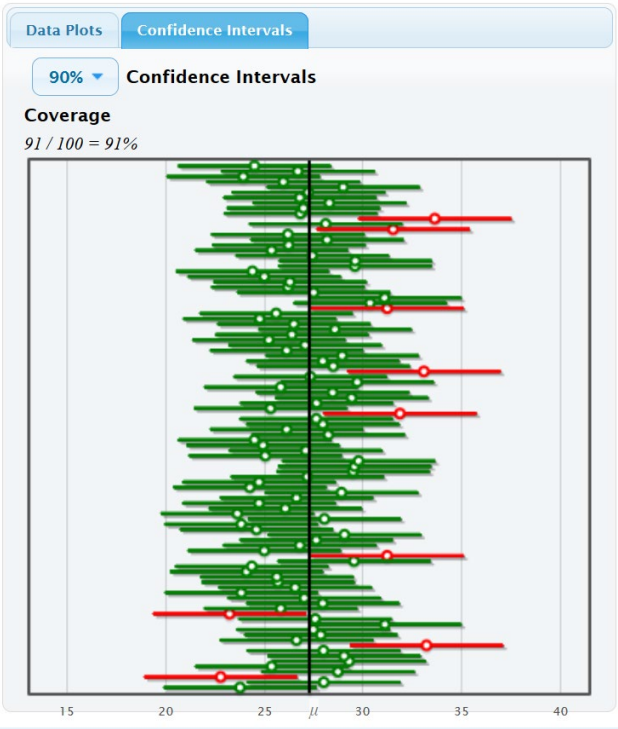
99% Confident: 99% of confidence intervals contain the population value and 1% of confidence intervals do not contain the population value.

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

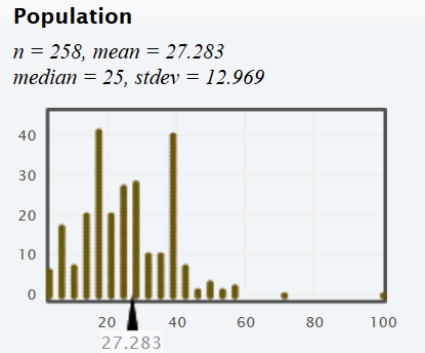
Example 1: 90% confidence level sampling distribution

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.



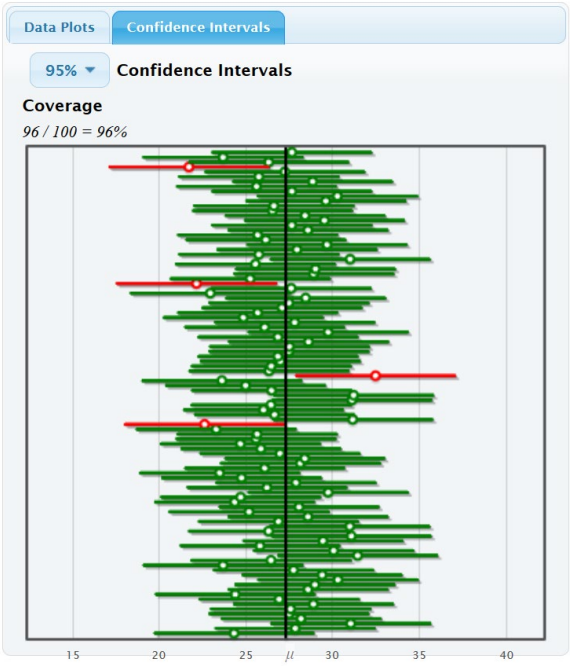
Notice when we use a 90% confidence level, about 90% of them were green (contained the population parameter) and about 10% of them were red (did not contain the population parameter). In other words, not all confidence intervals contain the population parameter! This is what the definition of 90% confidence 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

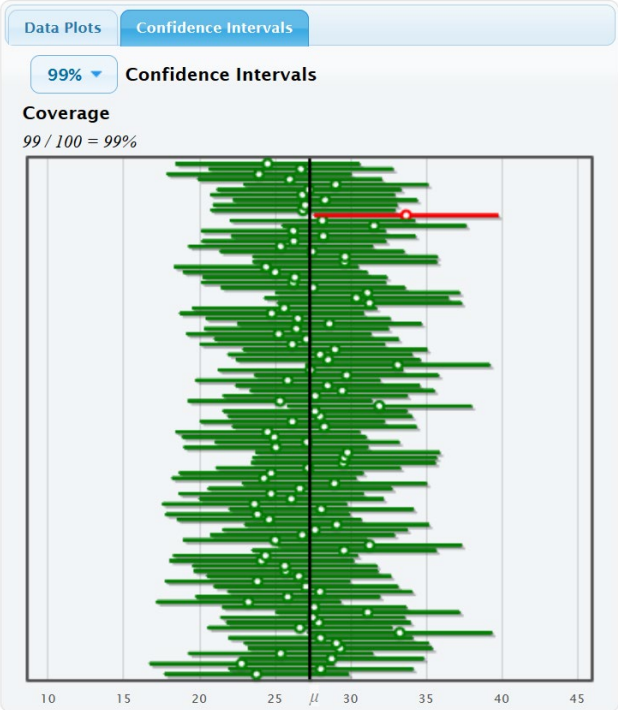
Work Hours per Week for working COC Statistics Students (Fall 2015 semester)



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.

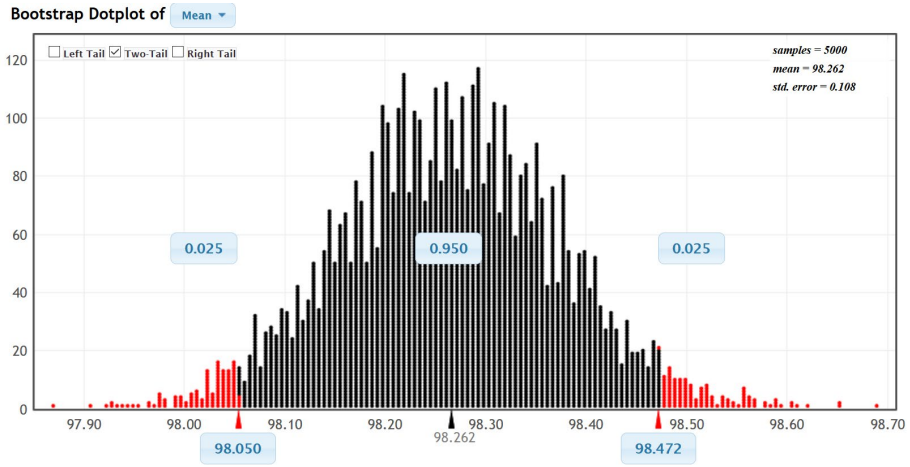
Confidence Interval Back-Solving Formula for Sample Statistic: $Sample\ Statistic = \frac{(Upper\ Limit + Lower\ Limit)}{2}$

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

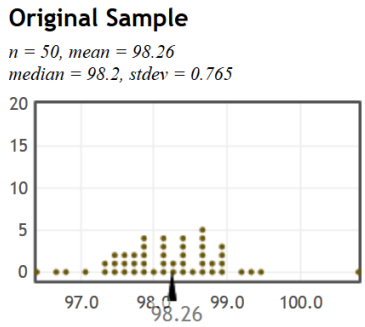
Example 1: Bootstrap Confidence Interval for Population Mean Average Body Temperature

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.

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

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

Problem Set Section 2D

(For #1-10) Add and subtract the given sample statistic and margin of error to find the confidence interval estimate of the population value. Then write the confidence interval using both inequality notation and using interval notation. Now write a sentence explaining the confidence interval to someone.

Confidence interval = Sample Statistic ± Margin of Error

1. "What is the population percent of the adult population is infected with this disease?"
   Sample percentage = 4.9%
   Margin of error = 1.3% (Found with 95% confidence level.)

2. "What is the population mean average height for men?"
   Sample mean = 68.335 inches
   Margin of error = 1.293 inches (Found with 99% confidence level.)

3. What is the population standard deviation for the systolic blood pressure in women?
   (Assume there was a normal sampling distribution.)
   Sample standard deviation = 17.11 mm of Hg
   Margin of error = 3.31 mm of Hg (Found with 90% confidence level.)

4. What is the population percentage of left-handed people get migraine headaches?
   Sample proportion = 0.088
   Margin of error = 0.027 (Found with 95% confidence level.)

5. What is the population mean average price of a used mustang car in thousands of dollars?
   Sample mean = 15.98 thousand dollars
   Margin of error = 3.78 thousand dollars (Found with 90% confidence level.)

6. "What is the population percentage of rabid animals are wild?"
   Sample proportion = 0.903
   Margin of error = 0.008 (Found with 95% confidence level.)

7. "What is the population mean average weight for men?"
   Sample mean = 172.55 pounds
   Margin of error = 11.272 pounds (Found with 99% confidence level.)

8. What is the population variance for the heights of men? (Assume there was a normal sampling distribution.)
   Sample variance = 10.177 square inches
   Margin of error = 3.661 square inches (Found with 90% confidence level.)

9. What is the population percentage of women in the U.S. are overweight?
   Sample percentage = 36.9%
   Margin of error = 1.44% (Found with 95% confidence level.)

10. What is the population mean average amount of tip in dollars at a particular restaurant?
    Sample mean = $3.849
    Margin of error = $0.504 (Found with 99% confidence level.)

(For #11-20) Write a sentence explaining each of the following confidence intervals. Then use the following formulas to identify the sample statistic ($\hat{p}$ or $\bar{x}$ or s) and the margin of error.

Sample Statistic = $\frac{(upper\ limit + lower\ limit)}{2}$ Margin of Error = $\frac{(upper\ limit - lower\ limit)}{2}$

11. A 95% confidence interval estimate of the population proportion of fat in the milk from Jersey cows is (0.046 , 0.052).

12. A 99% confidence interval estimate of the population mean number of miles is 13.4 miles < μ < 17.2 miles.

13. A 90% confidence interval estimate of the population proportion of people who will vote for the Independent party candidate is 0.068 < π < 0.083.

14. A 95% confidence interval estimate of the population mean amount of milk in gallons is (48.7 , 58.4).

15. A 99% confidence interval estimate of the population standard deviation for the height of men in inches is 2.34 < σ < 2.87. Assume there was a normal sampling distribution.

16. A 95% confidence interval estimate of the population proportion of peanuts in a can of mixed nuts is 0.4221 < π < 0.6179.

17. A 99% confidence interval estimate of the population mean pH of lakes in Florida is (6.118 , 7.064).

18. A 90% confidence interval estimate of the population proportion of home teams that win a soccer game is (0.5093 , 0.6574)

19. A 95% confidence interval estimate of the population mean average price of apartments in Manhattan, NY is $2514.36 < μ < $3798.64.

20. A 90% confidence interval estimate of the population variance for the pH of lakes in Florida is 1.2353 < σ^2 < 2.3675. Assume there was a normal sampling distribution.

(#21-26) Go to www.matt-teachout.org and click on "statistics" and then "data sets". Open the "coffee data" and copy and Columbian Mild price data. Now go to www.lock5stat.com and click on the "StatKey" button. Under the "sampling distribution" menu click on "mean". Under edit data, paste the Columbian Mild coffee data. Click on "samples of size n" and put in 30. Turn off the button that says, "First column is identifier" as we have only a single column of data. Now click ok. You are now ready to create your sampling distribution. This time we want the computer to create a confidence interval for each sample it takes. On the right side of the screen, click on the button that says confidence intervals and set the confidence level to 95%. StatKey will take a random sample from the population data, find the sample mean and place a dot for the sample mean in the distribution. It will also create a confidence interval from that sample mean. StatKey will keep track of whether the true population mean is actually contained in the confidence interval or not. Green means the confidence interval did contain the population value and red means that the confidence interval did not contain the population value. Now answer the following questions.

21. Notice the confidence intervals for sample means were different for each random sample. Discuss the implications of sampling variability on the accuracy of a confidence interval from a random sample.

22. What was the population mean in cents per pound? Did all the confidence intervals contain the population mean? What does it mean that the interval "contained" or "captured" the population mean?

23. How many total random samples did you take? How many of them contained the population mean? What percent of the confidence intervals contained the population mean?

24. How many confidence intervals did not contain the population mean? What percent of the confidence intervals did not contain the population mean?

25. As the number of random samples increased, did the percentage of confidence intervals that contained the population mean get closer or farther away from 95%? Why do you think that is?

26. Here is the definition of 95% confidence: "95% of confidence intervals contain the population parameter and 5% do not contain the population parameter". Explain this definition of 95% confidence in your own words.

(#27-32) Assume a fair coin has a 50% (0.5) chance of getting tails. If we take samples from that population, the sample proportions will usually not be 0.5. We want to look at lots of proportion confidence intervals from sample proportions. Go to www.lock5stat.com and click on the "StatKey" button. Under the "sampling distribution" menu click on "proportion". Under "edit proportion", put in 0.5 and then click ok. Under "sample size", set it to "n = 30". You are now ready to create your sampling distribution. We want the computer to create a confidence interval for each sample proportion. On the right side of the screen, click on the button that says confidence intervals and set the confidence level to 90%. StatKey will take a random sample from the population data, find the sample proportion and place a dot for the sample proportion in the distribution. It will also create a confidence interval from that sample. Remember that the population proportion for a fair coin is 0.5 (50%). StatKey will keep track of whether the true population proportion is actually contained in the confidence interval or not. Green means the confidence interval did contain the population value and red means that the confidence interval did not contain the population value. Now answer the following questions.

27. Notice the confidence intervals for sample proportions were different for each random sample. Discuss the implications of sampling variability on the accuracy of a confidence interval created from a random sample proportion.

28. Did all the confidence intervals contain the population proportion of 0.5? What does it mean that the interval "contained" or "captured" the population parameter?

29. How many total confidence intervals did you make? How many of them contained the population proportion 0.5? What percent of the confidence intervals contained the population proportion 0.5?

30. How many of the confidence intervals did not contain the population proportion 0.5? What percent of the confidence intervals did not contain the population proportion 0.5?

31. As the number of random samples increased, did the percentage of confidence intervals that contained the population proportion get closer or farther away from 90%? Why do you think that is?

32. Here is the definition of 90% confidence: "90% of confidence intervals contain the population parameter and 10% do not contain the population parameter". Explain this definition of 90% confidence in your own words.

Section 2E – One Population Mean & Proportion 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.

Sampling Distribution: Take many random samples from a population, calculate a sample statistic like a mean or percent from each sample and graph all of the sample statistics on the same graph. The center of the sampling distribution is a good estimate of the population parameter.

Sampling Variability: Random samples values and sample statistics are usually different from each other and usually different from the population parameter.

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. Can be calculated by either a bootstrap distribution or by adding and subtracting the sample statistic and the margin of error.

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.

Bootstrapping: Taking many random samples values from one original real random sample with replacement.

Bootstrap Sample: A simulated sample created by taking many random samples values from one original real random sample with replacement.

Bootstrap Statistic: A statistic calculated from a bootstrap sample.

Bootstrap Distribution: Putting many bootstrap statistics on the same graph in order to simulate the sampling variability in a population, calculate standard error, and create a confidence interval. The center of the bootstrap distribution is the original real sample statistic.

In the last section, we saw that if we have only one random sample from a population, we would not be able to find the population parameter exactly. The best we can do is create a confidence interval, which is two numbers that we think the population parameter is in between.

In this section, we will look at some of the famous formulas that statisticians use to estimate population parameters with confidence intervals. We will also look at sample data conditions in order to ensure the accuracy of the formula.

If our sampling distribution is normal, most one-population confidence interval formulas start from the following.

Sample Statistic ± Margin of Error

Early mathematicians and statisticians thought a lot about how to estimate the margin of error when you do not know the population parameter. The key was the sampling distribution. If a sampling distribution looked normal, then the empirical rule would suggest that the middle 95% would correspond to two standard deviations above and below the center. This gave rise to another common formula.

Sample Statistic ± (2 × Standard Error)

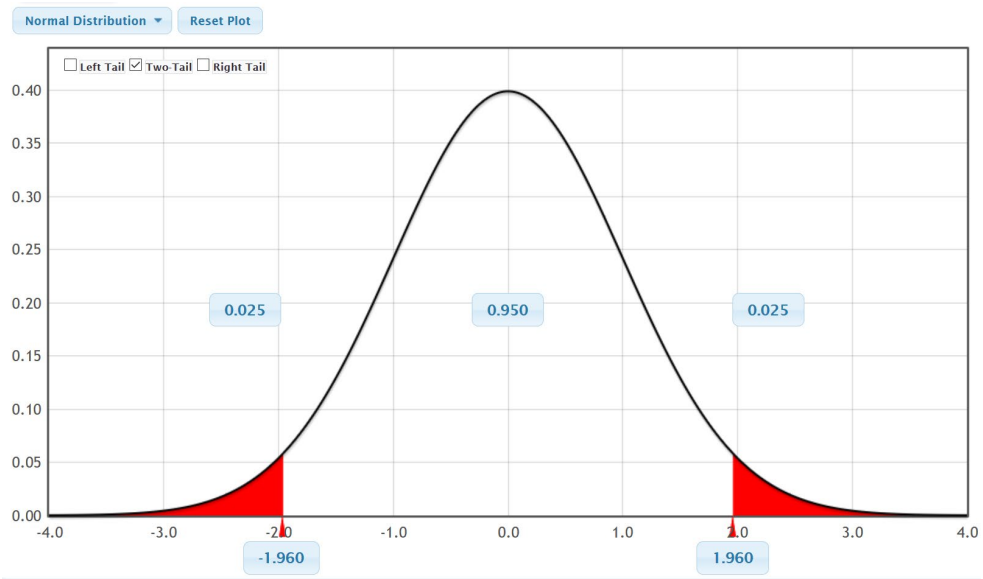
Critical Value Z-scores

What is the “2” representing in the following formula. It seems it is counting how many standard errors one is from the mean (center) of the sampling distribution. Does this remind you of a statistic we previously learned?

If you recall, the Z-score measures the number of standard deviations from the mean. So the “2” is really a Z-score. This gave rise to the idea of replacing the “2” with a Z-score. The Z-score can be adapted for 90%, 95% or 99%. Remember two standard deviation is just an approximation for 95%. If that is the case, can we get a more accurate number for 95%?

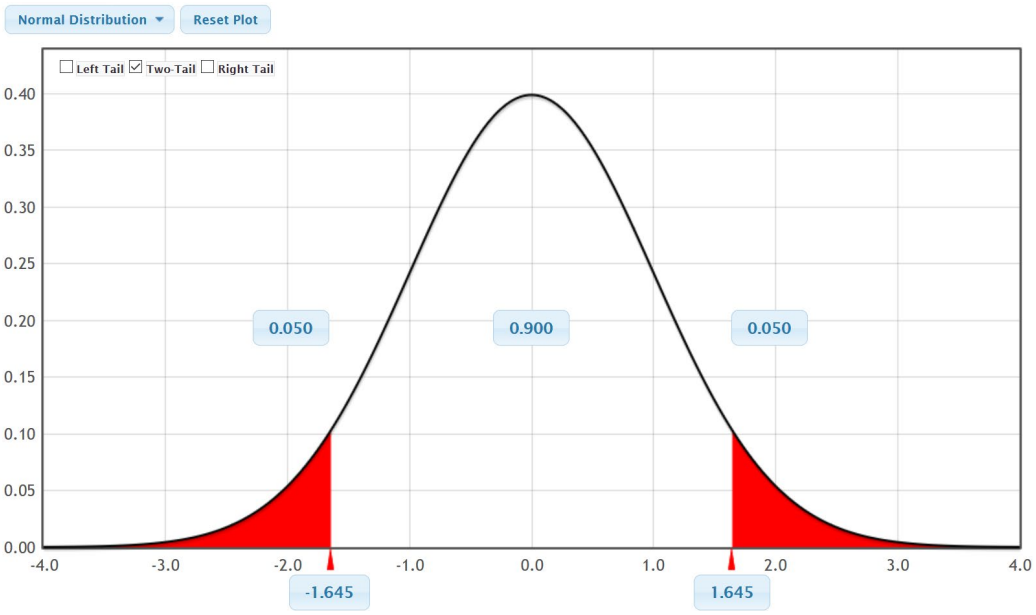
Using a normal calculator, we can calculate the Z-score for 90%, 95% and 99% confidence. These are very famous and are often referred to as “critical value Z-scores” or "Z_c" for short.

Go to www.lock5stat.com and open StatKey. Under the “theoretical distributions” menu, click on “normal”. If the mean is zero and the standard deviation is one, then this will calculate Z-scores. Click the “two-tail” button. The first Z-score calculated is for 95%.



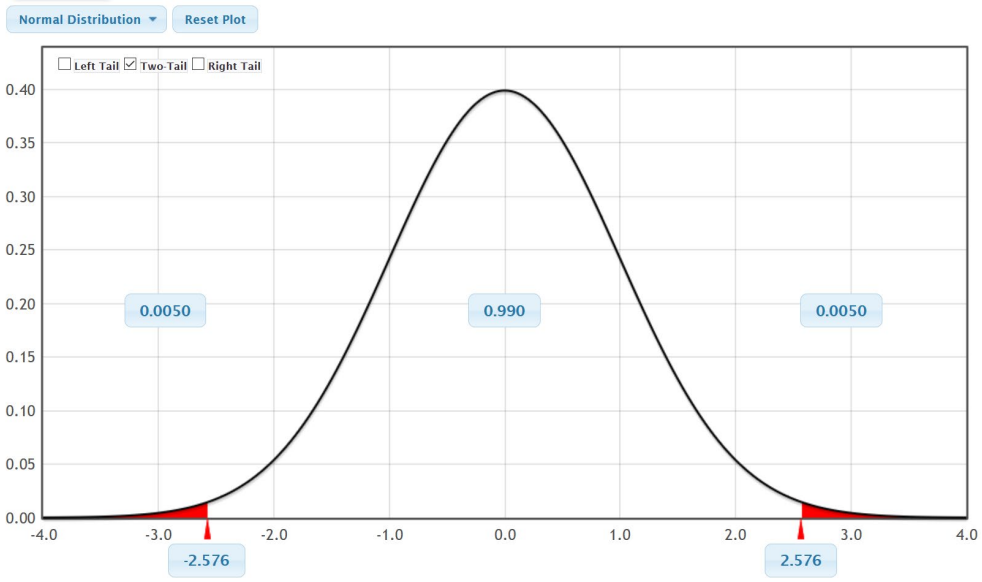
This is the most famous of all the critical value Z-scores. Remember, for the middle 95%, the empirical rule indicates that it will be “about” two standard deviations. It turns out, 1.96 standard deviations is more accurate. Notice that just like the confidence intervals have an upper limit and lower limit, so does the Z-score critical values.
For 95% confidence, we can replace the ±2 in the formula with ±1.96.

What about 90% confidence intervals? Go back to the normal calculator in StatKey and click on the “0.95” in the middle. Change it to 0.9 (90%).



Notice the Z-score for 90% confidence intervals is ±1.645. Notice that as the confidence interval decreases from 95% to 90%, the Z-score gets lower. This will cause the margin of error to decrease and the confidence interval to get narrower.

What about 99% confidence intervals? Go back to the normal calculator in StatKey and change the middle proportion into 0.99 (99%).



Notice the Z-score for 99% confidence intervals is ±2.576. Therefore, instead of being 1.645 standard errors away or 1.96 standard errors away, now we are 2.576 standard errors away. As the confidence interval increases from 95% to 99%, the Z-score gets larger. This will cause the margin of error to increase and the confidence interval to get wider.

Here are the famous critical value Z-scores.

- 90% confidence level: Z = ± 1.645
- 95% confidence level: Z = ± 1.96
- 99% confidence level: Z = ± 2.576

Let us summarize the progress of our one-population confidence interval formula. It is important to remember that these formulas only work if our sampling distribution looks normal. Z-scores calculate the number of standard deviations (standard errors) from the mean in a perfectly normal curve.

Sample Statistic ± Margin of Error

Sample Statistic ± (2 × Standard Error)

Sample Statistic ± (Z × Standard Error)

Statisticians discovered that as long as the sampling distribution was normal, the Z-scores were accurate for proportion (percentage) confidence intervals. The famous critical value Z-scores are still used to this day to calculate a confidence interval estimate of a population proportion (percentage).

One-Population Proportion Confidence Interval

Before computers were invented, it was very difficult to make sampling distributions. Yet it was vital to understanding sample statistics and calculating standard error. Early mathematicians and statisticians invented formulas to estimate the standard error.

Standard Error Estimation Formula for Proportions = $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Sample Proportion = $\hat{p}$
Sample Size = n

So now, we can finish our estimation formula for a confidence interval estimate of the population proportion. In order to estimate the margin of error, we multiply the standard error by the number of standard errors (Z-score).

Sample Statistic ± Margin of Error
Sample Statistic ± (2 × Standard Error)
Sample Statistic ± (Z × Standard Error)

$\hat{p} \pm \left(Z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$

Example 1: Calculating the confidence interval for a proportion

A random sample of 54 bears in a region of California showed that 19 of them were female. Find the sample proportion and use the formula above to calculate a 95% confidence interval estimate for the population proportion of female bears in this region of California.

Sample Proportion $(\hat{p})$ = $\frac{Amount\ of\ Female\ Bears}{Sample\ Size}$ = $\frac{19}{54} \approx 0.352$

Critical Value Z-score for 90% Confidence = ± 1.96

Now we will replace the Z-score with 1.96 and $\hat{p}$ with 0.352 and *n* with 54 into our formula and work it out. Remember to follow order of operations. Notice the standard error estimate is 0.065 (6.5%) and the margin of error estimate is 0.127 (12.7%).

$\hat{p} \pm \left(Z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$

$0.352 \pm \left(1.96\sqrt{\frac{0.352(1-0.352)}{54}} \right)$

$0.352 \pm \left(1.96\sqrt{\frac{0.352(0.648)}{54}} \right)$

$0.352 \pm (1.96 \times 0.065)$

$0.352 \pm (0.127)$

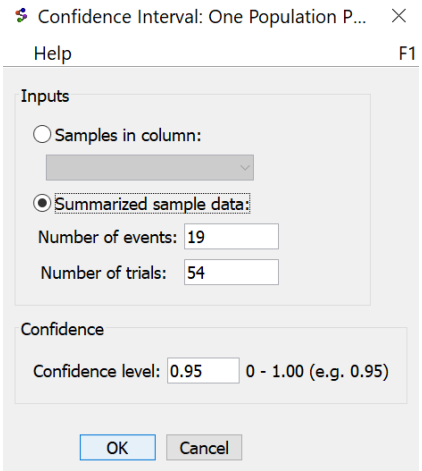
0.352 – 0.127 < Population Proportion of Female Bears (π) < 0.352 + 0.127

0.225 < Population Proportion of Female Bears (π) < 0.479

We are 95% confident that between 22.5% and 47.9% of all bears in this region of California are female.

Note: While it is important to understand formulas, data scientist today rely on computers to calculate confidence intervals. It is very difficult to calculate confidence intervals from large data sets with a formula and a calculator. The job of a data scientist, statistician, or data analyst is understand and explain the data, not to spend hours calculating something a computer can do in a split second.

To calculate this confidence interval with Statcato, we will click on the "statistics" menu and then "confidence intervals". Click on one-population proportion and under summary data; enter 19 for the number of events and 54 for the number of trials. Set the confidence level to 0.95 and click OK.



Here is the Statcato printout. Notice the computer calculation is almost the same as the one we did with the formula and calculator. However, it took a lot less time.

Confidence Interval - One population proportion: confidence level = 0.95
Input: Summary data

Number of trials	Number of Events	Sample proportion	Margin of Error	95.0%CI
54	19	0.352	0.127	(0.2245, 0.4792)

Key Question: How accurate is this confidence interval?

This confidence interval relies on a Z-score and the standard error so the sampling distribution for sample proportions must be normal for this formula to be accurate. If we look at the section on the central limit theorem, we remember that for a sampling distribution for random sample proportions to be normal, we need at least ten success and at least ten failures. This gives rise to the assumptions or conditions required for certain confidence interval calculations. For the formula approach to be accurate, the following must be true. If any of these assumptions are not met, then the confidence interval may not be accurate.

One-population Proportion Assumptions

1. The categorical sample data should be collected randomly or be representative of the population.
2. Data values within the sample should be independent of each other.
3. There should be at least ten successes and at least ten failures.

Let us check these assumptions in the previous confidence interval for the proportion of female bears.

1. Random Categorical Data? *Yes. This data was random and gender is a categorical variable.*

2. Data values within the sample should be independent of each other. *This can be difficult to determine. It should not be the same bear measured multiple times. In addition, if one bear is female it should not change the probability of other bears being female. It is likely safe to assume these are true in this case.*

3. At least ten successes? *Yes. There were 19 female bears in the data, which is more than ten.*
   At least ten failures? *Yes. There were 54−19 = 35 bears that were not female in the data which is more than ten.*

Overall, it appears the data does satisfy the requirements for using the formula and so the confidence interval will be relatively accurate.

Bootstrapping

Is there a way to make a confidence interval if the data did not meet the assumptions?

It depends on which assumptions. One technique that is sometimes used is called “Bootstrapping”. Bootstrapping does require the sample to be representative of the population. That usually means it was collected randomly with data values that are independent of each other. As long as you have those two assumptions, you can bootstrap.

One-population Bootstrap Assumptions

1. The sample data should be collected randomly or be representative of the population.
2. Data values within the sample should be independent of each other.

Bootstrapping does not use formula for standard error and critical values like Z-scores or T-scores. It calculates the middle 95%, 99% or 90% directly using a bootstrap sampling distribution. Since bootstrapping is not tied to formulas and critical values, it does not require the sampling distribution to be normal or to match up with a specific theoretical curve.

The idea of bootstrapping is to create a theoretical population by assuming that the population is just many copies of your one real representative random sample. In practice, bootstrapping uses computers to take thousands for random samples with replacement from your one representative random sample. It randomly selects a value from your data, but puts the value back before picking another value randomly. This allows us to get the same value in a bootstrap sample multiple times. It then calculates the statistic like the mean or proportion from all of the bootstrap samples. These are sometimes called "bootstrap statistics". Putting all the bootstrap statistics on the same graph gives a "bootstrap sampling distribution". If you find the computer find the cutoffs for the middle 95% of the bootstrap distribution, you have an estimated 95% confidence interval.

Bootstrapping: Taking many random samples values from one original real random sample with replacement.

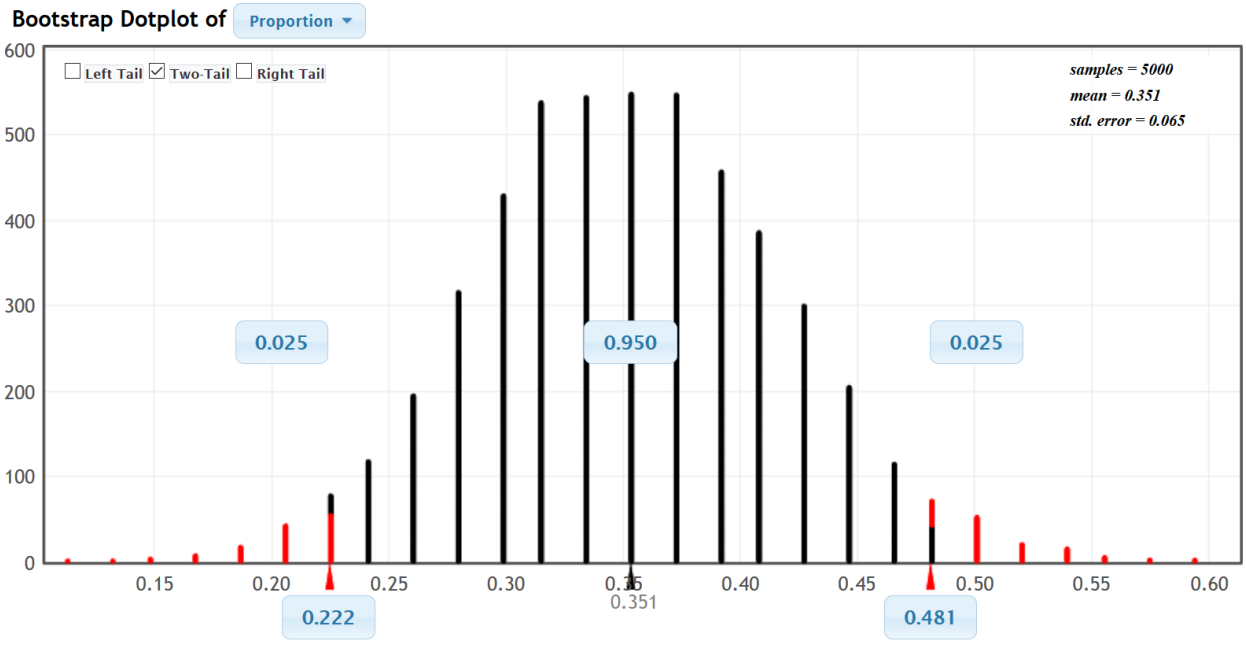
Bootstrap Sample: A simulated sample created by taking many random samples values from one original real random sample with replacement.

Bootstrap Statistic: A statistic calculated from a bootstrap sample.

Bootstrap Distribution: Putting many bootstrap statistics on the same graph in order to simulate the sampling variability in a population, calculate standard error, and create a confidence interval. The center of the bootstrap distribution is the original real sample statistic.

Female Bears Example

In the last example, we used the traditional Z critical value and standard error formula to create a confidence interval and estimate the population percentage of bears that are female. We could also use a bootstrap. Go to the "Bootstrap Confidence Interval" menu in StatKey at www.lock5stat.com and click on "CI for Single Proportion". Under "Edit Data" put in the random sample data count (19 female bears) and the total sample size (54 bears). Click "Generate 1000 Samples" a few times. Now click "Two-Tail". The default is 95%, but you can always change the middle proportion to 99% (0.99) or 90% (0.90) if needed. This problem was a 95% confidence interval, so we will leave the middle proportion as 0.95.



In a bootstrap confidence interval, the upper and lower limit of the confidence interval are found at the bottom right and left (0.222 and 0.481). Using these numbers, we are 95% confident that the population percentage of bears in this region of California that are female is between 22.2% and 48.1%. Notice that the upper limit, lower limit and standard error are very close to what we got by formula or Statcato. Notice that the shape of the bootstrap distribution is very normal. Though the bootstrap does not give us the margin of error like Statcato, we can use the formula we learned in the previous section. Remember the standard error and margin of error in this calculation are

only reasonably accurate if the distribution is normal. Notice the margin of error is close to what we got by formula or Statcato.

Margin of Error = $\frac{(Upper\ Limit - Lower\ Limit)}{2} = \frac{(0.481-0.222)}{2} \approx 0.1295$

Key Notes about Bootstrapping

- A bootstrap distribution attempts to estimate and visualize the sampling variability in the population by creating a simulated population. Remember that standard error and margin of error are only accurate if the distribution is normal. So while we can estimate standard error and margin of error from a bootstrap, they may not be accurate if the bootstrap distribution is not normal.
- While a bootstrap distribution may be similar to a true sampling distribution from the population, there are important differences. The center of a bootstrap distribution is the sample statistic from the original real random data set. This makes the bootstrap ideal for estimating the confidence interval. A true sampling distribution is taking thousands of real samples from the population, so the center of a sampling distribution is the population parameter. We should not treat a true sampling distribution from the population the same as a bootstrap. If you have a sampling distribution, then the center can get a very accurate estimate of the population parameter. If you know the population parameter, you do not need a confidence interval. The middle 95% of a sampling distribution from an actual population is not a confidence interval.

Critical Value T-scores

In 1908, a statistician named William Gosset discovered that while Z-scores were very accurate for proportions, they were not very accurate when estimating mean averages, especially if the sample size was small. Small samples should have a larger margin of error than those indicated by Z-scores. To deal with this problem, he invented T-scores. His idea was that each sample size should have a different number of standard deviations. When Gosset invented the T-distribution, he worked for Guinness Beer and was not allowed to publish his work. He therefore published under the pseudonym "student". To this day, the T-distribution is often called the "Student T-Distribution" since it was invented by a then unknown author named "student".

T-scores are the same as Z-scores in the sense that they count the number of standard deviations or standard errors from the mean. However, they have a built in error correction for smaller data sets. For very large sample sizes, T-scores and Z-scores are about the same. For example, if we are using a 95% confidence level and our sample size is very large, then the T-score will be close to the Z-score of ± 1.96 standard deviations. When sample sizes are small, the T-scores become significantly greater than the Z-scores. This causes the margin of error to increase for small sample sizes. Remember, less random data should result in more error. We usually use Z-scores when estimating population proportions or percentages. We prefer to use T-scores when estimating population mean averages.

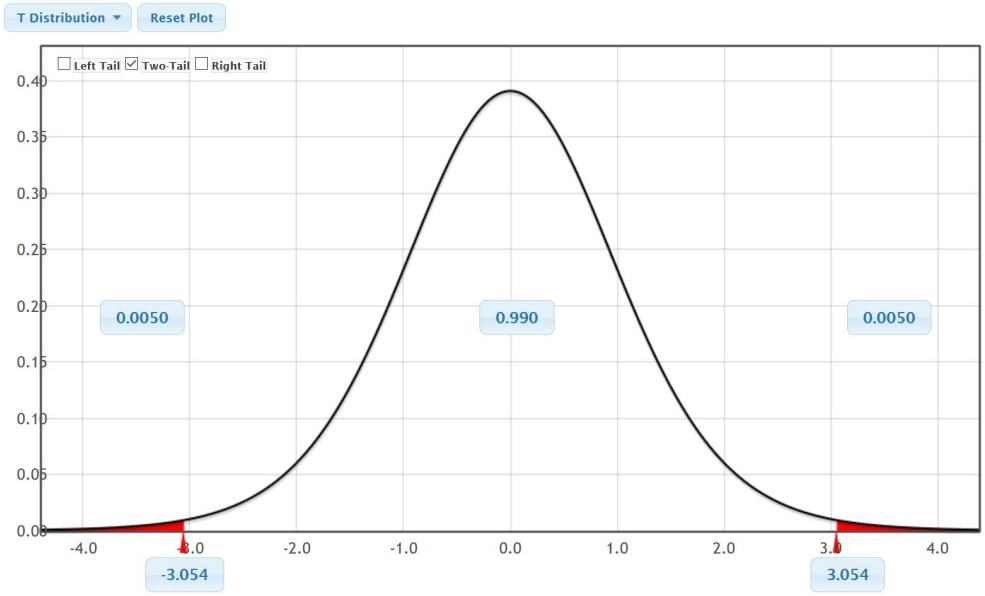
Note: You can use Z-scores for the mean if the sample size is large or if you know the population standard deviation exactly. However, we rarely know the population standard deviation with any certainty, especially when we do not even know the population mean. Also in large sample sizes, the T-scores are still accurate, so you might as well use the T-scores.

Degrees of Freedom

If you recall from previous sections, statistics like variance and standard deviation are based on a sum of squares divided by the degrees of freedom. For one sample, the degrees of freedom is usually equal to one less than the sample size $(df = n - 1)$. Because of this, Gosset organized his T-scores not by sample size, but by degrees of freedom. Gosset calculated his T-scores with calculus and wrote them on charts. Before computers were invented, a statistician would first calculate the degrees of freedom and then look up the correct T-score on these charts. In modern times, T-scores can be easily calculated with computer programs like StatKey.

Example 1: Calculate the T-score critical value for a sample size n = 13 and a 99% confidence level.

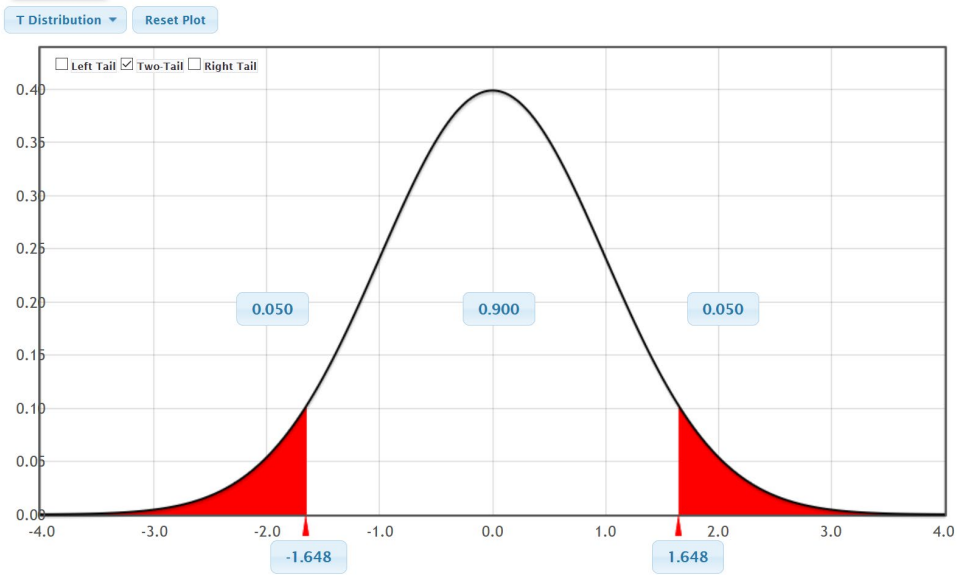
Go to www.lock5stat.com and click on "StatKey". Under the "theoretical distributions" menu, click on "t". Since the sample size is 13, the degrees of freedom will be df = 13−1 = 12. If we click on "two tail" and set the middle proportion to 0.99, we will get the following.



We see from the graph that critical value T-score for 99% confidence and 12 degrees of freedom is ± 3.054. Notice this is larger than the 99% confidence critical value Z-score (± 2.576). For smaller sample sizes, the T-scores are significantly greater than the Z-scores.

Example 2: Calculate the T-score critical value for a sample size n = 500 and a 90% confidence level.

Go to www.lock5stat.com and click on "StatKey". Under the "theoretical distributions" menu, click on "t". Since the sample size is 13, the degrees of freedom will be df = 500−1 = 499. If we click on "two tail" and set the middle proportion to 0.9, we will get the following.



We see from the graph that critical value T-score for 90% confidence and 499 degrees of freedom is ± 1.648. Notice this is very close to the 90% confidence critical value Z-score (± 1.645). For larger sample sizes, the T-scores and the Z-scores are about the same.

Summary of Critical Value T-scores

- T-scores (like Z-scores) count the number of standard deviations from the mean. In a sampling distribution of sample means, it counts how many standard errors we should be from the center of the sampling distribution for a given confidence level.
- T-scores are different for every sample size. They are usually organized by degrees of freedom. For one-population, the degrees of freedom is usually $df = n - 1$.
- T-scores are significantly larger than Z-scores for small sample sizes. The smaller the sample size, the larger the discrepancy between the T-score and Z-score.
- T-scores are about the same as Z-scores for large sample sizes.

<u>One-Population Mean Confidence Interval</u>

Let us look at the formula for calculating a one-population mean average confidence interval. Many computer programs to this day still use this formula.

Statisticians estimated the standard error for a sampling distribution for sample means with the following formula. The formula is surprisingly accurate and close to the standard error in an actual sampling distribution.

Standard Error Estimation Formula for Means = $\frac{s}{\sqrt{n}}$

Sample Standard Deviation = *s*
Sample Size = *n*

Here is the formula for a confidence interval estimate of the population mean. In order to estimate the margin of error, we multiply the standard error by the number of standard errors (T-score).

Sample Statistic ± Margin of Error
Sample Mean ± (T × Standard Error)

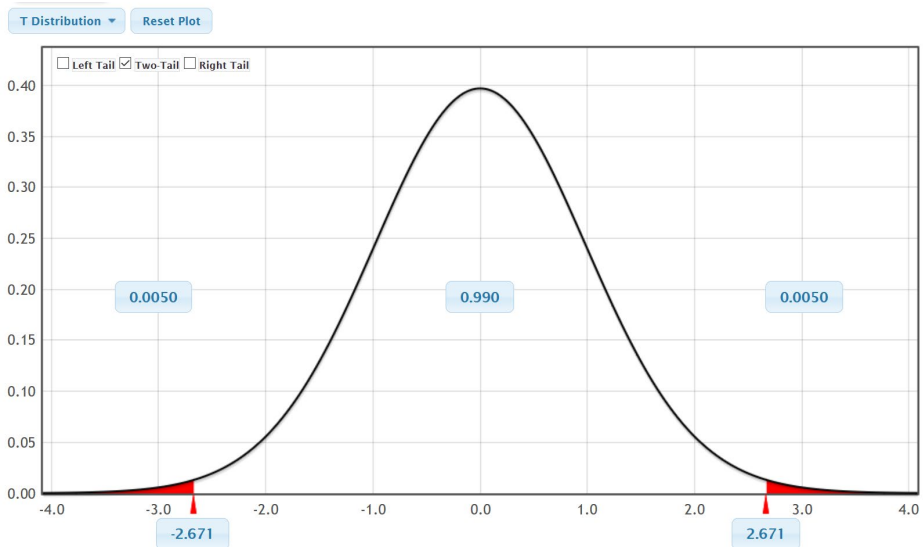
$\overline{x} \pm \left(T \frac{s}{\sqrt{n}}\right)$

Example 1: Calculating the confidence interval estimate of a population mean

A random sample of 54 bears in a region of California was taken. The weights of the bears showed a skewed right shape with a sample mean of 182.889 pounds and sample standard deviation of 121.801 pounds. Find the degrees of freedom and the critical value T-score. Then use the formula above to calculate a 99% confidence interval estimate for the population mean average weight of bears in this region of California.

Degrees of Freedom: df = n – 1 = 54 – 1 = 53.

Using the T-score calculator in StatKey we found that the critical Value T-score for 99% Confidence and 53 degrees of freedom is T = ± 2.671



Now we will replace the T-score with 2.671, $\bar{x}$ with 182.889, *n* with 54, and *s* with 121.801 into our formula and work it out. Remember to follow order of operations. Notice the standard error estimate is 16.575 pounds and the margin of error estimate is 44.272 pounds.

$\bar{x} \pm \left(T \frac{s}{\sqrt{n}}\right)$

$182.889 \pm 2.671 \times \frac{121.801}{\sqrt{54}}$

$182.889 \pm (2.671 \times 16.575)$

$182.889 \pm (44.272)$

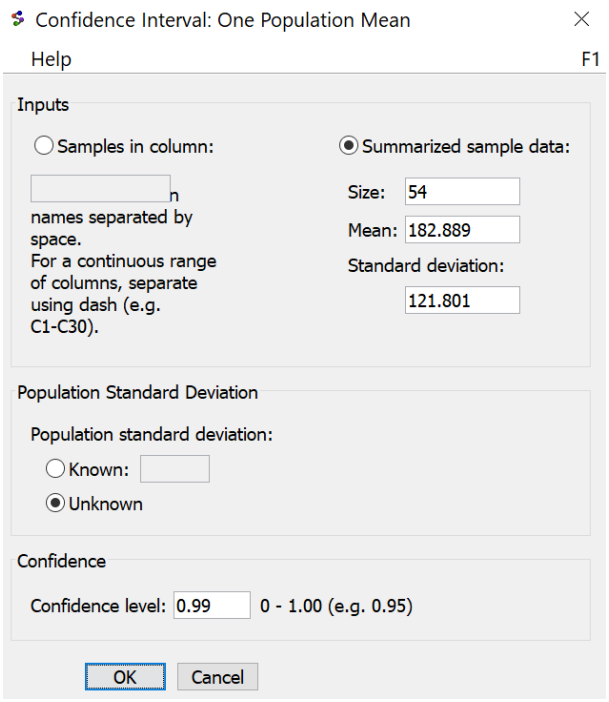
182.889 – 44.272 < Population Mean Average Weight of Bears in Pounds (μ) < 182.889 + 44.272

138.617 pounds < Population Mean Average Weight of Bears in Pounds (μ) < 272.161 pounds

We are 95% confident that the population mean average weight of bears in this region of California is in between 138.617 pounds and 272.161 pounds.

Note: While it is important to understand this formula, it is much easier to calculate this with a computer.

To calculate this confidence interval with Statcato, we will click on the "statistics" menu and then "confidence intervals". Click on "One-population mean". Under "Summary data", enter 182.889 for the mean, 121.801 for the standard deviation, and 54 for the number of trials. Set the confidence level to 0.99 and click OK. If we have the raw data, we could also put in the column "C1" where it says "samples in column".



Here is the Statcato printout. Notice the computer calculation is not exactly the same as the one we did with the formula and calculator. The computer did not round as much as we did. Computer calculations are usually much more accurate than calculator calculations because they tend to keep a lot more decimal places.

Confidence Interval - One population mean: confidence level = 0.99
Input: Summary data
σ unknown

Var	N	Mean	Stdev	Margin of Error	99.0%CI
summary	54.0	182.889	121.801	44.285	(138.6039, 227.1741)

It might be good to adjust our explanation sentence with the more accurate numbers from the computer.

We are 95% confident that the population mean average weight of bears in this region of California is in between 138.604 pounds and 272.174 pounds.

Key Question: How accurate is this confidence interval?

This confidence interval relies on a T-score and standard error so the sampling distribution for sample means must be normal for this formula to be accurate. If we look at the section on the central limit theorem, we remember that for a sampling distribution for random sample means to be normal, we need one of two things to be true. Either the data itself must be normal or the sample size must be at least 30. This gives rise to the assumptions or conditions required for mean average confidence interval calculations. For the formula approach to be accurate, the following must be true. If any of these assumptions are not met, then the confidence interval may not be accurate.

One-population Mean Assumptions

1. The quantitative sample data should be collected randomly or be representative of the population.
2. Data values within the sample should be independent of each other.
3. The sample size should be at least 30 or have a nearly normal shape.

Let us check these assumptions in the previous confidence interval for the mean average weight of bears.

1. Random Quantitative Data? *Yes. This data was random and weight in pounds is a quantitative variable.*

2. Data values within the sample should be independent of each other. *This can be difficult to determine. It should not be the same bear measured multiple times. These bears were probably tagged so they probably did not accidentally measure the same bear multiple times. Also, one bears weight should not change the probability of other bear having a certain weight. This data may not pass this assumption. Let us assume we see a bear that is eating well and is very heavy. Then there may be a higher probability of other bears being heavy in the same area.*

3. The sample data must be nearly normal or the sample size must be at least 30? *We see from the histogram that this data was skewed right, but the sample size was 54 (at least 30). Therefore, it does pass the 30 or normal requirement. Remember only one of the two need to be true for it to pass.*

The data did satisfy the random requirement and the at least 30 or normal requirement. If the data does satisfy the independence assumption, then the data would satisfy the overall requirements for using the formula and so the confidence interval will be relatively accurate.

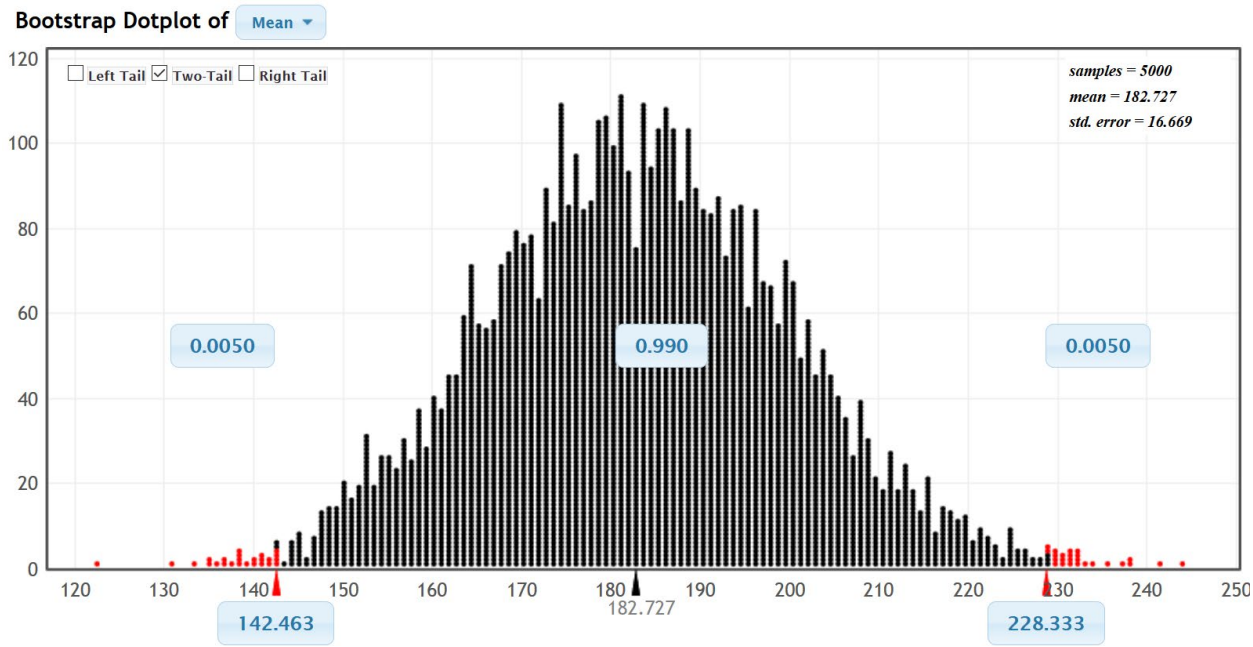
Could we have calculated this confidence interval with a bootstrap distribution?

Remember, the accuracy of a bootstrap is tied to the quality of the original sample data set. This data set was collected randomly but may fail the independence requirement.

Bear Weight Example

In this last example, we used the traditional T critical value and standard error formula to create a confidence interval and estimate the population mean average weight of bears. We could also use a bootstrap. First, go to the "Bear Data" at www.matt-teachout.org and copy the bear weight column of data. Now go to the "Bootstrap Confidence Interval" menu in StatKey at www.lock5stat.com and click on "CI for Single Mean, Median, St.Dev." Under "Edit Data", paste in the raw quantitative bear weight data. Make sure to check the "Header Row" box since this data set had a title and push "OK". Click "Generate 1000 Samples" a few times. Now click "Two-Tail".

The default is 95%, but you can change the middle proportion to 99%. This problem was a 99% confidence interval, so we will change the middle proportion to 99% (0.99).



We see that the bootstrap distribution is normally distributed. The confidence interval has a lower limit of 142.463 pounds, an upper limit of 228.333 pounds and a standard error of 16.669. Notice these numbers are relatively close to the same numbers we got by formula and Statcato. Since the confidence interval is normal, we can use the margin of error back-solving formula to find the approximate margin of error.

Margin of Error = $\frac{(Upper\ Limit - Lower\ Limit)}{2} = \frac{(228.333-142.463)}{2} \approx 42.935$

Problem Set Section 2E

Directions: Answer the following questions.

1. What are the assumptions necessary for making a one-population proportion confidence interval?

2. What are the assumptions necessary for making a one-population mean confidence interval?

3. What are the assumptions necessary for making a one-population bootstrap confidence interval?

4. An experiment was conducted to see what percentage of rats would show empathy toward fellow rats in distress. Of the 30 total rats in the study, 23 showed empathy. What was the sample proportion? What are the critical value Z-scores for 99% confidence? If you cannot remember them, open StatKey at www.lock5stat.com. Go to "theoretical distributions" and click on "normal". You can look up the critical value Z-scores. Use the critical values and the given standard error to calculate the margin of error and construct a 99% confidence interval estimate of the population proportion of rats that show empathy. Convert the upper and lower limits of your confidence interval into percentages.

Standard Error ≈ 0.07725

a) Sample Proportion $\hat{p} = \frac{Number\ of\ Success\ (events)}{Total\ Sample\ Size} =$

b) Critical value Z-scores = ±

c) Margin of Error = $Z \times Standard\ Error =$

d) Confidence Interval Lower Limit = $\hat{p} - (Margin\ of\ Error)$

e) Confidence Interval Upper Limit = $\hat{p} + (Margin\ of\ Error)$

5. Use the following Statcato printout to check your margin of error and confidence interval answers from the rat empathy data in number 4. Now check the assumptions and write sentences to explain the margin of error and confidence interval.

Number of trials	Number of Events	Sample proportion	Margin of Error	99.0%CI
30	23	0.767	0.199	(0.5678, 0.9656)

a) Check each of the assumptions for this problem. Assume the rats were randomly selected. Explain your answers.

b) Write a sentence to explain the margin of error in context.

c) Write a sentence to explain the confidence interval in context.

6. A study was done on the effectiveness of lie detector tests to catch someone that lies. In a random sample of 48 total lies, the machine identified only 31 of them. What was the sample proportion? What are the critical value Z-scores for 95% confidence? If you cannot remember them, open StatKey at www.lock5stat.com. Go to "theoretical distributions" and click on "normal". You can look up the critical value Z-scores. Use the critical values and the given standard error to calculate the margin of error and construct a 95% confidence interval estimate of the population proportion of lies caught be lie detector tests. Convert the upper and lower limits of your confidence interval into percentages.

Standard Error ≈ 0.0689

a) Sample Proportion $\hat{p} = \frac{Number\ of\ Success\ (events)}{Total\ Sample\ Size}$ =

b) Critical value Z-scores = ±

c) Margin of Error = $Z \times Standard\ Error$ =

d) Confidence Interval Lower Limit = $\hat{p} - (Margin\ of\ Error)$

e) Confidence Interval Upper Limit = $\hat{p} + (Margin\ of\ Error)$

7. Use the following Statcato printout to check your margin of error and confidence interval answers from the lie detector data in number 6. Now check the assumptions and write sentences to explain the margin of error and confidence interval.

Number of trials	Number of Events	Sample proportion	Margin of Error	95.0%CI
48	31	0.646	0.135	(0.5105, 0.7811)

a) Check each of the assumptions for this problem. Explain your answers.

b) Write a sentence to explain the margin of error in context.

c) Write a sentence to explain the confidence interval in context.

8. We want to determine what percentage of cereals the company Quaker makes. A random sample of 24 cereals found that Quaker made four of them. What was the sample proportion? What are the critical value Z-scores for 90% confidence? If you cannot remember them, open StatKey at www.lock5stat.com. Go to "theoretical distributions" and click on "normal". You can look up the critical value Z-scores. Use the critical values and the given standard error to calculate the margin of error and construct a 90% confidence interval estimate of the population proportion of cereals made by Quaker. Convert the upper and lower limits of your confidence interval into percentages.

Standard Error ≈ 0.076

a) Sample Proportion $\hat{p} = \frac{Number\ of\ Success\ (events)}{Total\ Sample\ Size}$ =

b) Critical value Z-scores = ±

c) Margin of Error = $Z \times Standard\ Error$ =

d) Confidence Interval Lower Limit = $\hat{p} - (Margin\ of\ Error)$

e) Confidence Interval Upper Limit = $\hat{p} + (Margin\ of\ Error)$

9. Use the following Statcato printout to check your margin of error and confidence interval answers from the cereal data in number 8. Now check the assumptions and write sentences to explain the margin of error and confidence interval.

Number of trials	Number of Events	Sample proportion	Margin of Error	90.0%CI
24	4	0.167	0.125	(0.0415, 0.2918)

a) Check each of the assumptions for this problem. Assume the cereal data was randomly selected. Explain your answers.

b) Write a sentence to explain the margin of error in context.

c) Write a sentence to explain the confidence interval in context.

10. If a cereal has more than 9 grams of sugar per serving, we consider it to have a high sugar content. We want to determine what percentage of cereals have a high sugar content. A random sample of 24 cereals found that 10 of them have a high sugar content. What was the sample proportion? What are the critical value Z-scores for 95% confidence? If you cannot remember them, open StatKey at www.lock5stat.com. Go to "theoretical distributions" and then click on "normal". You can look up the critical value Z-scores. Use the critical values and the given standard error to calculate the margin of error and construct a 95% confidence interval estimate of the population proportion of cereals made by Quaker. Convert the upper and lower limits of your confidence interval into percentages.

Standard Error ≈ 0.1006

a) Sample Proportion $\hat{p} = \frac{Number\ of\ Success\ (events)}{Total\ Sample\ Size}$ =

b) Critical value Z-scores = ±

c) Margin of Error = $Z \times Standard\ Error$ =

d) Confidence Interval Lower Limit = $\hat{p} - (Margin\ of\ Error)$

e) Confidence Interval Upper Limit = $\hat{p} + (Margin\ of\ Error)$

11. Use the following Statcato printout to check your margin of error and confidence interval answers from the cereal data in number 10. Now check the assumptions and write sentences to explain the margin of error and confidence interval.

Number of trials	Number of Events	Sample proportion	Margin of Error	95.0%CI
24	10	0.417	0.197	(0.2194, 0.6139)

a) Check each of the assumptions for this problem. Assume the cereal data was randomly selected. Explain your answers.

b) Write a sentence to explain the margin of error in context.

c) Write a sentence to explain the confidence interval in context.

12. A random sample of 45 high school students has a skewed left distribution. The sample mean average ACT exam score ($\bar{x}$) was 20.8 with a sample standard deviation of 9.868. What is the degrees of freedom? Open StatKey at www.lock5stat.com. Go to "theoretical distributions" and click on "T". Use the degrees of freedom and StatKey to look up the critical value T-scores for a 90% confidence level. Use the critical values and the given standard error to calculate the margin of error and construct a 90% confidence interval estimate of the population mean average ACT exam.

Standard Error = 1.471 ACT points

a) Degrees of Freedom = $n - 1$ =

b) Critical value T-scores = ±

c) Margin of Error = $T \times Standard\ Error$ =

d) Confidence Interval Lower Limit = $\bar{x} - (Margin\ of\ Error)$

e) Confidence Interval Upper Limit = $\bar{x} + (Margin\ of\ Error)$

13. Use the following Statcato printout to check your margin of error and confidence interval answers from the ACT data in number 12. Now check the assumptions and write sentences to explain the margin of error and confidence interval.

Var	N	Mean	Stdev	Margin of Error	90.0%CI
summary	45.0	20.8	9.868	2.472	(18.3284, 23.2716)

a) Check each of the assumptions for this problem. Explain your answers.

b) Write a sentence to explain the margin of error in context.

c) Write a sentence to explain the confidence interval in context.

14. A random sample of body temperatures in degrees Fahrenheit was taken from 50 randomly selected adults. The sample mean temperature of 98.26 °Fand a standard deviation of 0.765 °F. What is the degrees of freedom? Open StatKey at www.lock5stat.com. Go to "theoretical distributions" and click on "T". Use the degrees of freedom and StatKey to look up the critical value T-scores for a 95% confidence level. Use the critical values and the given standard error to calculate the margin of error and construct a 95% confidence interval estimate of the population mean average body temperature.

Standard Error = 0.1082 °F

a) Degrees of Freedom = $n - 1$ =

b) Critical value T-scores = ±

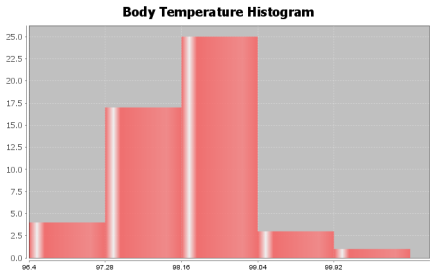
c) Margin of Error = $T \times Standard\ Error$ =

d) Confidence Interval Lower Limit = $\bar{x} - (Margin\ of\ Error)$

e) Confidence Interval Upper Limit = $\bar{x} + (Margin\ of\ Error)$

15. Use the following Statcato printout to check your margin of error and confidence interval answers from the temperature data in number 14. Now check the assumptions and write sentences to explain the margin of error and confidence interval. A histogram of the data has been created with Statcato.

Var	N	Mean	Stdev	Margin of Error	95.0%CI
summary	50.0	98.26	0.765	0.217	(98.0426, 98.4774)



a) Check each of the assumptions for this problem. Explain your answers.

b) Write a sentence to explain the margin of error in context.

c) Write a sentence to explain the confidence interval in context.

16. A random sample of cereal sugar content (grams per serving) was taken from 24 cereals. The sample mean average amount of sugar was of 7.208 grams per serving and a standard deviation of 4.634 grams per serving. What is the degrees of freedom? Open StatKey at www.lock5stat.com. Go to "theoretical distributions" and click on "T". Use the degrees of freedom and StatKey to look up the critical value T-scores for a 99% confidence level. Use the critical values and the given standard error to calculate the margin of error and construct a 99% confidence interval estimate of the population mean average amount of sugar in cereals.

Standard Error = 0.9459 grams

a) Degrees of Freedom = $n - 1$ =

b) Critical value T-scores = ±

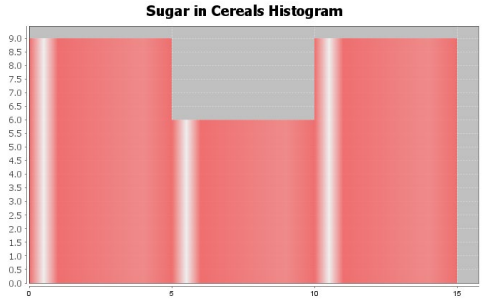
c) Margin of Error = $T \times Standard\ Error$ =

d) Confidence Interval Lower Limit = $\bar{x} - (Margin\ of\ Error)$

e) Confidence Interval Upper Limit = $\bar{x} + (Margin\ of\ Error)$

17. Use the following Statcato printout to check your margin of error and confidence interval answers from the sugar in cereals data in number 16. Now check the assumptions and write sentences to explain the margin of error and confidence interval. A histogram of the data has been created with Statcato.

Var	N	Mean	Stdev	Margin of Error	99.0%CI
Sugar (grams per serving)	24.0	7.208	4.634	2.656	(4.5527, 9.8639)



a) Check each of the assumptions for this problem. Explain your answers.

b) Write a sentence to explain the margin of error in context.

c) Write a sentence to explain the confidence interval in context.

18. A random sample of cereal carbohydrate content (grams per serving) was taken from 24 cereals. The sample mean average amount of carbs was of 15.043 grams per serving and a standard deviation of 3.596 grams per serving. What is the degrees of freedom? Open StatKey at www.lock5stat.com. Go to "theoretical distributions" and click on "T". Use the degrees of freedom and StatKey to look up the critical value T-scores for a 99% confidence level. Use the critical values and the given standard error to calculate the margin of error and construct a 99% confidence interval estimate of the population mean average amount of sugar in cereals.

Standard Error = 0.734 grams

a) Degrees of Freedom = $n - 1$ =

b) Critical value T-scores = ±

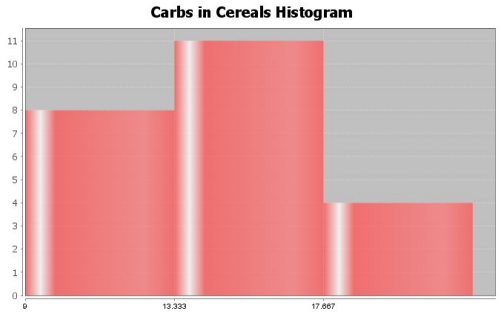
c) Margin of Error = $T \times Standard\ Error$ =

d) Confidence Interval Lower Limit = $\bar{x} - (Margin\ of\ Error)$

e) Confidence Interval Upper Limit = $\bar{x} + (Margin\ of\ Error)$

19. Use the following Statcato printout to check your margin of error and confidence interval answers from the carbohydrates in cereals data in number 18. Now check the assumptions and write sentences to explain the margin of error and confidence interval. A histogram of the data has been created with Statcato.

Var	N	Mean	Stdev	Margin of Error	99.0%CI
summary	24.0	15.043	3.596	2.061	(12.9824, 17.1036)



a) Check each of the assumptions for this problem. Explain your answers.

b) Write a sentence to explain the margin of error in context.

c) Write a sentence to explain the confidence interval in context.

One-Population Bootstrap Confidence Interval Practice Problems

20. An experiment was conducted to see what percentage of rats would show empathy toward fellow rats in distress. Of the 30 total rats in the study, 23 showed empathy. Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Proportion". Click on "Edit Data" and enter 23 for the "count" and 30 for the "sample size". Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the proportion. Use the bootstrap distribution to find a 99% confidence interval for the population proportion.

a) Does this data meet the assumptions for a bootstrap confidence interval? Explain your answer.

b) How many bootstrap samples did you take?

c) What is the shape of the bootstrap distribution?

d) Write the upper and lower limits of the bootstrap confidence interval. Compare the upper and lower limits of the bootstrap confidence interval to the ones found by the traditional formula with Statcato in #5. Are the close?

e) Write a sentence to explain the bootstrap confidence interval estimate of the population proportion.

21. A study was done on the effectiveness of lie detector tests to catch someone that lies. In a random sample of 48 total lies, the machine identified only 31 of them. Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Proportion". Click on "Edit Data" and enter 31 for the "count" and 48 for the "sample size". Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the proportion. Use the bootstrap distribution to find a 95% confidence interval for the population proportion.

a) Does this data meet the assumptions for a bootstrap confidence interval? Explain your answer.

b) How many bootstrap samples did you take?

c) What is the shape of the bootstrap distribution?

d) Write the upper and lower limits of the bootstrap confidence interval. Compare the upper and lower limits of the bootstrap confidence interval to the ones found by the traditional formula with Statcato in #7. Are the close?

e) Write a sentence to explain the bootstrap confidence interval estimate of the population proportion.

22. We want to determine what percentage of cereals the company Quaker makes. A random sample of 24 cereals found that Quaker made four of them. Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Proportion". Click on "Edit Data" and enter 4 for the "count" and 24 for the "sample size". Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the proportion. Use the bootstrap distribution to find a 90% confidence interval for the population proportion.

a) Does this data meet the assumptions for a bootstrap confidence interval? Explain your answer.

b) How many bootstrap samples did you take?

c) What is the shape of the bootstrap distribution?

d) Write the upper and lower limits of the bootstrap confidence interval. Compare the upper and lower limits of the bootstrap confidence interval to the ones found by the traditional formula with Statcato in #9. Are the close?

e) Write a sentence to explain the bootstrap confidence interval estimate of the population proportion.

23. If a cereal has more than 9 grams of sugar per serving, we consider it to have a high sugar content. We want to determine what percentage of cereals have a high sugar content. A random sample of 24 cereals found that 10 of them have a high sugar content. Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Proportion". Click on "Edit Data" and enter 10 for the "count" and 24 for the "sample size". Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the proportion. Use the bootstrap distribution to find a 95% confidence interval for the population proportion.

a) Does this data meet the assumptions for a bootstrap confidence interval? Explain your answer.

b) How many bootstrap samples did you take?

c) What is the shape of the bootstrap distribution?

d) Write the upper and lower limits of the bootstrap confidence interval. Compare the upper and lower limits of the bootstrap confidence interval to the ones found by the traditional formula with Statcato in #11. Are the close?

e) Write a sentence to explain the bootstrap confidence interval estimate of the population proportion.

24. Go to www.matt-teachout.org, click on "Statistics" and then "Data Sets". Open the "cereal data" in excel. Copy the column of data labeled "sugar (grams per serving)". Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Mean, Median, St.Dev." Click on "Bootstrap Dot plot of Mean". Now click on "Edit Data" and paste the sugar data into StatKey. Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the mean. Use the bootstrap distribution to find a 99% confidence interval for the population mean.

a) Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.

b) How many bootstrap samples did you take?

c) What is the shape of the bootstrap distribution for the mean?

d) Write the upper and lower limits of the bootstrap confidence interval for the population mean. Compare the upper and lower limits of the bootstrap confidence interval to the ones found by the traditional formula with Statcato in #17. Are the close?

e) Write a sentence to explain the bootstrap confidence interval estimate of the population mean.

We can also use bootstrapping to estimate the population median average amount of sugar in cereals. Click on "Bootstrap Dot plot of Median". Use the bootstrap distribution to find a 99% confidence interval for the population median.

f) What is the shape of the bootstrap distribution for the median?

g) Write the upper and lower limits of the bootstrap confidence interval for the population median.

h) Write a sentence to explain the bootstrap confidence interval estimate of the population median.

25. Go to www.matt-teachout.org, click on "Statistics" and then "Data Sets". Open the "cereal data" in excel. Copy the column of data labeled "carbs (grams per serving)". Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Mean, Median, St.Dev." Click on "Bootstrap Dot plot of Mean". Now click on "Edit Data" and paste the carb data into StatKey. Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the mean. Use the bootstrap distribution to find a 95% confidence interval for the population mean.

a) Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.

b) How many bootstrap samples did you take?

c) What is the shape of the bootstrap distribution for the mean?

d) Write the upper and lower limits of the bootstrap confidence interval for the population mean. Compare the upper and lower limits of the bootstrap confidence interval to the ones found by the traditional formula with Statcato in #19. Are the close?

e) Write a sentence to explain the bootstrap confidence interval estimate of the population mean.

We can also use bootstrapping to estimate the population median average amount of carbohydrates in cereals. Click on "Bootstrap Dot plot of Median". Use the bootstrap distribution to find a 95% confidence interval for the population median.

f) What is the shape of the bootstrap distribution for the median?

g) Write the upper and lower limits of the bootstrap confidence interval for the population median.

h) Write a sentence to explain the bootstrap confidence interval estimate of the population median.

26. Go to www.matt-teachout.org, click on "Statistics" and then "Data Sets". Open the "bear data" in excel. Copy the column of data labeled "weight in pounds". Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Mean, Median, St.Dev." Click on "Bootstrap Dot plot of Mean". Now click on "Edit Data" and paste the bear weight data into StatKey. Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the mean. Use the bootstrap distribution to find a 90% confidence interval for the population mean average weight of bears.

a) Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.

b) How many bootstrap samples did you take?

c) What is the shape of the bootstrap distribution for the mean?

d) Write the upper and lower limits of the bootstrap confidence interval for the population mean.

e) Write a sentence to explain the bootstrap confidence interval estimate of the population mean.

We can also use bootstrapping to estimate the population median average weight of bears. Click on "Bootstrap Dot plot of Median". Use the bootstrap distribution to find a 90% confidence interval for the population median.

f) What is the shape of the bootstrap distribution for the median?

g) Write the upper and lower limits of the bootstrap confidence interval for the population median.

h) Write a sentence to explain the bootstrap confidence interval estimate of the population median.

27. Go to www.matt-teachout.org, click on "Statistics" and then "Data Sets". Open the "bear data" in excel. Copy the column of data labeled "length in inches". Do not click on "head length" by mistake. We want the overall length of the bears. Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Mean, Median, St.Dev." Click on "Bootstrap Dot plot of Mean". Now click on "Edit Data" and paste the bear length data into StatKey. Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the mean. Use the bootstrap distribution to find a 99% confidence interval for the population mean average length of bears.

a) Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.

b) How many bootstrap samples did you take?

c) What is the shape of the bootstrap distribution for the mean?

d) Write the upper and lower limits of the bootstrap confidence interval for the population mean.

e) Write a sentence to explain the bootstrap confidence interval estimate of the population mean.

We can also use bootstrapping to estimate the population median average length of bears. Click on "Bootstrap Dot plot of Median". Use the bootstrap distribution to find a 99% confidence interval for the population median.

f) What is the shape of the bootstrap distribution for the median?

g) Write the upper and lower limits of the bootstrap confidence interval for the population median.

h) Write a sentence to explain the bootstrap confidence interval estimate of the population median.

Section 2F – Two-Population Mean & Proportion Confidence Intervals

Studying the differences between two populations is very common in statistics, however sampling variability makes it very difficult to determine. Think of it this way. We know that random samples are usually different from each other, so even if two populations were the same, the samples taken from those populations would be different. A key question to ask is why are the samples different? Are the samples different because the populations are different or are my samples different because of sampling variability? Here is another key question. Are my samples significantly different or only slightly different? Two-population confidence intervals are often used to answer these difficulty questions.

Before you can understand two-population confidence intervals, we have to take you back to arithmetic. A two-population confidence interval is the answer to a subtraction problem. Remember, the answer to a subtraction problem is often called the "difference". We need to understand how subtraction works and what a difference actually tells us.

Understanding Positive Differences

Suppose you subtract two numbers and the answer comes out positive. There is a positive difference. Is the first number bigger or smaller than the second number? Let us look at an example.

17 – 6 = +11

What does this tell us? Since the difference comes out positive, we know that the first number (17) is larger than the second number (6). It actually tells us more than this. The answer of +11 tells us that the first number (17) is 11 units larger than the second number (6).

How does this translate to a two-population confidence interval?

A two-population confidence interval does not measure population 1 or population 2 individually. Instead, it measures the difference between the population parameters. Two-population mean confidence intervals measure $\mu_1 - \mu_2$ (the difference between the population means). Two-population proportion confidence intervals measure $\pi_1 - \pi_2$ or $p_1 - p_2$ (the difference between the population proportions). The key is that the confidence interval is the answer to a subtraction problem.

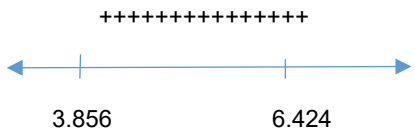
Example: We want to compare the population mean height of men and women. We used a random sample of 40 men's heights in inches and a random sample of 40 women's heights in inches. Putting the data into Statcato, we got the following two population mean confidence interval. Population 1 was men's heights and population 2 was women's heights. We will assume for now that this data did meet the assumptions to estimate the populations.

Confidence Intervals - Two population means: confidence level = 0.95
Samples of population 1 in C16 Men Ht (in)
Samples of population 2 in C2 Women Ht (in)

	N	Mean	Stdev
Population 1	40	68.335	3.020
Population 2	40	63.195	2.741

* Population standard deviations are unknown. *
DOF = 77
Margin of error = 1.284
95.0%CI = (3.8560, 6.4240)

First of all, 3.856 inches is NOT population 1 and 6.424 is NOT population 2. That is not how confidence intervals work. Remember this is an interval. It represents all of the numbers in between 3.856 and 6.424 inches and difference between the population means ($\mu_1 - \mu_2$) could be any of them. Think of the number line. Notice that all of the numbers between 3.856 and 6.424 are positive!



So while we do not know what the population difference is exactly, we do know that the difference is positive. Think back. Remember if the difference is positive, then the first number must be larger than the second. In this case, the mean average of population 1 (men's heights) is likely to be larger than the mean average of population 2 (women's heights). Remember the positive difference tells you how much larger.

Sentence to explain the confidence interval: We are 95% confident that the population mean average height of men (population 1) is between 3.856 and 6.424 inches larger than the population mean average height of women (population 2).

Note: You may also see the two-population confidence interval sentence written this way. We are 95% confident that the difference between the population mean average heights of men and women is between 3.856 and 6.424 inches. This can be a confusing way to explain the confidence interval though as people rarely understand the implications of that sentence.

Significance

Notice that the sample mean average height for the 40 men in the sample data was 68.335 inches and the sample mean average height for the 40 women in the sample data was 63.195 inches. Are these sample mean's significantly different? Yes. If both the upper and lower limits of your two population confidence interval are positive (+,+), then that does indicate that your sample statistic from group 1 is significantly higher than the sample statistic from group 2.

Positive Difference Two-population Confidence Intervals (+,+)

Sentence: "We are #% confident that the parameter from population 1 is between # and # larger than the parameter from population 2."

Significance: There is a significant difference between the two samples. The sample statistic for group 1 is significantly higher than group 2. This indicates that the parameter for population 1 might be higher than for population 2.

.

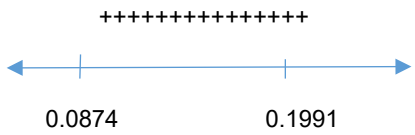
Example 2: Suppose we want to compare the percentage of statistics students that are democrat and the percentage of statistics students that are republican. We used the fall 2015 COC survey data to create the following confidence interval. For now, we will assume the problem met the assumptions for estimating the populations. Population 1 was democratic COC statistics students and population 2 was republican COC statistics students. We used a 90% confidence level and Statcato to calculate the following two-population proportion confidence interval.

Confidence Interval - Two population proportions: confidence level = 0.9

	Number of Events	Number of trials	Proportion
Sample 1	110	328	0.335
Sample 2	63	328	0.192

Sample proportion difference = 0.143
Margin of error = 0.056
90.0%CI = (0.0874, 0.1991)

Notice that both of the numbers in the two-population confidence interval are positive. These proportions can be converted to their percentage equivalent (+8.74% , +19.91%). Again 8.74% is NOT population 1 and 19.91% is NOT population 2. That is not how two-population confidence intervals work. The difference between the population proportions $(\pi_1 - \pi_2)$ could be any of the numbers between 0.0874 and 0.1991. Notice again that all of the numbers in between 0.0874 and 0.1991 are positive.



So the population proportion difference $\pi_1 - \pi_2$ is positive. This tells us that the population proportion (and percentage) of COC statistics students that are democrat (population 1) is likely to be larger than the population proportion or percentage of COC statistics students that are republican (population 2). Remember the positive difference tells you how much larger.

Sentence:

We are 90% confident that the population percentage of COC statistics students that are democratic (population 1) is between 8.74% and 19.91% higher than the percentage of COC statistics students that are republican (population 2).

OR

We are 90% confident that the population proportion of COC statistics students that are democratic (population 1) is between 0.0874 and 0.1991 higher than the population proportion of COC statistics students that are republican (population 2).

Understanding Negative Differences

Suppose you subtract two numbers and the answer comes out negative. (There is a negative difference.) Is the first number bigger or smaller than the second number? Let us look at an example.

5 – 13 = −8

What does this tell us? Since the difference comes out negative, we know that the first number (5) is smaller than the second number (13). It actually tells us more than this. The answer of −8 tells us that the first number (5) is eight units smaller than the second number (13). Notice we did not say that the first number is −8 units smaller. The difference of −8 tells us that the first number is eight units smaller than the second number.

How does this translate to a two-population confidence interval?

Remember, a two-population confidence interval does not measure population 1 or population 2 individually. Instead, it measures the difference between the population parameters. Two-population mean confidence intervals measure $\mu_1 - \mu_2$ (the difference between the population means). Two-population proportion confidence intervals measure $\pi_1 - \pi_2$ or $p_1 - p_2$ (the difference between the population proportions).

Example: We want to compare the population mean weight of women and men. We used a random sample of 40 women's weights in pounds and a random sample of 40 men's weights in pounds. Putting the data into Statcato, we got the following two population mean confidence interval. Population 1 was women's weights and population 2 was men's weights. We will assume for now that this data did meet the assumptions to estimate the populations.

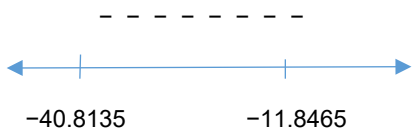
Confidence Intervals - Two population means: confidence level = 0.95
Samples of population 1 in C3 Women Wt (Lbs)
Samples of population 2 in C17 Men Wt (Lbs)

	N	Mean	Stdev
Population 1	40	146.220	37.621
Population 2	40	172.55	26.327

* Population standard deviations are unknown. *
DOF = 69
Margin of error = 14.484
95.0%CI = (-40.8135, -11.8465)

Remember, −40.8135 pounds is NOT population 1 and −11.8465 pounds is NOT population 2. That is not how two-population confidence intervals work. Remember this is an interval. It represents all of the numbers in between −40.8135 and −11.8465 pounds and difference between the population means ($\mu_1 - \mu_2$) could be any of them. Notice that the lower limit is now −40.8135 on the left and the upper limit is −11.8465 on the right. Many students are confused by this, but that is how the number line works. The more negative a number is, the smaller it is. Therefore, −40.8135 is smaller −11.8465.

How do we interpret this? Think again of the number line. Notice that all of the numbers between −40.8135 and −11.8465 are negative!



So while we do not know what the population difference is exactly, we do know that the difference is negative. Remember if the difference is negative, then the first number must be smaller than the second number. In this case, the mean average of population 1 (women's weights) is likely to be smaller than the mean average of population 2 (men's weights). The negative difference tells you how much smaller.

Sentence to explain the confidence interval: We are 95% confident that the population mean average weight of women (population 1) is between 11.8465 pounds and 40.8135 pounds less than the population mean average height of men (population 2).

Significance

Notice that the sample mean average weight for the 40 women in the sample data was 146.220 pounds and the sample mean average height for the 40 men in the sample data was 172.55 pounds. Are these sample mean's significantly different? Yes. If both the upper and lower limits of your two population confidence interval are negative (−,−), then that does indicate that your sample statistic from group 1 is significantly lower than the sample statistic from group 2.

Negative Difference Two-population Confidence Intervals (−,−)

Sentence: "We are #% confident that the parameter from population 1 is between # and # lower than (or less than) the parameter from population 2."

Significance: There is a significant difference between the two samples. The sample statistic for group 1 is significantly lower than group 2. This indicates that parameter for population 1 is probably lower than for population 2.

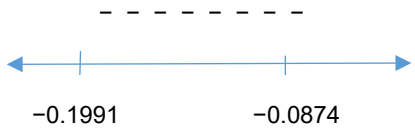
Example 2: In a previous example, we compared the percentage of statistics students that are democrat and the percentage of statistics students that are republican. We assigned democrat to be population 1 and republican to be population 2. What would happen if we reverse that? Suppose we let population 1 to be republican COC statistics students and population 2 to be democrat COC statistics students. We used a 90% confidence level and Statcato to calculate the following two-population proportion confidence interval. Assume the problem met the assumptions for estimating the populations.

Confidence Interval - Two population proportions: confidence level = 0.9

	Number of Events	Number of trials	Proportion
Sample 1	63	328	0.192
Sample 2	110	328	0.335

Sample proportion difference = -0.143
Margin of error = 0.056
90.0%CI = (-0.1991, -0.0874)

Notice that the sample difference is now negative, but the margin of error is the same. Both of the numbers in the two-population confidence interval are now negative. These proportions can be converted to their percentage equivalent (−19.91%, −8.74%). Notice that these are the same percentages, but have opposite signs. Notice that the lower limit is now −19.91% on the left and the upper limit is −8.74% on the right. Remember, the more negative a number is, the smaller it is, so −19.91% is smaller −8.74%. The difference between the population proportions $(\pi_1 - \pi_2)$ could be any of the numbers between −0.1991 and −0.0874. Notice again that all of the numbers in between −0.1991 and −0.0874 are negative.



So the population proportion difference $\pi_1 - \pi_2$ is negative. This tells us that the population proportion (and percentage) of COC statistics students that are republican (population 1) is likely to be smaller than the population proportion or percentage of COC statistics students that are democrat (population 2). The confidence interval being negative tells you how much smaller.

Sentence:

We are 90% confident that the population percentage of COC statistics students that are republican (population 1) is between 8.74% and 19.91% lower than the percentage of COC statistics students that are democrat (population 2).

OR

We are 90% confident that the population proportion of COC statistics students that are republican (population 1) is between 0.0874 and 0.1991 lower than the population proportion of COC statistics students that are democrat (population 2).

Significance:

Since both the upper and lower limits of the confidence interval were negative, this suggests that the sample percentage for group 1 (republican) was significantly lower than the sample percentage for group 2 (democrat). This indicates that the population percentage for republican COC statistics students is likely to be lower than the percentage for democratic COC statistics students.

Zero Difference

If we subtract two numbers and the answer is zero, the two numbers must be the same. Look at the following example.

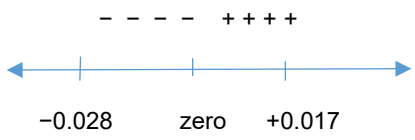
13 – 13 = 0

The zero difference tells us that the first number (13) is the same as the second number (13).

Example: Suppose 95% two-population proportion confidence interval came out to be (−0.028 , +0.017). Remember, −0.028 is NOT population 1 and +0.017 is NOT population 2. This confidence interval tells us that the difference between the population proportions $(\pi_1 - \pi_2)$ is somewhere between −0.028 and +0.017. Some people will write the sentence as follows.

Sentence: We are 95% confident that the population proportion difference is between −0.028 and +0.017.

What does that even mean? Is population 1 lower or higher than population 2? How much lower or higher? To answer these questions, we need to examine the number line between −0.028 and +0.017. Notice that there are many negative numbers in this interval, so population 1 may be lower than population 2. There are also many positive numbers in this interval, so population 1 may be higher than population 2. Zero is also in the interval, so it is also a possibility. Remember if the difference is zero, then population 1 and population 2 could be the same. This interval tells us that we really do not know which population is larger. When the upper and lower limits of a two-population confidence interval have opposite signs, this means there is no significant difference between the populations. The sample statistics for the two groups are so close, that we cannot tell if population 1 is lower or higher than population 2. They could be the same.

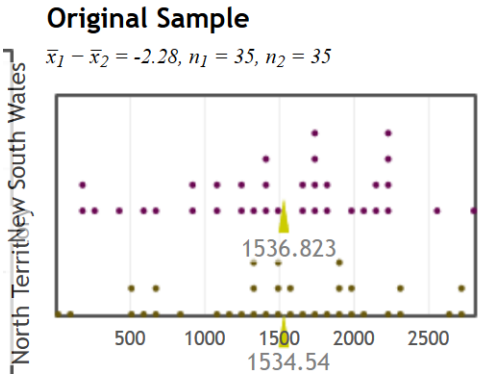


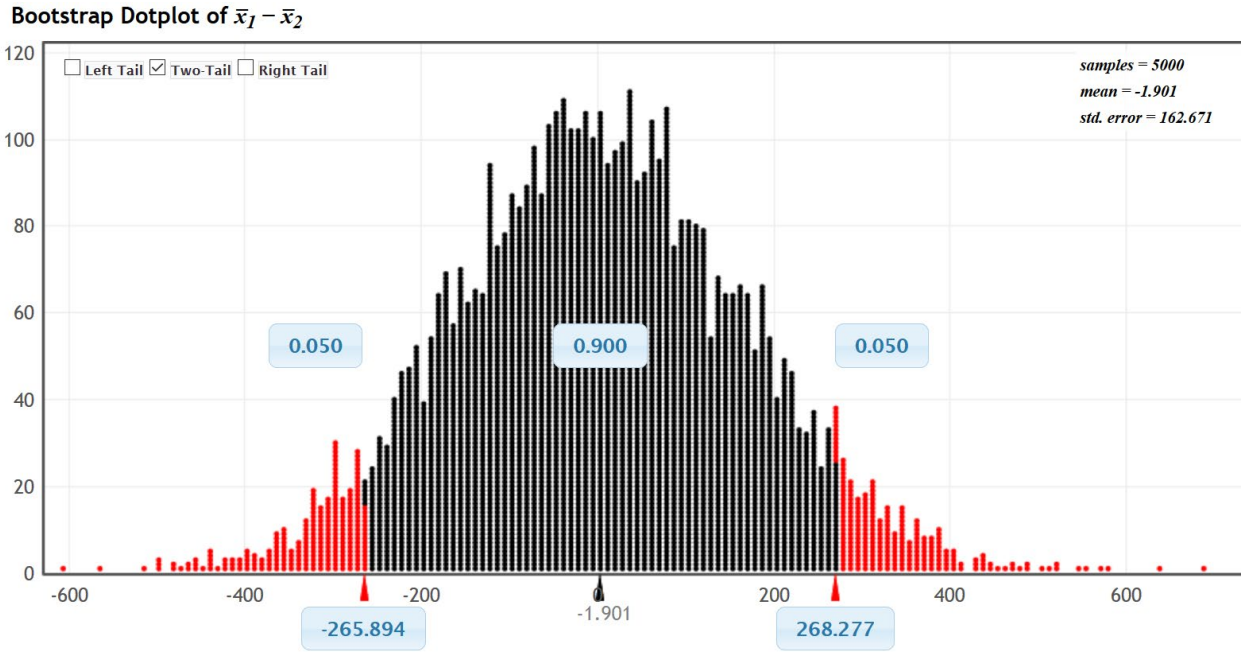
Two-population Confidence Intervals (− , +)

Sentence: We are #% confident that there is no significant difference between the parameter for population 1 and parameter for population 2.

Significance: When the upper and lower limits for the two-population confidence interval have opposite signs, then that indicates that the sample statistics for the two groups are not significantly different.

Example: Let us compare the population mean average salary of people living in Northern Territory, Australia (μ_1) to people living in New South Wales, Australia (μ_2). We used StatKey and random sample data to create the following two-population mean bootstrap 90% confidence interval. Assume the data met all of the assumptions.





From the bootstrap, we see that the 90% confidence interval is (−265.894 , +268.277). Notice the upper and lower limit have opposite signs. This tells us that the sample mean average salary for people living in the Northern Territory ($1534.54) is not significantly different from the sample mean average salary for people living in New South Wales ($1536.82). Since our sample means are so close, we cannot tell which population has a higher population mean average salary.

Confidence Interval Sentence: We are 90% confident that the difference between the population mean average salary of people living in Northern Territory, Australia and those living in New South Wales, Australia is between −$265.894 and +$268.277. *(This sentence tends to be confusing.)*

Better Confidence Interval Sentence: We are 90% confident that there is no significant difference between the population mean average salary of people living in Northern Territory, Australia and those living in New South Wales, Australia.

Note: We could also have calculated the 90% confidence interval with Statcato. Notice the upper and lower limits of the bootstrap are similar to what Statcato calculated.

Confidence Intervals - Two population means: confidence level = 0.9

Samples of population 1 in C1 North Territory ...

Samples of population 2 in C2 New South Wales ...

	N	Mean	Stdev
Population 1	35	1534.540	701.525
Population 2	35	1536.823	677.140

* Population standard deviations are unknown. *

DOF = 67

Margin of error = 274.883

90.0%CI = (-277.1660, 272.5998)

Calculating Two-population Mean and Proportion Confidence Intervals

We will now discuss the formulas and calculations for two-population mean and proportion confidence intervals. It is important to understand the formulas and be able to explain them. However, no statistician or data scientist calculates these by hand with a formula. We virtually always use computer software to calculate any difficult calculations like confidence intervals.

Two-population Mean Confidence Intervals

There are two types of two-population mean confidence intervals, independent groups and matched pairs. Matched pair data is a one-to-one pairing between the two groups. Matched pair data usuallyfrom the same person measured twice. For example, the first number in the first data set comes from the same person as the first number in the second data set. The second numbers in each data set come from the same person and so on. Matched pairs do not have to be the same person measured twice. It could also be comparing husbands and wives, or sisters and brothers. You could be comparing two football teams and comparing the salary for each position: the starting quarterbacks, the starting running backs, the starting right guard, etc. Notice that in matched pairs, the sample sizes for the two groups are the same.

Use independent groups when you are comparing separate groups. For example, like comparing a random sample of men to a random sample of women or comparing a random sample of people from California to a random sample of people from Arizona.

Example 1 (Matched Pair): Let's use the random sample health data and a 99% confidence interval to compare the population mean systolic and diastolic blood pressure for men. Since these values come from the same 40 men, they are matched pairs.

Population 1: Men's Systolic Blood Pressure (mm of Hg)
Population 2: Men's Diastolic Blood Pressure (mm of Hg)

For independent groups, we calculate sample mean and sample standard deviation separately for each group and then subtract the sample means. For matched pair, we subtract the ordered pairs first, and then calculate the mean of the difference ($\bar{d}$) and the standard deviation of the difference (s_d).

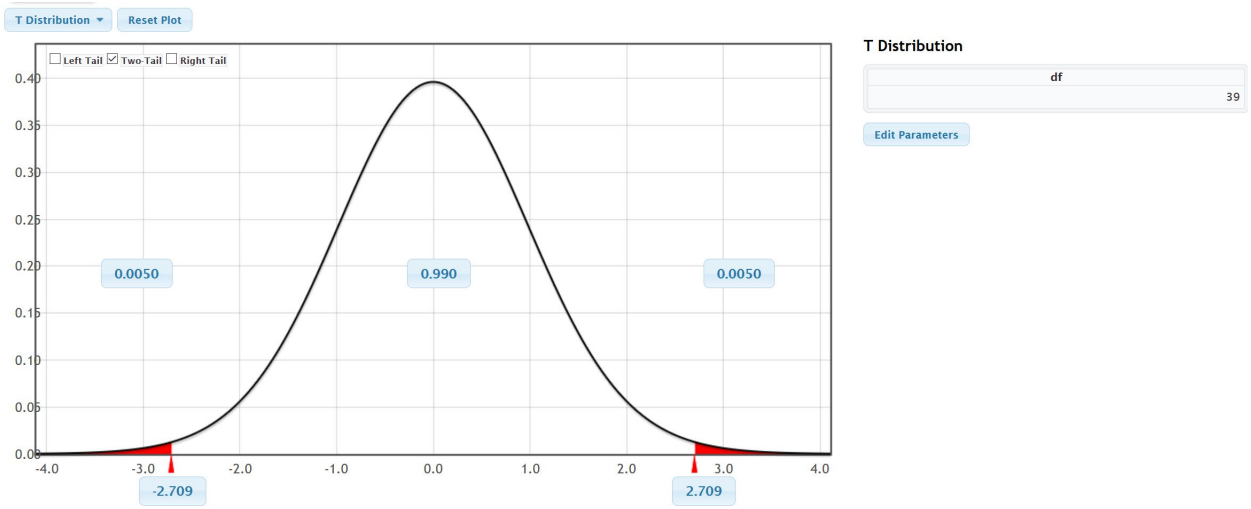
Since the systolic and diastolic blood pressure for these 40 men were matched pairs, notice we subtracted each pair and created a new column of data called the "difference" column. A two-population mean matched pair confidence interval is calculating a one-population confidence interval using just the difference column.

Men Syst BP (mm of Hg)	Men Diast BP (mm of Hg)	Difference between pairs
125	78	47
107	54	53
126	81	45
110	68	42
110	66	44
107	83	24
113	71	42
126	72	54
137	85	52
110	71	39
109	65	44
153	87	66
112	77	35
119	81	38
113	82	31
125	76	49
131	80	51
121	75	46
132	81	51
112	44	68
121	65	56
116	64	52
95	58	37
110	70	40
110	66	44

Descriptive Statistics

Variable	Mean	Standard Deviation
C3 Difference between pairs	45.675	9.352

The sample mean of the difference ($\bar{d}$) is 45.675, the sample standard deviation of the difference (s_d) is 9.352 and the sample size (n) is 40. These are used in the confidence interval calculation. We will also need to look up the T critical value. In matched pair, the sample size is the number of pairs (40), so the degrees of freedom is 39. We can use the theoretical T distribution function in StatKey to look up the critical value.



Notice that the T critical values are ±2.709. Here is the formula and calculation for the two-population mean matched pair confidence interval. Notice that the sample mean difference is 45.675 mm of Hg and the margin of error is 4.0057 mm of Hg. This gave us a confidence interval of

$\overline{d} \pm T \frac{s_d}{\sqrt{n}}$

$45.675 \pm 2.709 \frac{9.352}{\sqrt{40}}$

45.675 ± 4.0057

(41.6693 , 49.6807)

Notice that the upper and lower limits of the confidence interval are both positive. This tells us that the population 1 (men's systolic blood pressure) is higher than population 2 (men's diastolic blood pressure).

Sentence: We are 99% confident that the population mean systolic blood pressure for men is between 41.67 mm of Hg and 49.68 mm of Hg higher than the population mean diastolic blood pressure for men.

We can use Statcato to calculate this for us. Just go to the "statistics" menu in Statcato, click on "confidence intervals" and then matched pair. You can put in the summary data (sample mean difference 45.675, sample standard deviation of the differences 9.352, and sample size 40). You can also copy and paste the two quantitative data sets and then click the "samples in two columns" button.

Statcato => Statistics => Confidence Intervals => Matched Pairs

Confidence Interval - Matched Pairs: confidence level = 0.99

Sample 1: C1 Men Syst BP (mm ...

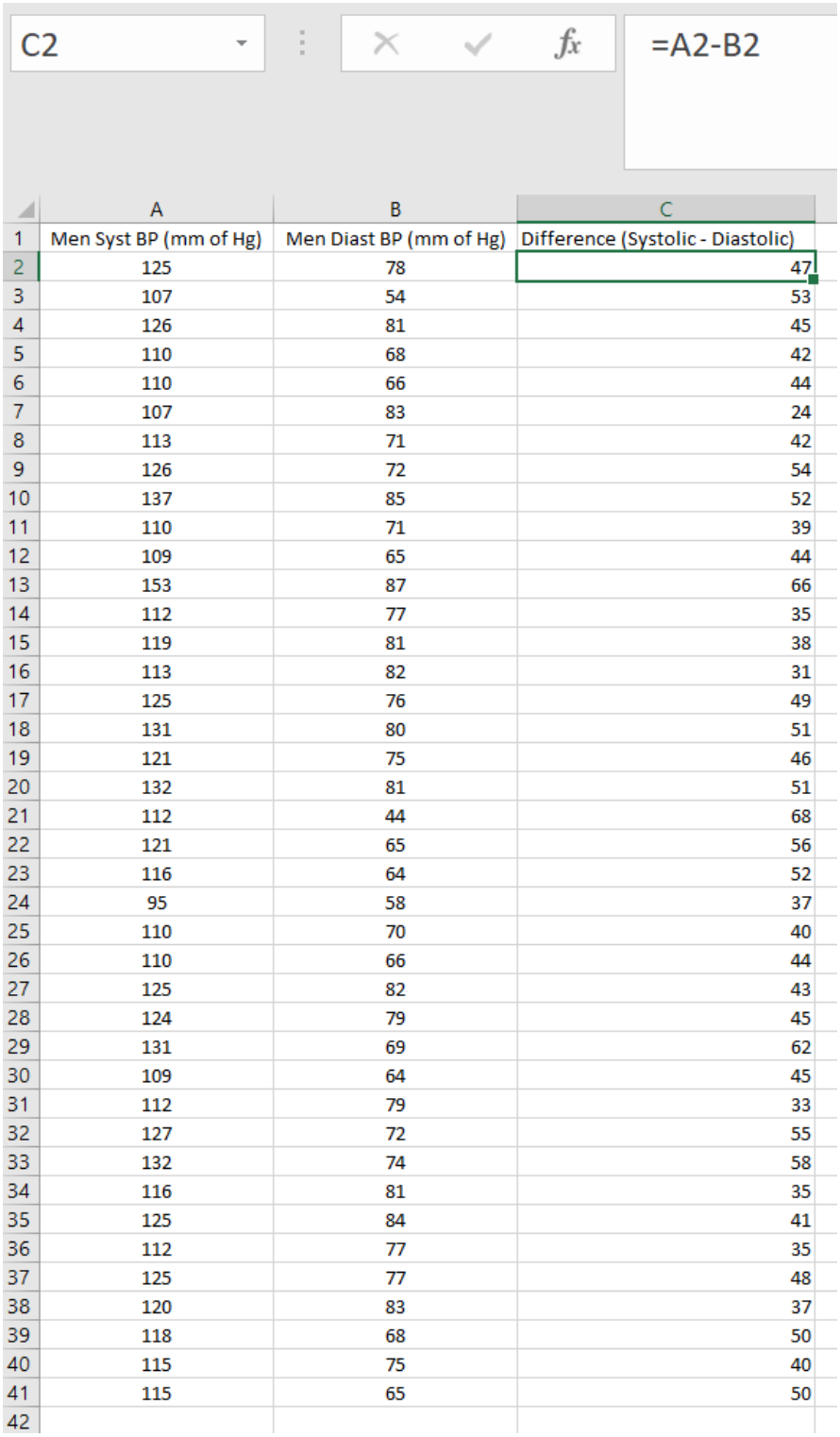
Sample 2: C2 Men Diast BP (mm...

Difference of Matched Pairs C1 Men Syst BP (mm ... - C2 Men Diast BP (mm...

N	Mean	Stdev	Margin of Error	99.0%CI
40	45.675	9.352	4.004	(41.6710, 49.6790)

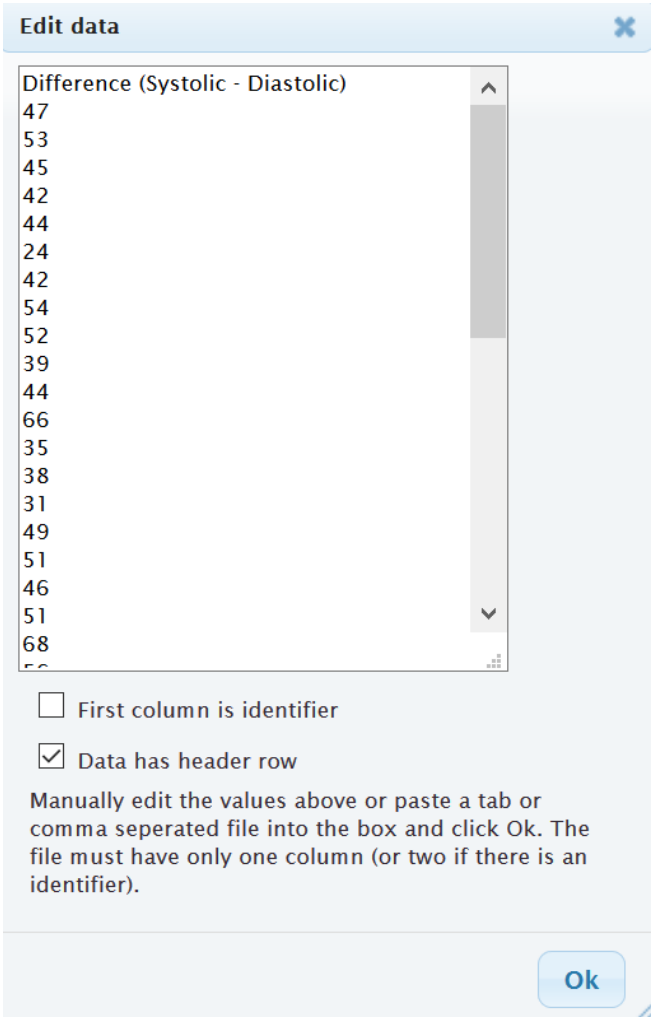
Notice that the confidence interval in Statcato is virtually the same as our formula calculation above.

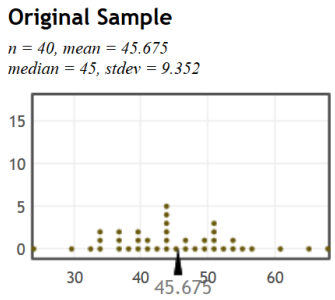
We can also use bootstrapping in StatKey to calculate this confidence interval. Remember a matched pair is calculated as a one-population mean bootstrap from the differences between the pairs. Let us start by calculating the difference column in Excel. Copy and pasted the two data sets into excel. In cell "C2" type in "=B2-C2" and push enter. Hold your curser on the bottom right corner until it turns into a "+". Double click and the formula will be applied to the rest of the data. You can also click and drag.



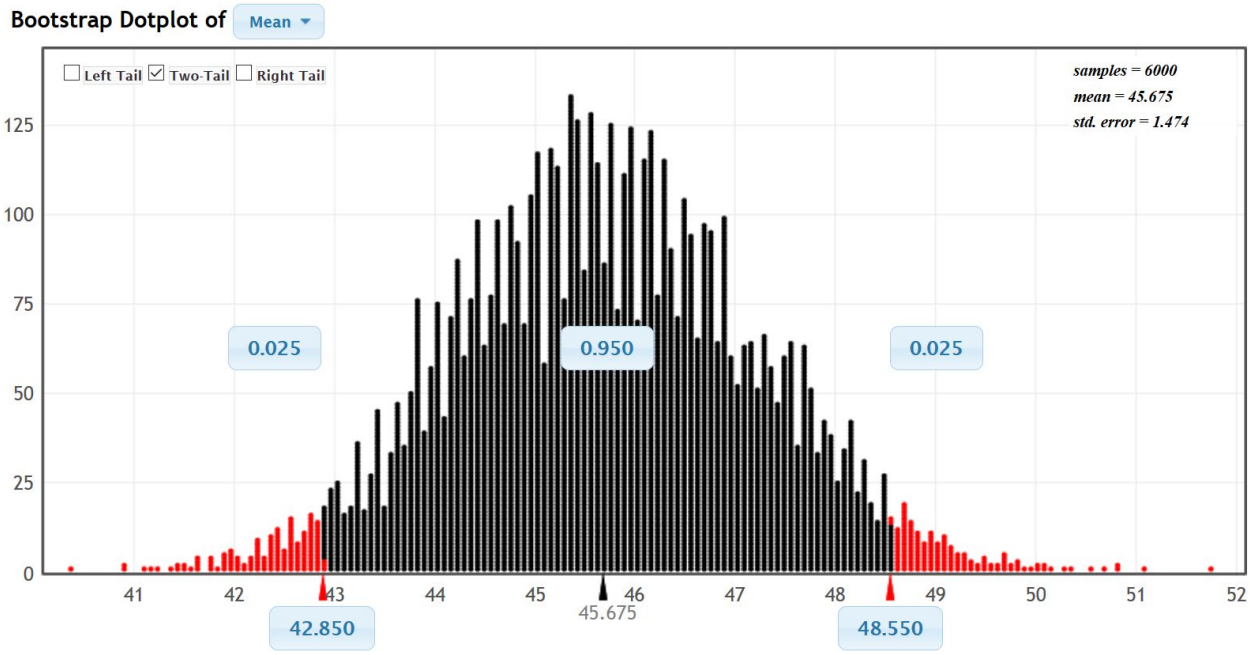
	A	B	C
1	Men Syst BP (mm of Hg)	Men Diast BP (mm of Hg)	Difference (Systolic - Diastolic)
2	125	78	47
3	107	54	53
4	126	81	45
5	110	68	42
6	110	66	44
7	107	83	24
8	113	71	42
9	126	72	54
10	137	85	52
11	110	71	39
12	109	65	44
13	153	87	66
14	112	77	35
15	119	81	38
16	113	82	31
17	125	76	49
18	131	80	51
19	121	75	46
20	132	81	51
21	112	44	68
22	121	65	56
23	116	64	52
24	95	58	37
25	110	70	40
26	110	66	44
27	125	82	43
28	124	79	45
29	131	69	62
30	109	64	45
31	112	79	33
32	127	72	55
33	132	74	58
34	116	81	35
35	125	84	41
36	112	77	35
37	125	77	48
38	120	83	37
39	118	68	50
40	115	75	40
41	115	65	50
42			

Open StatKey at www.lock5stat.com and click on “CI for Single Mean, Median, St.Dev.” under the “bootstrap confidence interval” menu. Make sure the bootstrap dot plot is set to “mean”. Click on edit data. Copy and paste the “difference” column only and push “Ok”.





Now create the bootstrap distribution by clicking on “generate 1000 samples” a few times and click on two-tail. The default is 95% but you can change the middle proportion to 0.90 or 0.99 if needed. Notice the 95% bootstrap confidence interval is (+42.85 , +48.55). This is similar to our formula calculations above.



Important Notes about two-population bootstraps: Remember bootstrap confidence intervals will always come out slightly different because of sampling variability. Also that though we used a one-population bootstrap, this was not a one-population confidence interval. It measured the difference between the populations and must be interpreted accordingly. Remember to keep track of population 1 and population 2 and the signs of the confidence intervals.

Example 2 (Two-population mean from Independent Groups): Earlier we used the health data to calculate the following two-population confidence interval to compare the population mean average weight of women and men. Notice these groups are independent and not a one-to-one pairing. The upper and lower limits were negative, indicating that we are 95% confident that the population mean average weight of women is between 11.8465 pounds and 40.8135 pounds less than the population mean average weight of men.

Confidence Intervals - Two population means: confidence level = 0.95

Samples of population 1 in C3 Women Wt (Lbs)

Samples of population 2 in C17 Men Wt (Lbs)

	N	Mean	Stdev
Population 1	40	146.220	37.621
Population 2	40	172.55	26.327

* Population standard deviations are unknown. *

DOF = 69

Margin of error = 14.484

95.0%CI = (-40.8135, -11.8465)

Let us discuss how Statcato calculated this confidence interval. Let us start with the degrees of freedom. For independent groups, the degrees of freedom calculation is much more difficult. There are many free online calculators for degrees of freedom. I like to use this one. You will need to enter the sample size and sample standard deviation for each of your two samples. Notice the degrees of freedom calculator gave 69.809. It is usually common to round down the degrees of freedom to account for possible greater variability. Notice Statcato rounded this degree of freedom down to 69 even though it was closer to 70.

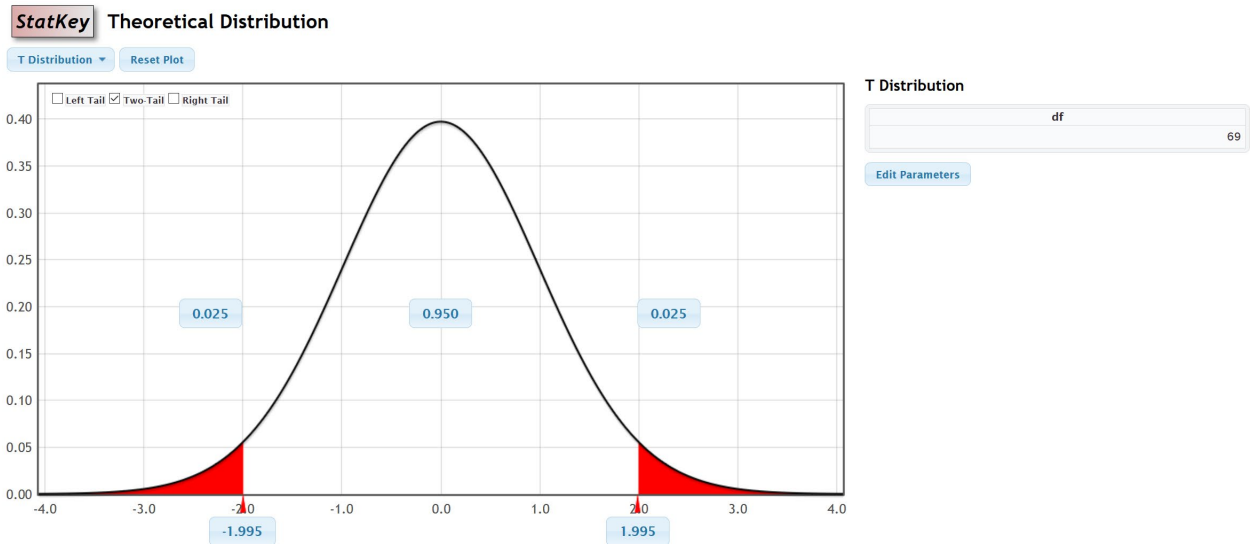
http://web.utk.edu/~cwiek/TwoSampleDoF

Compute Degrees of Freedom for t-test comparing means of two independent samples

Enter in the sample sizes (n1, n2) and sample standard deviations (s1, s2) and click "Compute DF" to get the degrees of freedom describing the sampling distribution of the difference in sample means.

n1: 40 n2: 40 s1: 37.621 s2: 26.327 Compute DF DF: 69.809

We can now look up the critical value T-scores for this confidence interval with the Theoretical Distribution T-score calculator in StatKey. Notice the critical value T-scores for 69 degrees of freedom are ± 1.995.



Here is the formula for the two-population mean confidence interval for independent groups. We see that the sample mean weight for the women $(\bar{x}_1)$ was 146.220 pounds, the sample mean weight for the men $(\bar{x}_2)$ was 172.55 pounds, the sample standard deviation for the women's weights (s_1) was 37.621 pounds and the sample standard deviation for the men's weights (s_2) is 26.327 pounds. While both sample sizes are 40 in this example, it is common for independent groups to have different sample sizes.

$$(\bar{x}_1 - \bar{x}_2) \pm T\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

$$(146.22 - 172.55) \pm 1.995\sqrt{\left(\frac{37.621^2}{40} + \frac{26.327^2}{40}\right)}$$

−26.33 ± 1.995 (7.26)

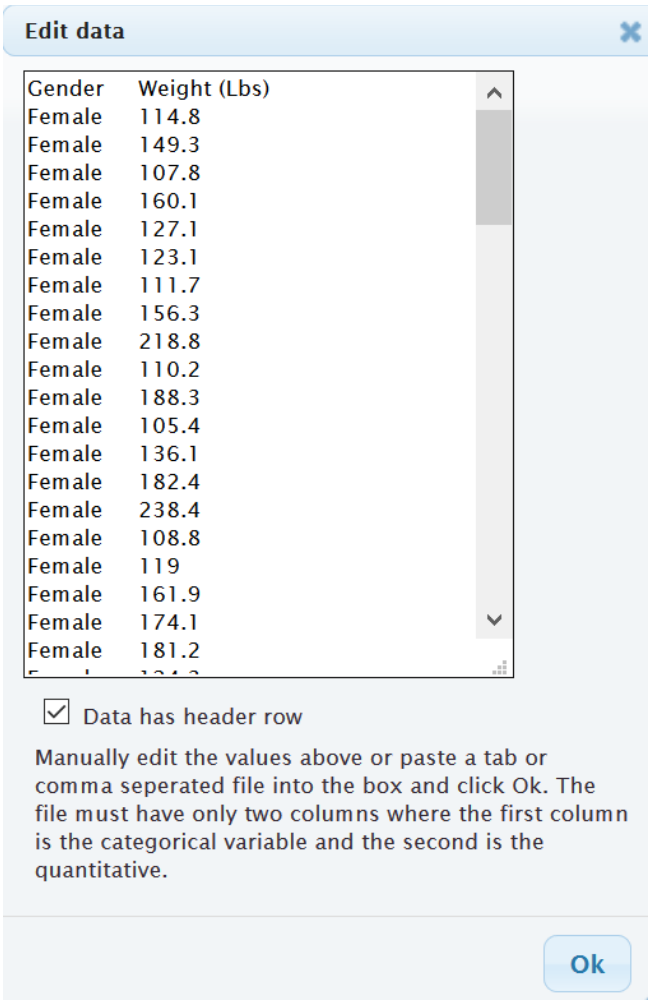
−26.33 ± 14.48

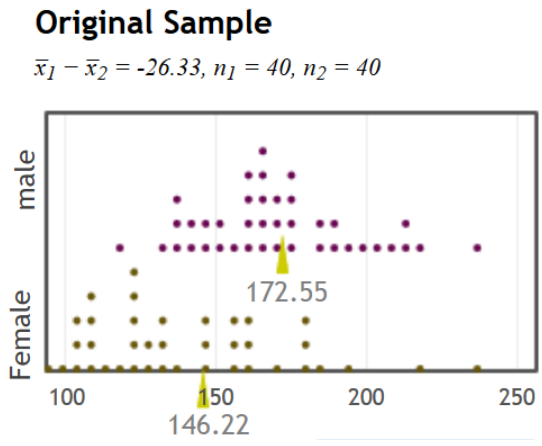
$$(-40.81\,, -11.85)$$

Note about Pooling the Variances: Statisticians sometimes pool the variances when comparing the population means from two populations. It requires the population variances to be the same. For students new to stats, it is better not to pool the variances.

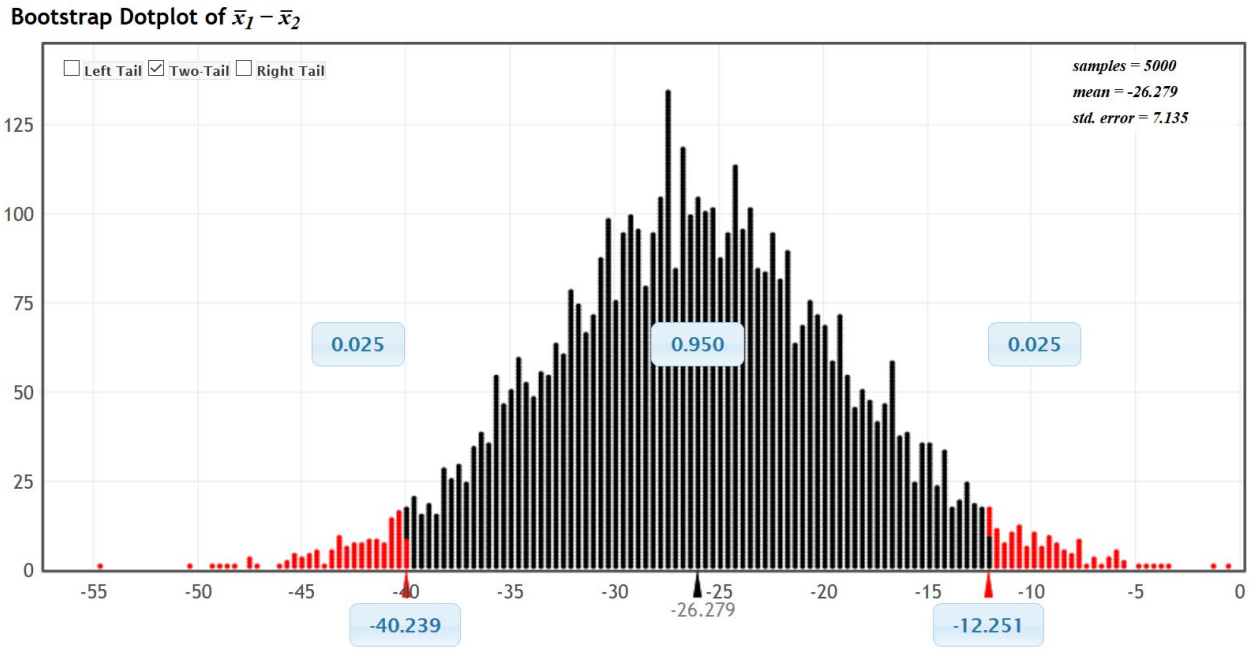
We can also calculate this confidence interval with bootstrapping. Go to www.lock5stat.com and click on StatKey. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Difference in Means". This menu is for independent groups. While Statcato prefers the data to be separated by group, StatKey prefers to have the categorical and quantitative data. Copy and paste the raw gender and weight data into a new excel spreadsheet first. They need to be next to each other. Now click on "Edit Data". Copy and paste the two columns into StatKey and push "Ok".

A	B
Gender	Weight (Lbs)
Female	114.8
Female	149.3
Female	107.8
Female	160.1
Female	127.1
Female	123.1
Female	111.7
Female	156.3
Female	218.8
Female	110.2
Female	188.3
Female	105.4
Female	136.1
Female	182.4
Female	238.4
Female	108.8
Female	119
Female	161.9
Female	174.1
Female	181.2
Female	124.3
Female	255.9
Female	106.7
Female	149.9
Female	163.1
Female	94.3
Female	159.7
Female	162.8
Female	130
Female	179.9
Female	147.8
Female	112.9
Female	195.6
Female	124.2
Female	135
Female	141.4
Female	123.9
Female	135.5
Female	130.4
Female	100.7
male	169.1
male	144.2
male	179.3
male	175.8
male	152.6
male	166.8
male	135
male	201.5
male	175.2
male	139
male	156.3
male	186.6
male	191.1
male	151.3
male	209.4
male	237.1
male	176.7
male	220.6
male	166.1
male	137.4
male	164.2
male	162.4
male	151.8
male	144.1
male	204.6
male	193.8
male	172.9
male	161.9
male	174.8
male	169.8





Now create the bootstrap distribution by clicking on "generate 1000 samples" a few times and click on two-tail. The default is 95% but you can change the middle proportion to 0.90 or 0.99 if needed. Notice the 95% bootstrap confidence interval is (−40.239 , −12.251). This is similar to our formula calculations above.



Important notes about two-population bootstraps: Remember bootstrap confidence intervals will always come out slightly different because of sampling variability. The two numbers at the bottom of the bootstrap distribution are the upper and lower limits of the confidence interval. For two-population, we need to keep track of population 1 and population 2 and the signs of the confidence intervals. In this case, population 1 was women's weights and population 2 was men's weights and the upper and lower limits were both negative.

Example 3 (Two-population proportion): Let's use the Math 140 survey data fall 2015 to compare the population percentage (proportion) of COC statistics students born in June (population 1) and the percentage (proportion) of COC statistics students born in December (population 2). We will assume the data met all of the assumptions for a two-population proportion confidence interval. The sample data showed that of the 336 total COC statistics students, 15 were born in June and 41 were born in December. We can use Statcato and a 90% confidence level to calculate the two-population proportion confidence interval. Just go to the "statistics" menu, and then click on "confidence intervals" and then "two-population proportion". Some computer programs will ask if you want to pool the samples. This means that you combine the two samples before calculating the standard error. Pooling is a technique used in hypothesis testing, but we do not pool the samples for two-population proportion confidence intervals.

Statcato => Statistics => Confidence Intervals => Two population Proportion

Note: Do not pool the sample proportions for confidence intervals.

Confidence Interval - Two population proportions: confidence level = 0.9

	Number of Events	Number of trials	Proportion
Sample 1	15	336	0.045
Sample 2	41	336	0.122

Sample proportion difference = -0.077
Margin of error = 0.035
90.0%CI = (-0.1121, -0.0427)

We see that the upper and lower limits of the confidence interval are both negative. This indicates that the proportion of COC statistics students born in June (population 1) is between 0.0427 and 0.1121 lower than the proportion of COC statistics students born in December (population 2).

Two-population proportion formula: Let us look at how this was calculated. Here is the two-population proportion formula. The sample proportion for group 1 $(\hat{p}_1)$ was 15 ÷ 336 ≈ 0.04464 and the sample proportion for group 2 $(\hat{p}_2)$ was 41 ÷ 336 ≈ 0.12202. Remember the famous Z-score critical value for 90% confidence is Z = ± 1.645

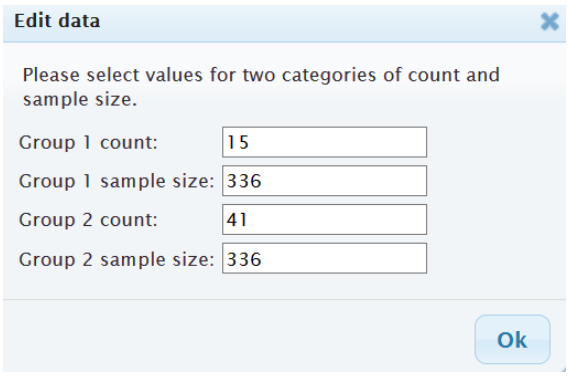
$$(\hat{p}_1 - \hat{p}_2) \pm Z\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}$$

$$(0.04464 - 0.12202) \pm 1.645\sqrt{\frac{0.04464(1-0.4464)}{336} + \frac{0.12202(1-0.12202)}{336}}$$

$$-0.07738 \pm 0.034731$$

$$(-0.1121\ ,\ -0.04265)$$

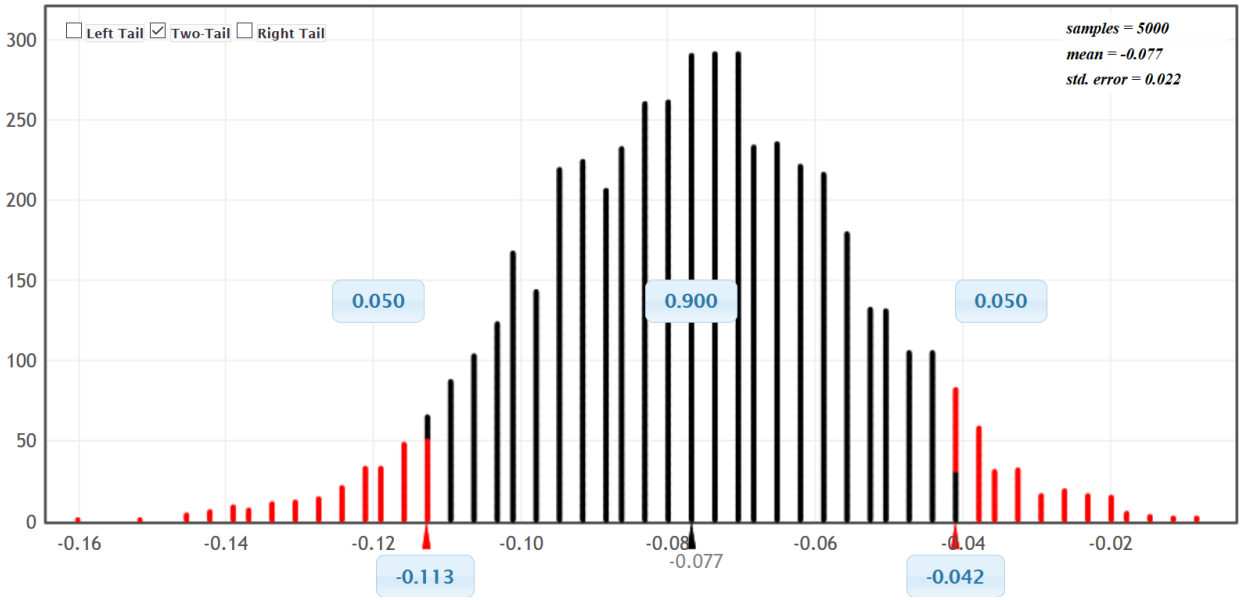
Bootstrapping: We can also calculate this confidence interval with bootstrapping. Go to www.lock5stat.com and click on "StatKey". Under the "bootstrap confidence interval" section click on "difference in proportions". Click on the "Edit Data" button, and then enter the counts and sample sizes for both groups. Remember June was group 1 and December was group 2. Then push "Ok". Now generate a few thousand bootstrap samples, click two tail, and then put "0.90" for the middle proportion.



Original Sample

Group	Count	Sample Size	Proportion
Group 1	15	336	0.045
Group 2	41	336	0.122
Group 1-Group 2	-26	n/a	-0.077

Bootstrap Dotplot of $\hat{p}_1 - \hat{p}_2$



Notice the bootstrap distribution is normal and centered close to the sample proportion difference of −0.077. It also indicates that the 90% confidence interval is (−0.113 , −0.042). This is close to what we got from the formula and Statcato.

Checking Assumptions

In order to compare populations, our sample data must be representative of the population. We usually require both samples to be large, random, and unbiased. The following assumptions are often used to check whether the sample data represents the population or not. Remember, if the sample data does not meet all of the assumptions, then we will not be able to draw any conclusions about the populations. It is also important to remember that these assumptions do not address all possible sources of bias.

Note about Independence:

- It is difficult to know for sure whether samples or individuals are indeed independent of each other. Individuals taken from two simple random samples from large populations will most likely be independent. A simple random sample of 50 people taken from a population of millions, will probably pass the individuals independent requirement. It is unlikely that we accidentally got people from the same family or people that work for the same company.

- Data collected conveniently or from voluntary response may fail the independence requirements. For example, if the sample data was collected from people in the same coffee shop or store, or on the same Facebook page, then they may be related or friends. This data would probably fail the independence requirements.

- Matched pair data means that there is a one-to-one pairing between the two samples. Usually it is the same people or objects measured twice. If the data is not matched pair, we often use the formulas for independent samples.

Notes about Bootstrapping:

- Bootstrapping does not require as many assumptions as traditional formula approaches and is often used when sample data fails the sample size requirements. However, bootstrapping does require the random and independence assumptions.

Notes about Experiments:

- Two-population confidence intervals can also be used in experimental design in order to prove cause and effect. In an experiment, the groups will not be random samples. They will need to be randomly assigned instead. Random assignment controls confounding variables.
- If the experiment uses random assignment, passes the assumptions, and shows a significant difference between the groups, then it indicates cause and effect.

Two-population Mean Assumptions (Matched Pair)

- The quantitative ordered pair sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- There should be at least thirty ordered pairs or the differences should have a nearly normal shape.

Two-population Mean Assumptions (Not Matched Pair)

- The two quantitative samples should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- Data values between the two samples should be independent of each other.
- The sample sizes should be at least 30 or have a nearly normal shape.

Two-population Proportion Assumptions

- The two categorical samples should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- Data values between the samples should be independent of each other.
- There should be at least ten successes and at least ten failures.

Two-population Bootstrap Assumptions

- The sample data should be collected randomly or be representative of the population.
- Data values within each sample should be independent of each other.
- If multiple samples were collected that were not matched pair, then the data values between the samples should be independent of each other.

Checking Assumptions Example 1: Earlier, we used the Math 140 survey data from fall 2015 to compare the population percentage (proportion) of COC statistics students born in June (population 1) and the percentage (proportion) of COC statistics students born in December (population 2). The sample data showed that of the 336 total COC statistics students, 15 were born in June and 41 were born in December. Would this data meet all of the assumptions for two-population confidence intervals with the traditional formula approach?

Traditional Formula Assumptions for Comparing Two-Population Proportions (Percentages)

- The two categorical samples should be collected randomly or be representative of the population.

  No. The month a person is born in is categorical; however, the sample data was not collected randomly. It was a census of all statistics students in the fall 2015 semester. Occasionally, data that is not collected randomly may still be representative. If we consider our population of interest as all statistics students from all semesters then this data may still be representative of the population of interest, even though it is not random.

- Both samples should have at least 10 successes and at least 10 failures.

  Yes. There were 15 students born in June and 41 in December. Both of these are greater than 10. There were 336 – 15 = 321 students NOT born in June and 336 – 41 = 295 NOT born in December. Both of these are greater than 10.

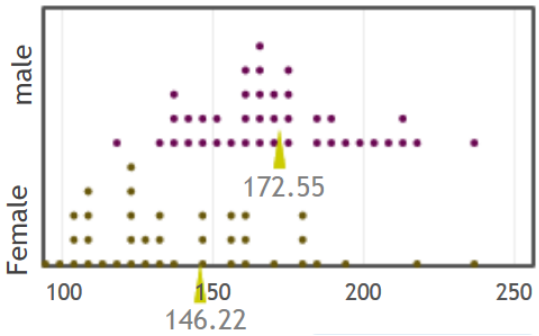
- Data values within each sample and between the samples should be independent of each other.

  No. This data likely fails the independence requirements. Many students came from the same statistics classes.

Checking Assumptions Example 2: earlier in this section, we used the randomly collected health data to compare the population mean weight of women to the population mean weight of men. Were these groups matched pair? Would this data pass the traditional formula assumptions for comparing the means? The sample data is given below.

Original Sample

$\bar{x}_1 - \bar{x}_2 = -26.33, n_1 = 40, n_2 = 40$



A random sample of men and women are not matched pair. They were not husband and wife or brother and sister. Therefore, we will proceed with checking the assumptions for independent groups.

Traditional Formula Assumptions for Comparing Two-Population Means from Independent Groups

- Both samples should be random and quantitative.

  Yes. Weights are quantitative and these were both random samples.

- Each sample should be Either Nearly Normal (almost bell shaped) OR have a Sample size of at least 30.

Women: Yes. The dot plot for the women's weight data shows that it is skewed right. So our sample size must be over 30 for it to pass. The sample size is 40, which is greater than the 30 requirement. Even though the shape was not normal, it still passes the at least 30 or normal requirement.

Men: Yes. The dot plot for the men's weight data shows that it is nearly normal. The sample size is 40, which is greater than the 30 requirement.

- Data values within the samples and between the samples should be independent of each other.

  Yes. A small random sample of 40 men and 40 women taken from millions of men and women in the population, are not likely to be related or know each other.

Problems Section 2F

1. What are the assumptions we should check if we want to use sample data to calculate a two-population proportion confidence interval?

2. What are the assumptions we should check if we want to use sample data to calculate a two-population mean confidence interval?

3. What are the assumptions we should check if we want to use sample data to calculate a two-population bootstrap confidence interval?

(#4-12) Answer the following questions. Assume the confidence intervals met the assumptions.

4. A two-population mean confidence interval is (+3.4 kg , +5.9 kg). They used a 90% confidence level.

   a) Is the mean from population 1 significantly higher, significantly lower, or not significantly different from the mean from population 2? Explain how you know.
   b) How much higher could the population mean from population 1 than the population mean from population 2?
   c) Write the two-population confidence interval sentence explaining this confidence interval.

5. A two-population proportion confidence interval is (−0.115 , −0.068). They used a 95% confidence level.

   a) Is the percentage from population 1 significantly higher, significantly lower, or not significantly different from the percentage from population 2? Explain how you know.
   b) How much lower could the percentage from population 1 be than the percentage from population 2?
   c) Write the two-population confidence interval sentence explaining this confidence interval.

6. A two-population mean confidence interval is (−16.4°F , +8.2°F). They used a 99% confidence level.

   a) Is the mean from population 1 significantly higher, significantly lower, or not significantly different from the mean from population 2? Explain how you know.
   b) Write the two-population confidence interval sentence explaining this confidence interval.

7. A two-population proportion confidence interval is (−0.045 , +0.038). They used a 90% confidence level.

   a) Is the percentage from population 1 significantly higher, significantly lower, or not significantly different from the percentage from population 2? Explain how you know.
   b) Write the two-population confidence interval sentence explaining this confidence interval.

8. A two-population mean confidence interval is (−$185.71 , −$103.62). They used a 95% confidence level.

   a) Is the mean from population 1 significantly higher, significantly lower, or not significantly different from the mean from population 2? Explain how you know.
   b) How much lower could the population mean from population 1 than the population mean from population 2?
   c) Write the two-population confidence interval sentence explaining this confidence interval.

9. A two-population proportion confidence interval is (+0.049 , +0.058). They used a 99% confidence level.

   a) Is the percentage from population 1 significantly higher, significantly lower, or not significantly different from the percentage from population 2? Explain how you know.
   b) How much higher could the percentage from population 1 be than the percentage from population 2?
   c) Write the two-population confidence interval sentence explaining this confidence interval.

10. A two-population mean confidence interval is (−6.233℃ , −4.718℃). They used a 90% confidence level.

    a) Is the mean from population 1 significantly higher, significantly lower, or not significantly different from the mean from population 2? Explain how you know.
    b) How much lower could the population mean from population 1 than the population mean from population 2?
    c) Write the two-population confidence interval sentence explaining this confidence interval.

11. A two-population proportion confidence interval is (−0.071 , +0.068). They used a 95% confidence level.

    a) Is the percentage from population 1 significantly higher, significantly lower, or not significantly different from the percentage from population 2? Explain how you know.
    b) Write the two-population confidence interval sentence explaining this confidence interval.

12. A two-population mean confidence interval is (+32.8 cm , +37.1 cm). They used a 99% confidence level.

    a) Is the mean from population 1 significantly higher, significantly lower, or not significantly different from the mean from population 2? Explain how you know.
    b) How much higher could the population mean from population 1 than the population mean from population 2?
    c) Write the two-population confidence interval sentence explaining this confidence interval.

(#13-20) Directions: Use the following Statcato and StatKey printouts and answer the following questions.

a) Does the data meet the assumptions for inference with two population proportions or two population means? If it is two means, are the groups independent or matched pair? List the assumptions needed and how the problem meets them or does not meet them.

b) Give the sample means or sample proportions for the two groups. Are they close or significantly different? Explain how you know. If they are significantly different, which group has a significantly higher sample mean or sample proportion?

c) Does the confidence interval indicate that the mean or percentage from population 1 is higher, lower, or not significantly different from population 2? Explain how you know. If the mean or percentage from population 1 is higher than population 2, then how much higher could it be? If the mean or percentage from population 1 is lower than population 2, then how much lower could it be?

d) Write the two-population confidence interval sentence explaining this confidence interval.

13. The ACT exam is used by many colleges to test the readiness of high school students for college. Many high school students are now taking ACT prep classes. A local high school offers an ACT prep class, but wants to know if it really helps. Twenty-eight students were randomly selected. They took the ACT exam before and after taking the ACT prep class. Population 1 is the ACT scores after taking the prep class and population 2 is the ACT scores before taking the prep class. The sample mean of the differences was 5.8 ACT points and the sample standard deviation of the differences was 4.3 ACT points. A histogram of the differences was normal. We created a 90% confidence interval for matched pairs with Statcato.

Confidence Interval - Matched Pairs: confidence level = 0.9
Input: Summary data
Difference of Matched Pairs -

N	Mean	Stdev	Margin of Error	90.0%CI
28	5.8	4.3	1.384	(4.4159, 7.1841)

14. We want to compare the population percentage of women that have at least one tattoo (π_1) and the population percentage of men that have at least one tattoo (π_2). A random sample of 794 women found that 137 of them had at least one tattoo. A random sample of 857 men found that 146 of them had at least one tattoo. Go to www.lock5stat.com and use StatKey to create a 99% two-population proportion bootstrap confidence interval.

15. Cotinine is an alkaloid found in tobacco and is used as a biomarker for exposure to cigarette smoke. It is especially useful in examining a person's exposure to second hand smoke. A random sample of 90 non-smoking American adults was collected. These adults were not smokers and did not live with any smokers. The average cotinine level for this sample was 7.2 ng/mL with a standard deviation of 5.8 ng/mL. A second sample of 85 non-smoking American adults was then collected. These adults did not smoke themselves, but did live with one or more smokers. The average cotinine level for this sample was 28.5 and had a standard deviation of 11.4. Population 1 was people that do NOT live with smokers (μ_1) and population 2 was people that DO live with smokers (μ_2). We used Statcato to create the following 95% two-population mean confidence interval for independent groups.

Confidence Intervals - Two population means: confidence level = 0.95

	N	Mean	Stdev
Population 1	90	7.2	5.8
Population 2	85	28.5	11.4

* Population standard deviations are unknown. *
DOF = 123
Margin of error = 2.730
95.0%CI = (-24.0304, -18.5696)

16. A body mass index of 20-25 indicates that a person is of normal weight. Use the following 90% two-population proportion confidence interval to compare the percentage of men with a normal BMI (π_1) and the percentage of women with a normal BMI (π_2). A random sample of 745 women and 760 men found that 198 of the women and 273 of the men had a normal BMI score.

Confidence Interval - Two population proportions: confidence level = 0.9

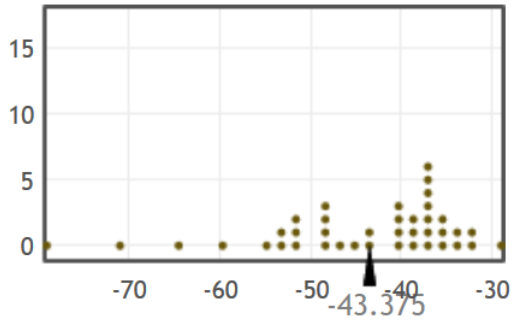
	Number of Events	Number of trials	Proportion
Sample 1	273	760	0.359
Sample 2	198	745	0.266

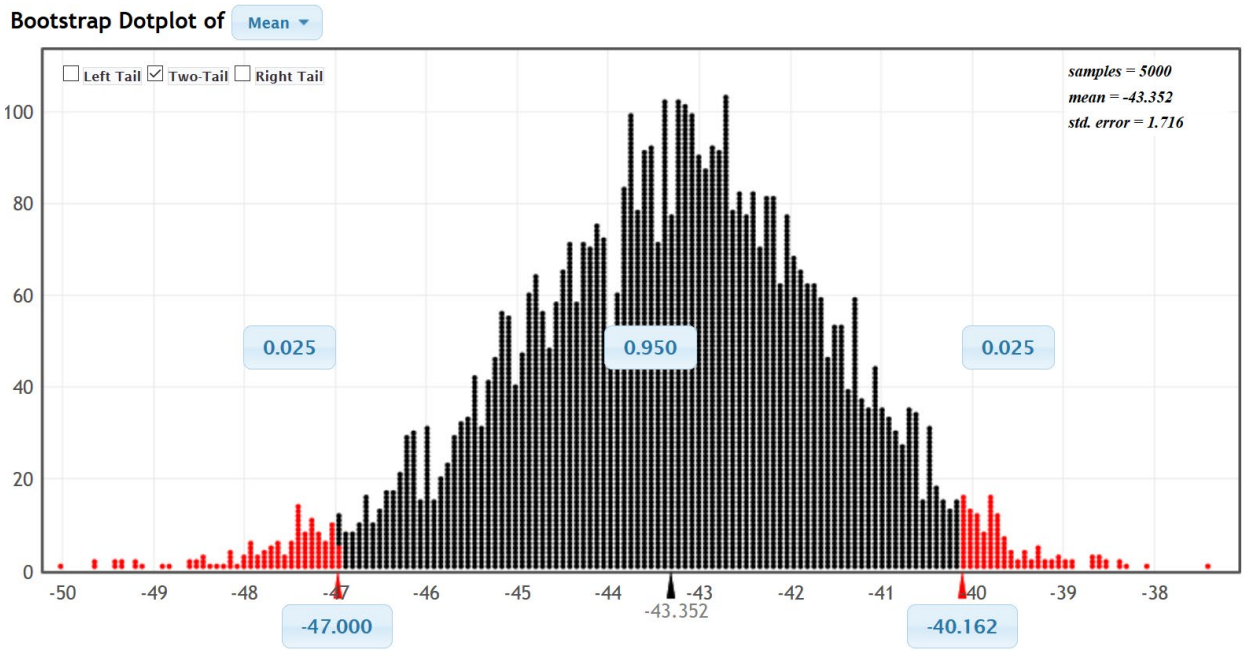
Sample proportion difference = 0.093
Margin of error = 0.039
90.0%CI = (0.0543, 0.1325)

17. We used the random health data at www.matt-teachout.org to compare the population mean average systolic and diastolic blood pressures for women. Population 1 was women's diastolic blood pressure and population 2 was women's systolic blood pressure. We used StatKey to create the following 95% bootstrap confidence interval of the differences between the matched pairs.

Original Sample

n = 40, mean = -43.375
median = -40, stdev = 10.748





18. An experiment was done to test the effectiveness of medicine that lowers cholesterol. An experiment was conducted and adults were randomly assigned into two groups. The groups had similar gender, ages, exercise patterns and diet. Of the 410 adults in the treatment group, 49 of them showed a decrease in cholesterol. Of the 420 adults in the placebo group, 38 of them showed a decrease in cholesterol. Was the medicine effective in lowering cholesterol? Use the following 99% confidence interval from Statcato to determine if the percentage of people on the medicine that have a decrease in cholesterol (population 1) is higher than the percentage from the placebo group (population 2).

Confidence Interval - Two population proportions: confidence level = 0.99

	Number of Events	Number of trials	Proportion
Sample 1	49	410	0.120
Sample 2	38	420	0.090

Sample proportion difference = 0.029
Margin of error = 0.055
99.0%CI = (-0.0258, 0.0838)

19. Open the Health data at www.matt-teachout.org. Copy and paste the gender data and cholesterol data into a new excel spreadsheet so that they are next to each other. Go to www.lock5stat.com can click on StatKey. Under the Bootstrap Confidence Interval menu, click on "CI for Difference in Means". Under the "edit data" menu, copy and paste the gender and cholesterol data into StatKey. Construct a 95% two-population mean bootstrap confidence interval estimate of the difference between women's population mean average cholesterol (μ_1) and men's population mean average cholesterol (μ_2).

20. In March 2003, a research group asked 2400 randomly selected Americans whether they believe that the U.S. made the right or wrong decision to use military force in Iraq. Of the 2400 adults, 1862 said that they believed that the U.S. did make the correct decision. In February 2008, the question was asked again to 2180 randomly selected Americans and 684 of them said that the U.S. did make the correct decision. Go to www.lock5stat.com and use StatKey to create a 90% two-population proportion bootstrap confidence interval to compare the population percentage of people that agree with war in 2008 (π_1) and the population percentage in 2003 (π_2).

Section 2G – One-Population Variance & Standard Deviation Confidence Intervals

It is often vital to estimate the standard deviation of a population. However, it can be very difficult to estimate with any accuracy, especially if we only have one random sample. Remember our principle of sampling variability. We saw in previous sections that sample standard deviations (s) will usually be very different from each other and can be very different from the population standard deviation (σ).

One-Population Variance and Standard Deviation Confidence Intervals

Recall that the population standard deviation is the square root of the population variance (σ^2). So we often estimate the population variance and then simply take the square root of the variance to get the standard deviation. The principle of sampling variability also applies to variance. Sample variances (s^2) will usually be very different from each other and may be very different from the population variance (σ^2).

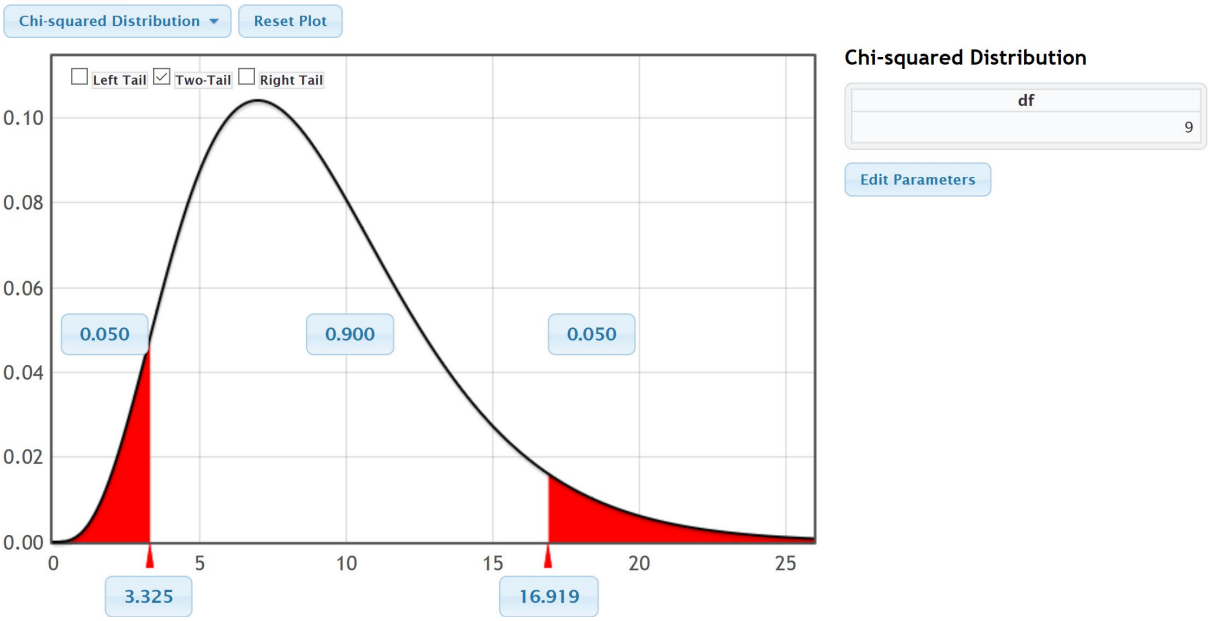
Sampling distributions for variance are usually skewed to the right and rarely have a normal shape. Since the sampling distribution is not normal or symmetric, we cannot use the traditional formula approach of the sample statistic ± margin of error. That formula will not work.

It is important to note that even though the quantitative data itself may be normal, the sampling distribution for variance still may be skewed to the right. Statisticians discovered that as long as the quantitative data itself was normal, the sampling distribution for sample variance follows a Chi-Squared distribution with degrees of freedom $(df = n - 1)$. So the formula for making a confidence interval to estimate population variance uses Chi-Squared critical values (χ^2). It is important to note that no matter what the sample size is, the sample data must be normal for this formula to work. If the data is not normal, we must resort to another technique like bootstrapping.

Calculating Chi-Squared Critical Values

Example 1: Calculate the Chi-Squared (χ^2) critical values for a sample size n = 10 and a 90% confidence level.

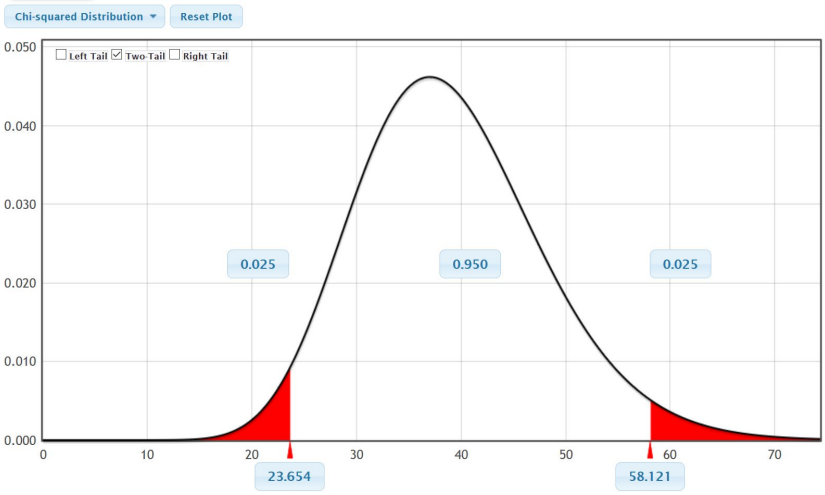
Go to www.lock5stat.com and click on "StatKey". Under the "theoretical distributions" menu, click on "χ^2". Since the sample size is 10, the degrees of freedom will be df = 10 − 1 = 9. If we click on "two tail" and set the middle proportion to 0.90, we will get the following. Variance is calculated with a sum of squares. That makes it impossible to ever be negative. That also means the upper and lower Chi-Squared critical values will be very different. The upper Chi-Squared critical value will be the larger number on the right and the lower Chi-Squared critical value will be the smaller number on the left. Both will be positive. You can see that Chi-Squared looks skewed to the right for nine degrees of freedom.



We see from the graph that upper critical value for 90% confidence and 9 degrees of freedom is 16.919 and the lower critical value for 90% confidence and 9 degrees of freedom is 3.325.

Example 2: Calculate the Chi-Squared (χ^2) critical values for a sample size n = 40 and a 95% confidence level.

Go to www.lock5stat.com and click on "StatKey". Under the "theoretical distributions" menu, click on "χ^2". Since the sample size is 40, the degrees of freedom will be df = 40 − 1 = 39. If we click on "two tail" and set the middle proportion to 0.95, we will get the following. Notice that the upper and lower Chi-Squared critical values will both be positive, but will be very different. Also, notice that as the degrees of freedom increases, the chi-squared distribution looks less skewed to the right.



We see from the graph that upper critical value for 95% confidence and 39 degrees of freedom is 58.121 and the lower critical value for 95% confidence and 39 degrees of freedom is 23.654.

Confidence Interval Formulas for Variance and Standard Deviation

Here is the confidence interval formulas for estimating population variance. Taking the square root gives us the formula for estimating population standard deviation as well. Notice that the upper critical value is on the left and lower critical value is on the right. When you divide by a larger number, the overall fraction is smaller.

$\frac{s^2(n-1)}{\chi^2_{upper}}$ < Population Variance (σ^2) < $\frac{s^2(n-1)}{\chi^2_{lower}}$

$\sqrt{\frac{s^2(n-1)}{\chi^2_{upper}}}$ < Population Standard Deviation (σ) < $\sqrt{\frac{s^2(n-1)}{\chi^2_{lower}}}$

Example 1: We measured the heights in inches of 40 randomly selected men. The data showed a normal shape. The sample standard deviation was 3.020 inches. Use the Chi-squared critical values and the formulas above to create a 95% confidence interval for the population variance and the population standard deviation.

We calculated the Chi-squared critical values in the previous example. The upper critical value was 58.121 and the lower critical value was 23.654.

Sample Variance $(s^2) = (3.020)^2 = 9.1204$

Degrees of Freedom $(n-1) = 40 - 1 = 39$

$\frac{s^2(n-1)}{\chi^2_{upper}}$ < Population Variance (σ^2) < $\frac{s^2(n-1)}{\chi^2_{lower}}$

$\frac{9.1204\ (40-1)}{58.121}$ < Population Variance (σ^2) < $\frac{9.1204\ (40-1)}{23.654}$

$\frac{9.1204\ (39)}{58.121}$ < Population Variance (σ^2) < $\frac{9.1204\ (39)}{23.654}$

6.11992 < Population Variance (σ^2) < 15.03744

Variance Confidence Interval Sentence: We are 95% confident that the population variance for all men is between 6.11992 and 15.03744 square inches. *(Notice that variance is in square units since it is the standard deviation squared.)*

If we take the square root of our answers, we can get an estimate of the population standard deviation.

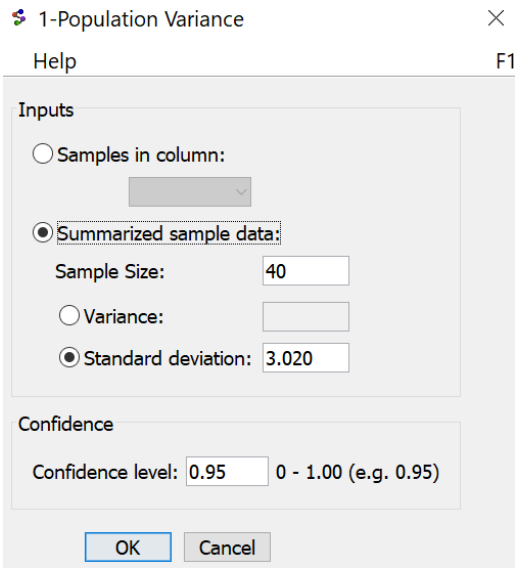
$\sqrt{6.11992}$ < Population Standard Deviation (σ) < $\sqrt{15.03744}$

2.47 inches < Population Standard Deviation (σ) < 3.88 inches

Standard Deviation Confidence Interval Sentence: We are 95% confident that the population standard deviation for all men is between 2.47 inches and 3.88 inches.

As with all calculations, it is much easier and more accurate to calculate these with a computer program.

In Statcato, we can go to the "Statistics" menu and click on confidence intervals. If we click on "1-population variance" and enter the sample size (40) and sample standard deviation 3.020 under summary data, we get the following. Notice you can also calculate the confidence interval from raw data or by entering the sample variance.



Notice Statcato gave us almost the same confidence intervals for variance and standard deviation as we calculated with the formula.

Confidence Interval - One population variance: confidence level = 0.95

Input: Summary data

N	Variance	Stdev	95.0%CI Variance	95.0%CI Stdev
40	9.120	3.02	(6.1200, 15.0372)	(2.4739, 3.8778)

Here are the assumptions for making a confidence interval to estimate population variance or standard deviation.

<u>One-Population Variance or Standard Deviation Confidence Interval Assumptions</u>

1. The quantitative sample data should be collected randomly or be representative of the population.
2. Data values within the sample should be independent of each other.
3. The sample data must be normal.

Does the men's height data meet these assumptions? Let us check them.

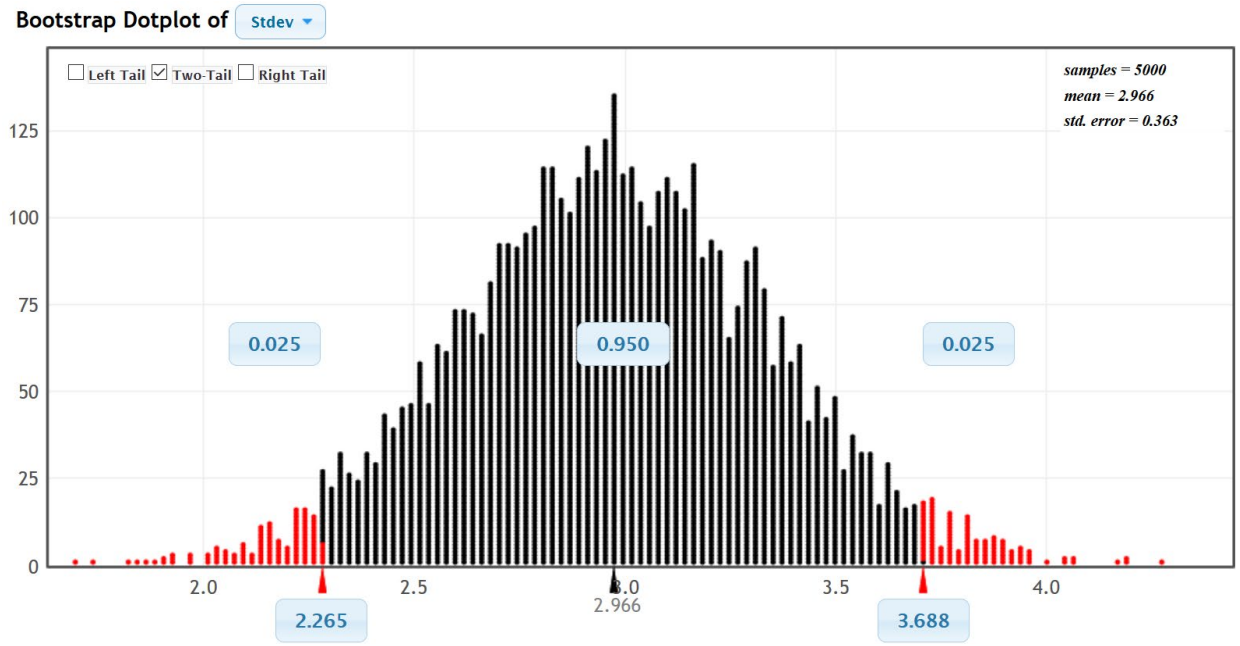
1. Is this random quantitative sample data or sample data that represents the population? Yes. Height is quantitative and this was a random sample.

2. Are the data values within the sample independent of each other? Yes. A random sample out of a large population will be unlikely to accidently get men that are family members. One man's height should not change the probability of another man's height.

3. Is the sample data normal? We did not see the histogram, but the problem did state that the data was normal.

Men's Height Bootstrap Example

Let us use a bootstrap distribution to estimate the confidence interval for population standard deviation for men's height. First go to the "Health Data" at www.matt-teachout.org and copy the men's height column of data. Now go to the "Bootstrap Confidence Interval" menu in StatKey at www.lock5stat.com and click on "CI for Single Mean, Median, St.Dev." Under "Edit Data", paste in the raw quantitative men's height data. Make sure to check the "Header Row" box since this data set had a title and push "OK". Under the "Generate Dot plot of" menu, click on "St.Dev." (standard deviation). Now click "Generate 1000 Samples" a few times. Then click "Two-Tail". Make sure the middle proportion is 0.95 (95%).



Notice that the upper and lower limits of the confidence interval are close to what we got with formula or Statcato.

Problems Section 2G

1. What assumptions should we check if we want to use sample data to create a confidence interval to estimate a population standard deviation or variance?

2. A random sample of 45 high school students ACT exams has a skewed left distribution with a sample standard deviation (s) of 9.868. What is the sample variance (s^2)? What is the degrees of freedom? Open StatKey at www.lock5stat.com, go to "Theoretical Distributions" and then click on "χ^2". Look up the Chi-squared critical values. Use the critical values, degrees of freedom (n−1) and sample variance to construct a 90% confidence interval estimate of the population variance for all ACT exams. Take the square root of your variance confidence interval to calculate a 90% confidence interval estimate for the population standard deviation.

Variance Confidence Interval Formula:

$\frac{s^2(n-1)}{\chi^2_{upper}}$ < Population Variance (σ^2) < $\frac{s^2(n-1)}{\chi^2_{lower}}$

a) Sample Variance = $(Sample\ Standard\ Deviation)^2$ =

b) Degrees of Freedom = $n-1$ =

c) Chi-squared upper critical value =

d) Chi-squared lower critical value =

e) Variance Confidence Interval =

f) Standard Deviation Confidence Interval =

3. Use the following Statcato printout to check your variance confidence interval answer and your standard deviation confidence interval answer from the random sample ACT data in number 2. Check the assumptions for a variance confidence interval. Remember the data was skewed left. Write down a sentence to explain the population variance confidence interval. Write down a sentence to explain the population standard deviation confidence intervals.

Confidence Interval - One population variance: confidence level = 0.9

Input: Summary data

N	Variance	Stdev	90.0%CI Variance	90.0%CI Stdev
45	97.377	9.868	(70.8423, 143.8391)	(8.4168, 11.9933)

a) Check each of the assumptions for this problem. Explain your answers.

b) Write down a sentence to explain the population variance confidence interval.

c) Write down a sentence to explain the population standard deviation confidence interval.

4. A random sample of body temperatures in degrees Fahrenheit was taken from 50 randomly selected adults. Assume the sample data was normally distributed. The sample standard deviation of 0.765 °F. What is the sample variance (s^2)? What is the degrees of freedom? Open StatKey at www.lock5stat.com, go to "Theoretical Distributions" and then click on "χ^2". Look up the Chi-squared critical values. Use the critical values, degrees of freedom (n−1) and sample variance to construct a 99% confidence interval estimate of the population variance for all body temperatures. Take the square root of your variance confidence interval to calculate a 99% confidence interval estimate for the population standard deviation.

Variance Confidence Interval Formula:

$\frac{s^2(n-1)}{\chi^2_{upper}}$ < Population Variance (σ^2) < $\frac{s^2(n-1)}{\chi^2_{lower}}$

a) Sample Variance = $(Sample\ Standard\ Deviation)^2$ =

b) Degrees of Freedom = $n-1$ =

c) Chi-squared upper critical value =

d) Chi-squared lower critical value =

e) Variance Confidence Interval =

f) Standard Deviation Confidence Interval =

5. Use the following Statcato printout to check your variance confidence interval answer and your standard deviation confidence interval answer from the random sample body temperature data in number 4. Check the assumptions for a variance confidence interval. Remember the data is normally distributed. Write down a sentence to explain the population variance confidence interval. Write down a sentence to explain the population standard deviation confidence intervals.

Confidence Interval - One population variance: confidence level = 0.99

Input: Summary data

N	Variance	Stdev	99.0%CI Variance	99.0%CI Stdev
50	0.585	0.765	(0.3666, 1.0524)	(0.6054, 1.0258)

a) Check each of the assumptions for this problem. Explain your answers.

b) Write down a sentence to explain the population variance confidence interval.

c) Write down a sentence to explain the population standard deviation confidence interval.

6. Go to www.matt-teachout.org, click on "Statistics" and then "Data Sets". Open the "cereal data" in excel. Copy the column of data labeled "sugar (grams per serving)". Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Mean, Median, St.Dev." Click on "Bootstrap Dot plot of Stdev". Now click on "Edit Data" and paste the sugar data into StatKey. Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the standard deviation. Use the bootstrap distribution to find a 99% confidence interval for the population standard deviation.

a) Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.

b) How many bootstrap samples did you take?

c) What is the shape of the bootstrap distribution for the standard deviation?

d) Write the upper and lower limits of the bootstrap confidence interval for the population standard deviation.

e) Write a sentence to explain the bootstrap confidence interval estimate of the population standard deviation.

7. Go to www.matt-teachout.org, click on "Statistics" and then "Data Sets". Open the "cereal data" in excel. Copy the column of data labeled "carbs (grams per serving)". Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Mean, Median, St.Dev." Click on "Bootstrap Dot plot of Stdev". Now click on "Edit Data" and paste the carb data into StatKey. Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the standard deviation. Use the bootstrap distribution to find a 95% confidence interval for the population standard deviation.

a) Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.

b) How many bootstrap samples did you take?

c) What is the shape of the bootstrap distribution for the standard deviation?

d) Write the upper and lower limits of the bootstrap confidence interval for the population standard deviation.

e) Write a sentence to explain the bootstrap confidence interval estimate of the population standard deviation.

8. Go to www.matt-teachout.org, click on “Statistics” and then “Data Sets”. Open the “bear data” in excel. Copy the column of data labeled “weight in pounds”. Go to www.lock5stat.com and click on the “StatKey” tab. Under the “Bootstrap Confidence Intervals” menu, click on “CI for Single Mean, Median, St.Dev.” Click on “Bootstrap Dot plot of Stdev”. Now click on “Edit Data” and paste the bear weight data into StatKey. Click on “Generate 1000 Samples” a few times to create the bootstrap sampling distribution for the standard deviation. Use the bootstrap distribution to find a 90% confidence interval for the population standard deviation for the weight of bears.

a) Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.

b) How many bootstrap samples did you take?

c) What is the shape of the bootstrap distribution for the standard deviation?

d) Write the upper and lower limits of the bootstrap confidence interval for the population standard deviation.

e) Write a sentence to explain the bootstrap confidence interval estimate of the population standard deviation.

9. Go to www.matt-teachout.org, click on “Statistics” and then “Data Sets”. Open the “bear data” in excel. Copy the column of data labeled “length in inches”. Do not click on “head length” by mistake. We want the overall length of the bears. Go to www.lock5stat.com and click on the “StatKey” tab. Under the “Bootstrap Confidence Intervals” menu, click on “CI for Single Mean, Median, St.Dev.” Click on “Bootstrap Dot plot of Stdev”. Now click on “Edit Data” and paste the bear length data into StatKey. Click on “Generate 1000 Samples” a few times to create the bootstrap sampling distribution for the standard deviation. Use the bootstrap distribution to find a 99% confidence interval for the population standard deviation for the length of bears.

a) Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.

b) How many bootstrap samples did you take?

c) What is the shape of the bootstrap distribution for the standard deviation?

d) Write the upper and lower limits of the bootstrap confidence interval for the population standard deviation.

e) Write a sentence to explain the bootstrap confidence interval estimate of the population standard deviation.

Chapter 2 Review Problems

Topics to Study

- Confidence Interval Key Terms
- Statistics and Parameters
- Sampling Distributions
- Know how to interpreting confidence intervals
- T-distribution
- Table summarizing critical value, standard error, margin of error and confidence intervals
- Confidence Interval Assumptions
- Bootstrapping
- Two-population confidence intervals

1. Determine if each of the following symbols are a mean, standard deviation, proportion, slope, or correlation coefficient. Also, decide if it is a sample statistic or a population parameter.

(N , n , π , $\hat{p}$, μ , $\bar{x}$, σ , s , ρ , r , β_1 , b_1, σ^2, s^2)

2. For each number determine the symbol used from the following list and if it is a statistic or a parameter.

(N , n , π , $\hat{p}$, μ , $\bar{x}$, σ , s , ρ , r , β_1 , b_1, σ^2, s^2)

a) "We tested a sample of incoming college freshman and found that their sample mean average IQ was 101.9, a sample standard deviation of 14.8 and a sample variance of 219.04. We think the population mean IQ is 100, the population standard deviation for IQ scores is 15, and the population variance is 225."

b) "We want to check and see if the population correlation coefficient could be zero and the population slope could be about 20 pounds per degree Fahrenheit. The sample correlation coefficient was 0.338 and the sample slope was 13.79 pounds per degree Fahrenheit."

c) "Our study found that of the people tested in the sample, only 3% showed side effects to the medication. We think the population percentage of side effects is closer to 1.5%".

d) "We took a random sample of 238 people from a population of about 5 million people."

3. List the assumptions that need to be checked before you make a one-population mean confidence interval.

4. List the assumptions that need to be checked before you make a one-population variance or standard deviation confidence interval.

5. List the assumptions that need to be checked before you make a one-population proportion confidence interval.

6. List the assumptions that need to be checked before you make a two-population mean confidence interval.

7. List the assumptions that need to be checked before you make a two-population proportion confidence interval.

8. List the assumptions for a bootstrap confidence interval.

9. Define the following terms: Population, Census, Sample, Statistic, Parameter, Sampling Distribution, Sampling Variability, Point Estimate, Margin of Error, Standard Error, Confidence Interval, 95% Confident, 90% Confident, 99% Confident, Bootstrapping, Bootstrap Sample, Bootstrap Statistic, Bootstrap Distribution

10. Write a sentence to explain the following confidence intervals. Assume the confidence intervals came from unbiased random sample data that met all of the assumptions.

a) Explain the one-population mean confidence interval (55.6 pounds, 69.4 pounds). *Confidence Level = 99%*

b) Explain the one-population proportion confidence interval (0.352 , 0.411). *Confidence Level = 90%*

c) Explain the one-population standard deviation confidence interval (3.1 pounds, 4.7 pounds). *Confidence Level = 95%*

d) Explain the one-population variance confidence interval (461.8 square inches, 591.3 square inches). *Confidence Level = 99%*

e) Explain the two-population mean confidence interval (+13.2 kg, +14.8 kg). *Confidence Level = 95%* Is there a significant difference between the populations? Explain why.

f) Explain the two-population mean confidence interval (−$3.79, +$4.13). *Confidence Level = 90%* Is there a significant difference between the populations? Explain why.

g) Explain the two-population proportion confidence interval (−0.024, +0.017). *Confidence Level = 95%* Is there a significant difference between the populations? Explain why.

h) Explain the two-population proportion confidence interval (−0.072, −0.057). *Confidence Level = 99%* Is there a significant difference between the populations? Explain why.

11. Explain what a sampling distribution is and how we can use it to find the population parameter, standard error and better understand sampling variability.

12. Explain how a critical value Z-score or T-score and standard error can be used to calculate the margin of error. How can we use margin of error to make the confidence interval.

13. In one-population variance confidence intervals, how does the computer use the chi-squared critical values, the degrees of freedom and the sample variance to make the confidence interval?

14. Answer the following questions about the T-distribution.

a) Who invented the T-distribution?
b) What company did he work for?
c) Why did he invent T-scores?
d) Why did he have to publish the T-distribution discovery under a pseudonym?
e) What pseudonym did he use?
f) When are T-scores significantly larger than Z-scores?
g) When are T-scores and Z-scores almost the same?
h) What types of confidence intervals use Z-scores?
i) What types of confidence intervals use T-scores?
j) How is degrees of freedom usually calculated for one quantitative data set?

15. Explain the ideas behind the Central Limit Theorem.

16. Explain the process of bootstrapping and how a bootstrap distribution may be used to calculate a confidence interval without a formula. What assumptions are necessary to make a bootstrap confidence interval? How is a bootstrap distribution different from a sampling distribution?
