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

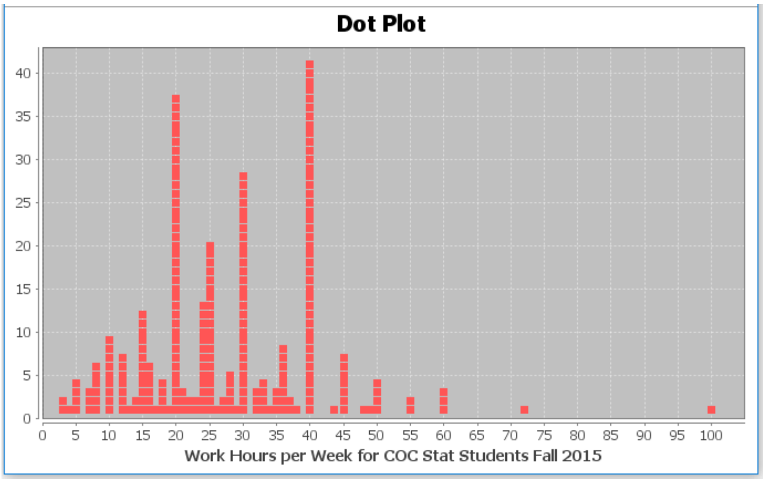
Suppose we take many, many random samples from a population. From each random sample, we calculate a statistic like the sample mean average. If we put all of those sample means on the same graph, we have created a "sampling distribution".   Sampling distributions are one of the best ways to understand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

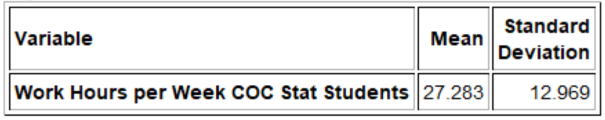
Example: Work Hours per Week for working COC Statistics Students (Fall 2015 semester)

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.

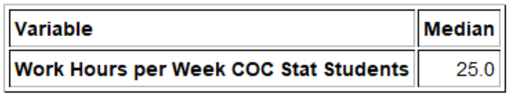
Census Data (Work Hours per Week for working COC Stat Students Fall 2015)



Population Parameters



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



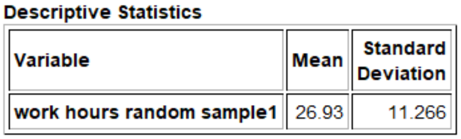
Variable	Median
Work Hours per Week COC Stat Students	25.0

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

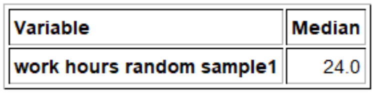
We learned in chapter 1 that random samples tend to minimize sampling bias, so are better representations of the populations than other samples that are not random. Does this mean that random samples are perfect representations of the population? Let us see.

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

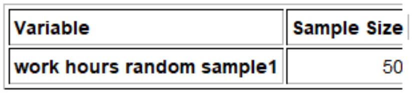


Descriptive Statistics

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

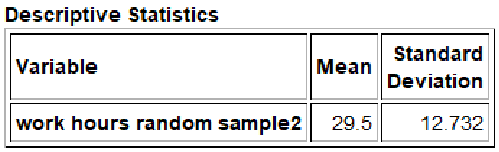
We see that the sample mean was 26.93 hours per week, the sample standard deviation was 11.266 hours per week, and the sample median was 24 hours per week. Notice that all of these sample statistics are different from the population parameters.

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

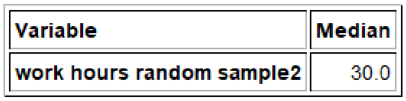
Sample 1 median = 24 hours per week
Population median = 25 hours per week

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

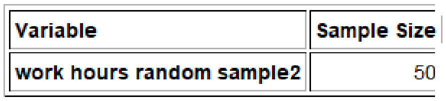


Descriptive Statistics

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.

<u>Sampling Variability</u>: 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".

<u>Margin of Error</u>: 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.

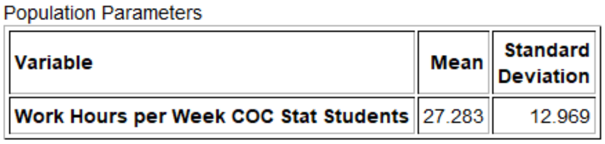
In an article published by a health website, the author states that the population average weight of all men in America is 196 pounds. As with most articles, this is a guess about the population average and is not the actual population average weight of men. We call this a "point estimate". Someone took 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.

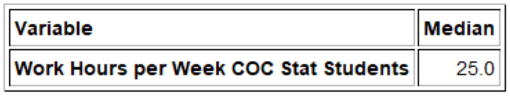
Sampling Distributions for Sample Mean Averages

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.



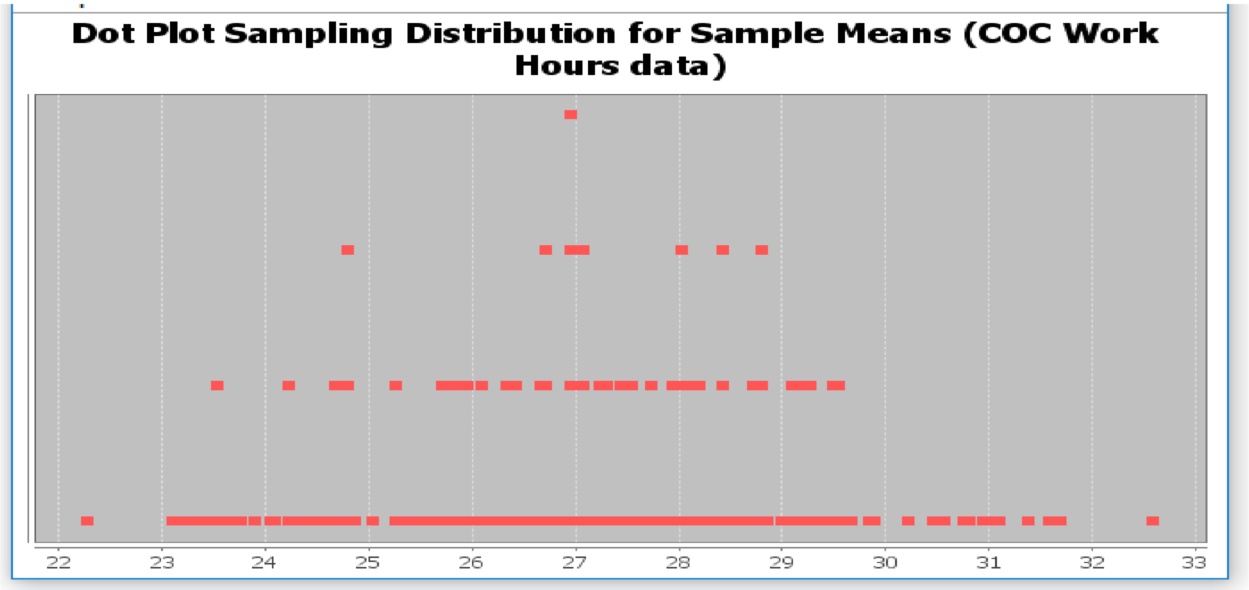
Population Parameters

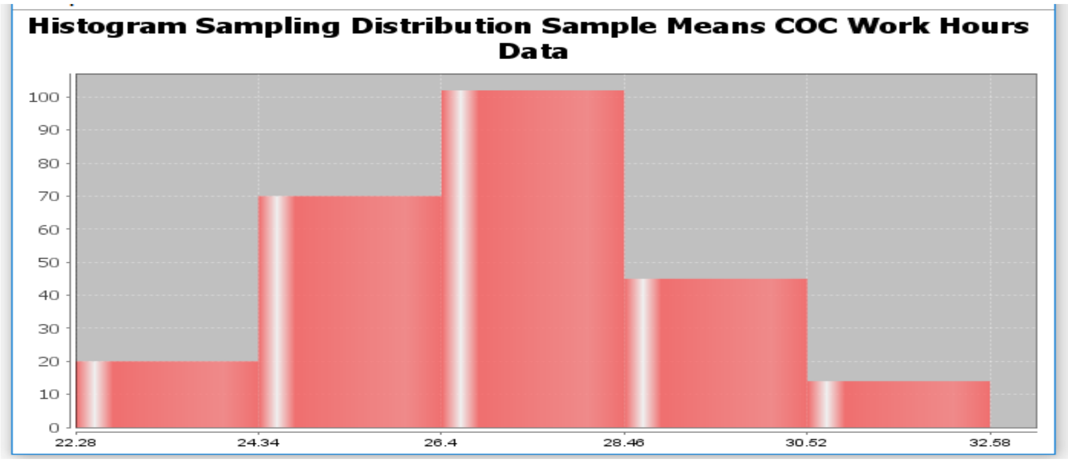
Variable	Mean	Standard Deviation
Work Hours per Week COC Stat Students	27.283	12.969



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

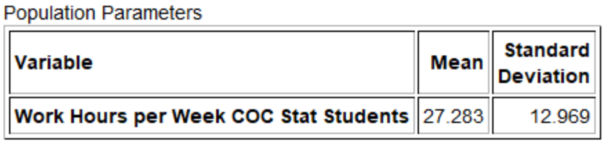
- We took 251 random samples and calculated 251 sample means. We see sampling variability in action. The population mean is 27.283 hours per week but sample means ranged between 22.28 hours and 32.58 hours. Random sample means are usually not the same as each other and can be very different from the 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

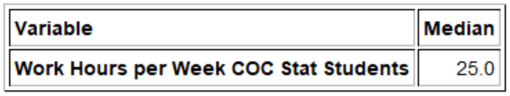
Sampling Distributions for Sample Standard Deviations

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.



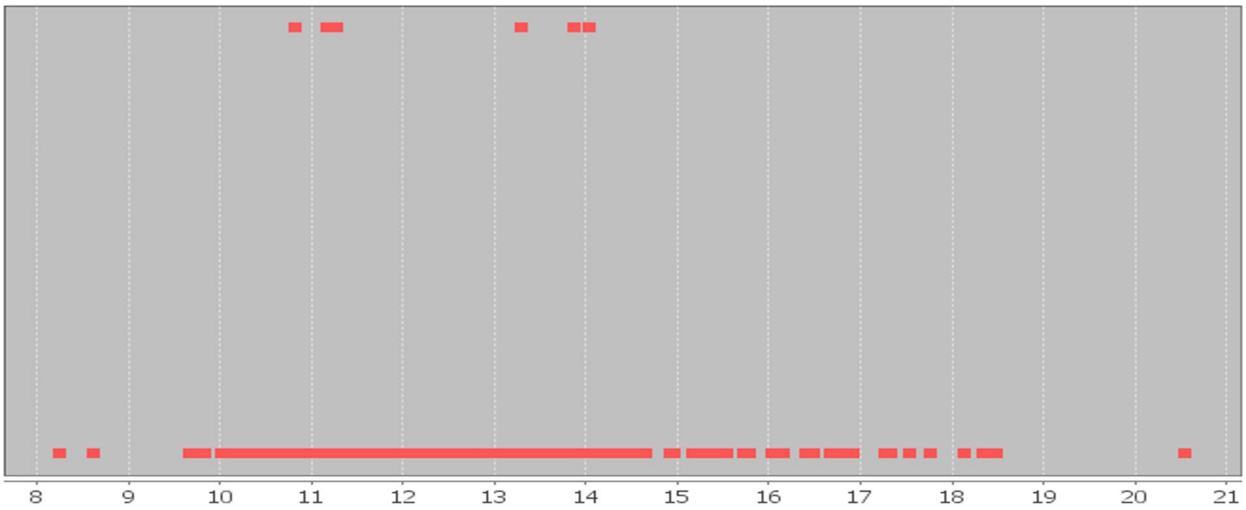
Population Parameters

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

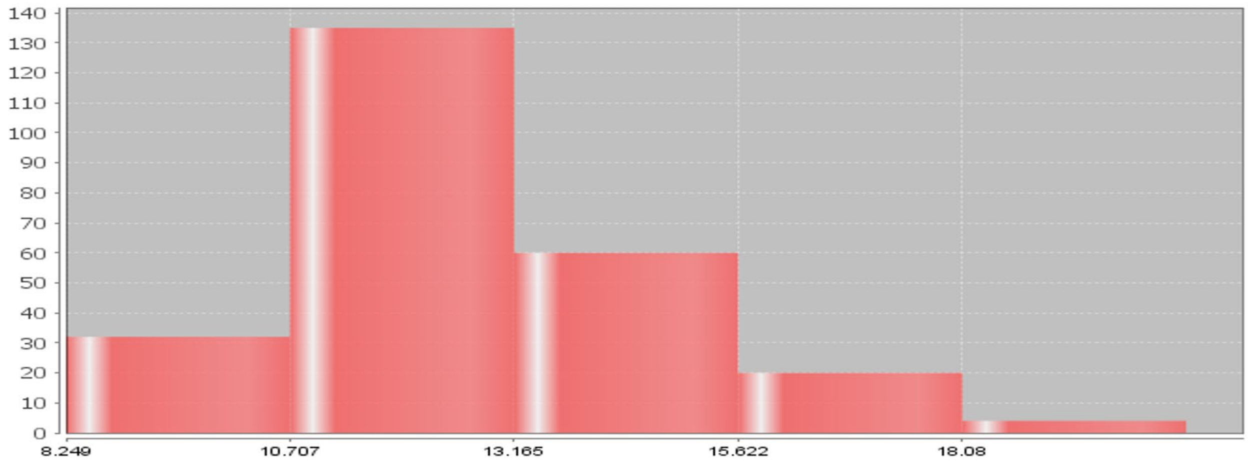


Variable	Median
Work Hours per Week COC Stat Students	25.0

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

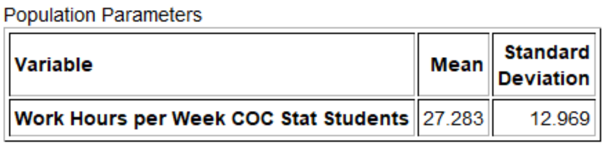
Variable	Total Number of Random Samples
Sampling Distribution for Sample Standard Deviations (COC stat students work hours)	251

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.

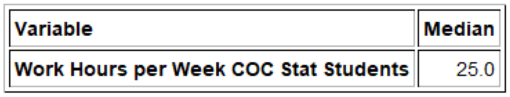
Sampling Distributions for Sample Median Averages

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.

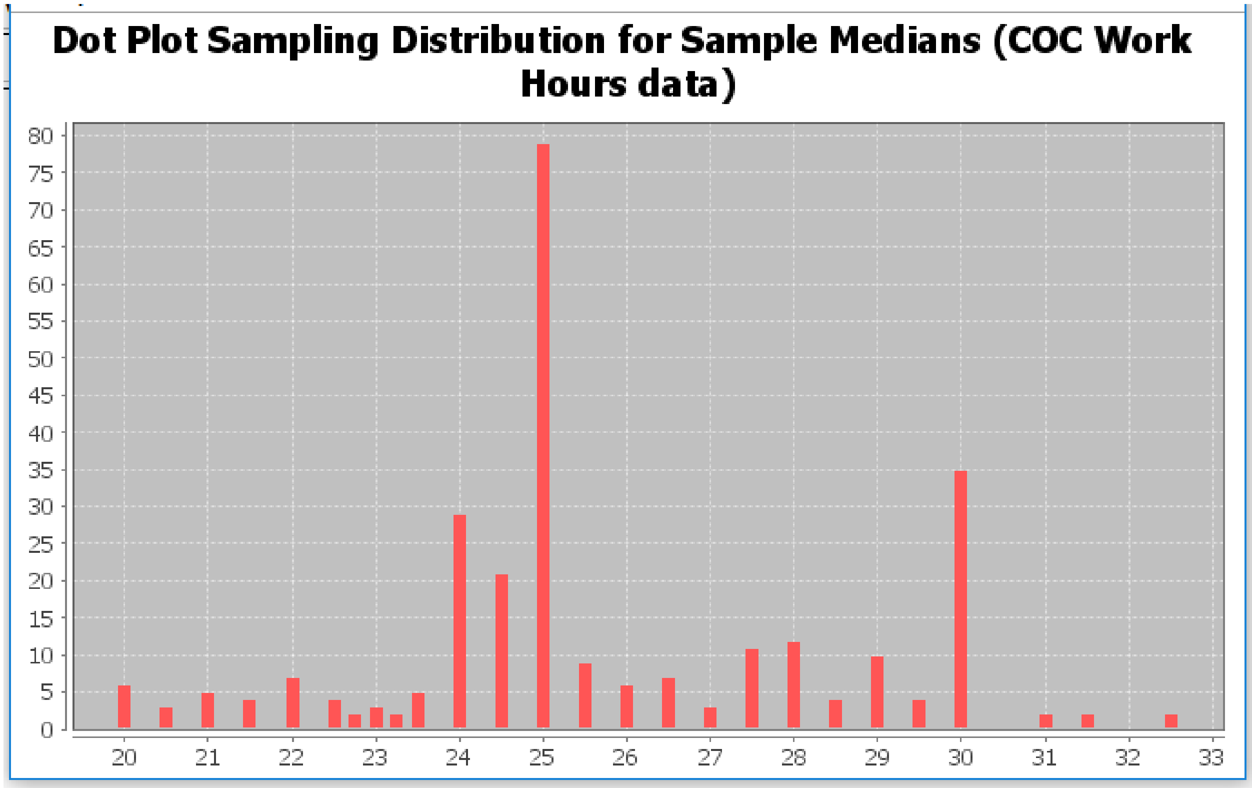


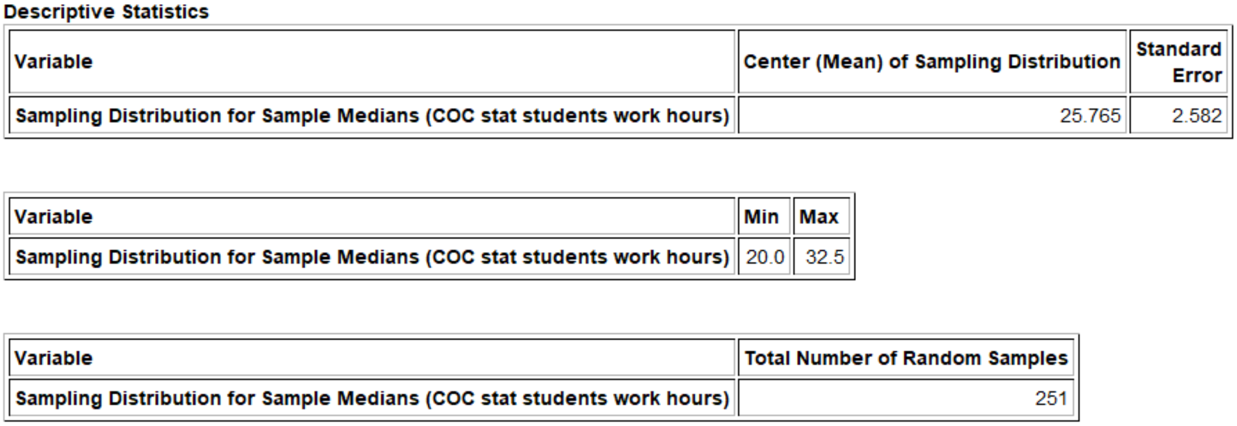
Population Parameters

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.

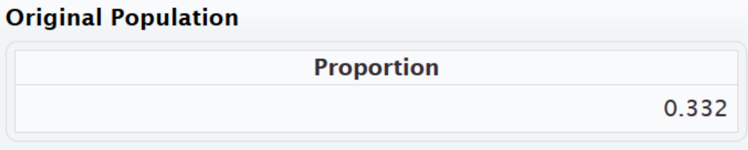
- Again, we calculated the approximate center of the sampling distribution. This is the mean average of all of the sample medians. Notice that the center of the sampling distribution is 25.765 hours and is closer to the population median of 25 hours per week. Since this data was skewed to the right, 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.

<u>Sampling Distributions for Sample Proportions (Sample Percentages)</u>

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?

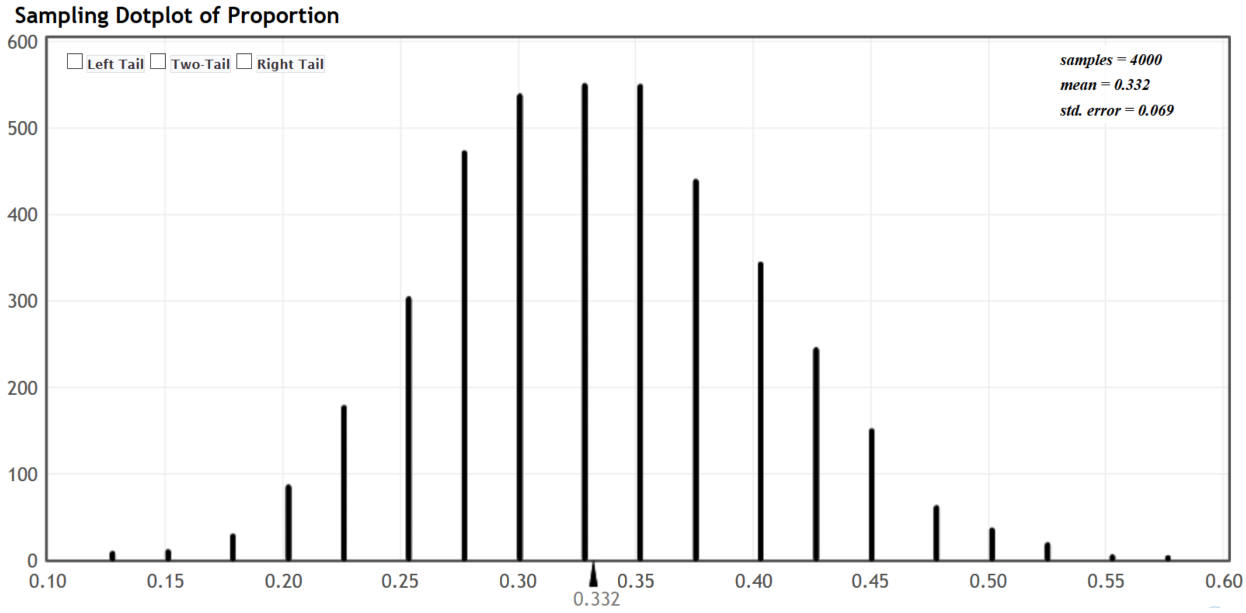
To understand sampling variability for sample percentages we will again chose an example where we have census data and therefore know the population parameter. College of the Canyons (COC) has two campuses in the Santa Clarita Valley, the Valencia campus and the Canyon Country campus. We want to know 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

samples = 4000
mean = 0.332
std. error = 0.069

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.

2. 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.

3. 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.