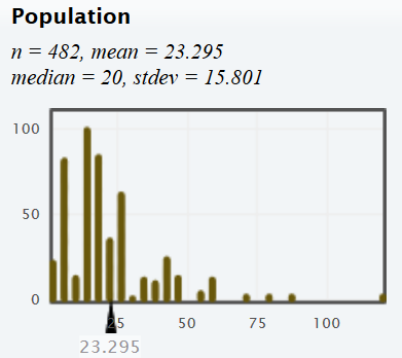
Section 2C – The Central Limit Theorem

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)



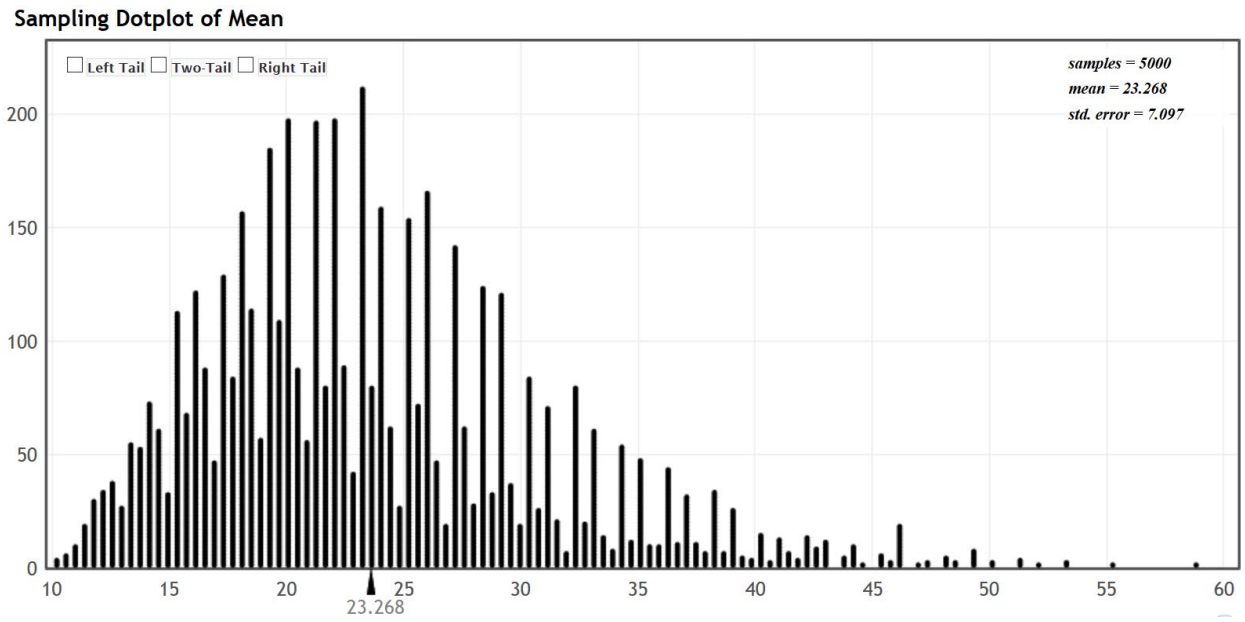
We see that the population is skewed with a population mean average commute time of 23.295 minutes. We will assume that the census is unbiased and that the population mean is really 23.295 minutes.

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.



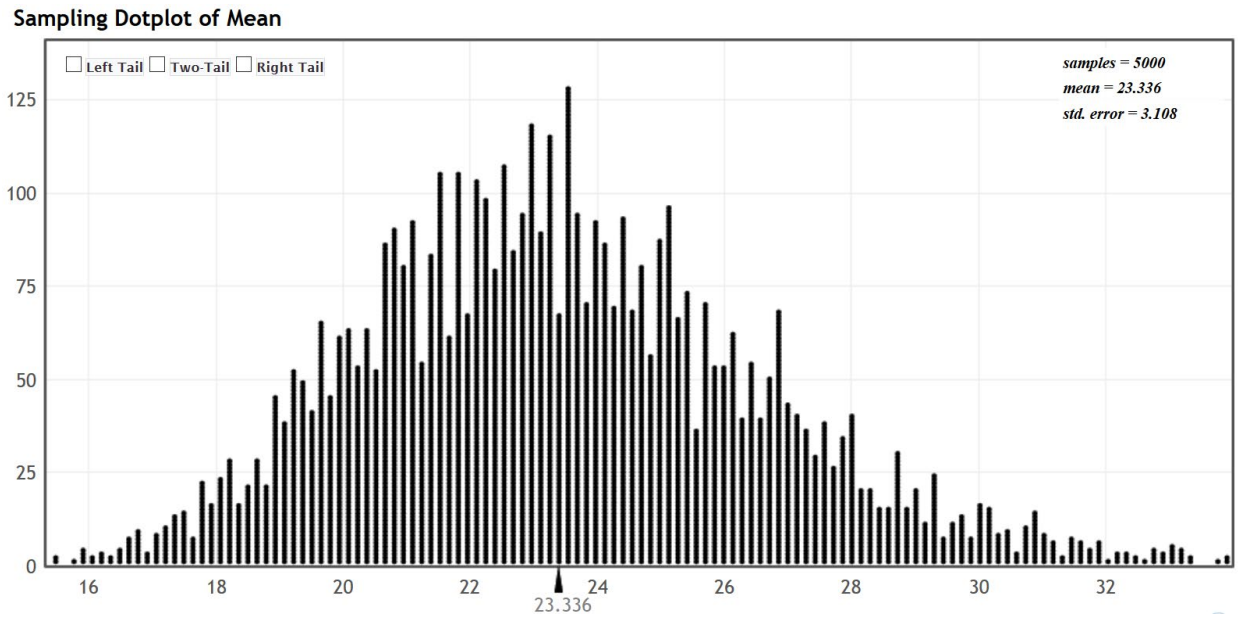
Mean Example 2: Sample Size of Twenty-Five from a Skewed Population

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



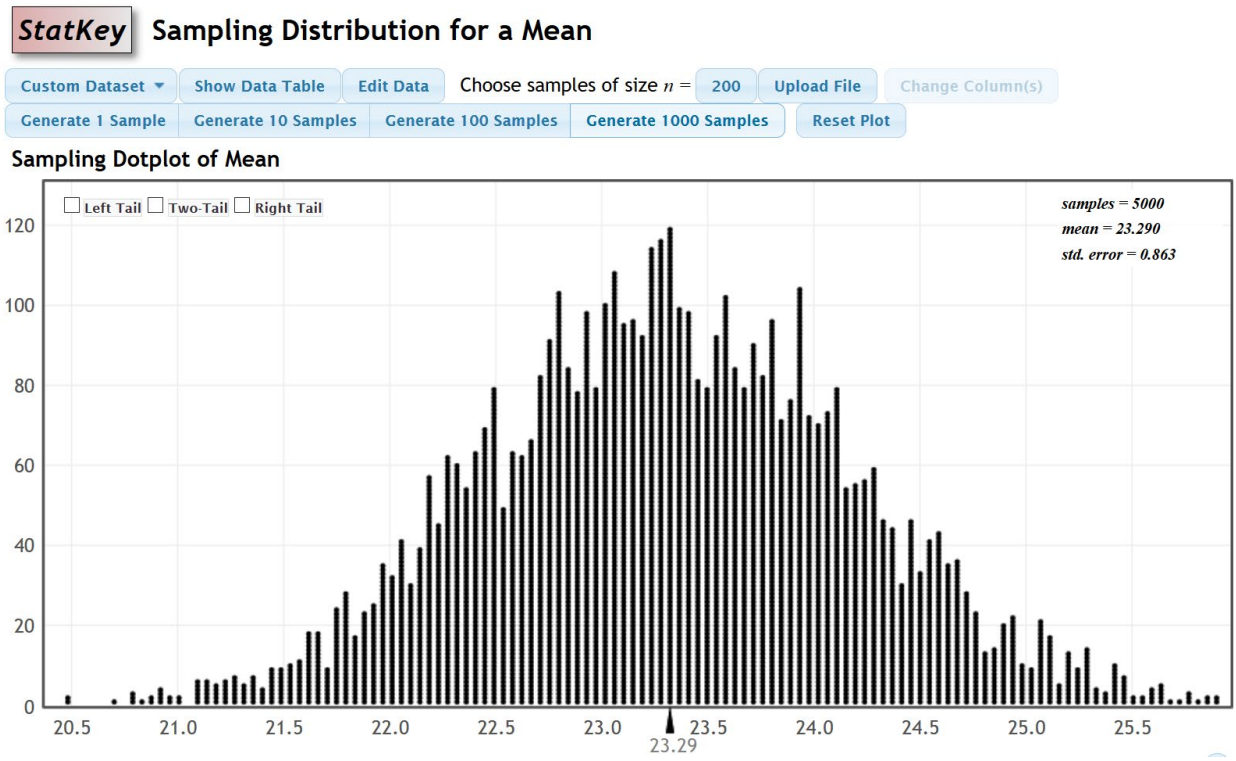
Mean Example 3: Sample Size of Two Hundred from a Skewed Population

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.

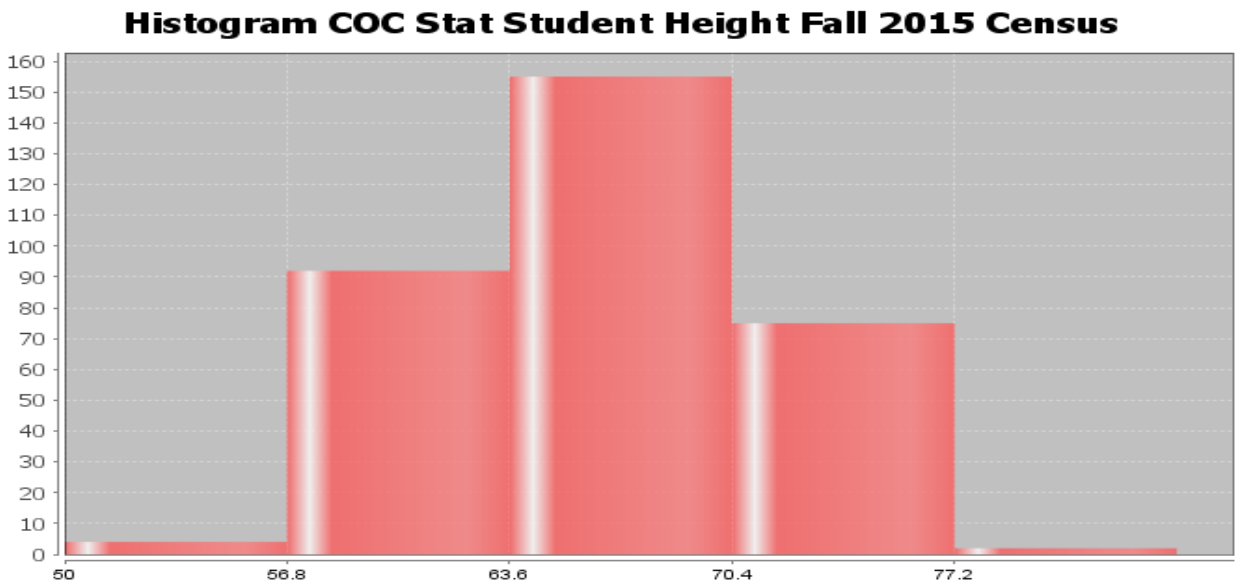
Key Question: What would happen if the population were already normal?

Mean Example 4: Sampling Distribution from a normal population.

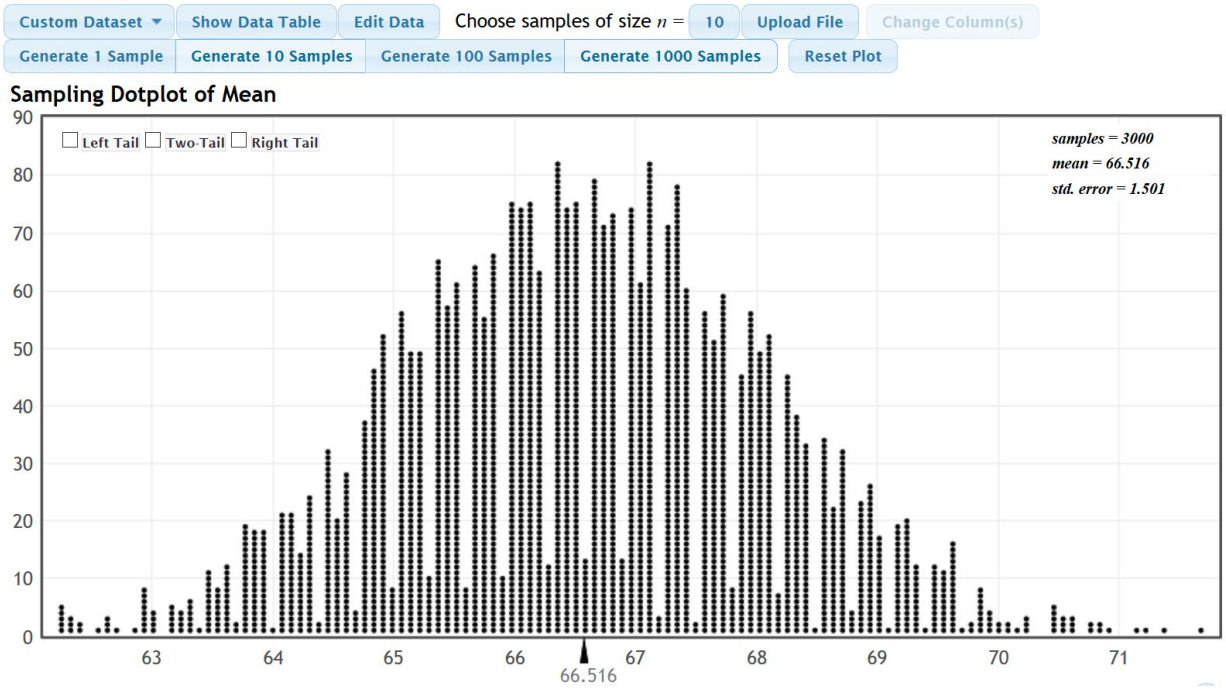
Let us now look at an example of a census with a normal shape. In the fall 2015 semester, we took a census of all of the statistics students at COC and asked them their heights in inches. We will assume this was an unbiased census. This population looked very normal with a population mean average height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

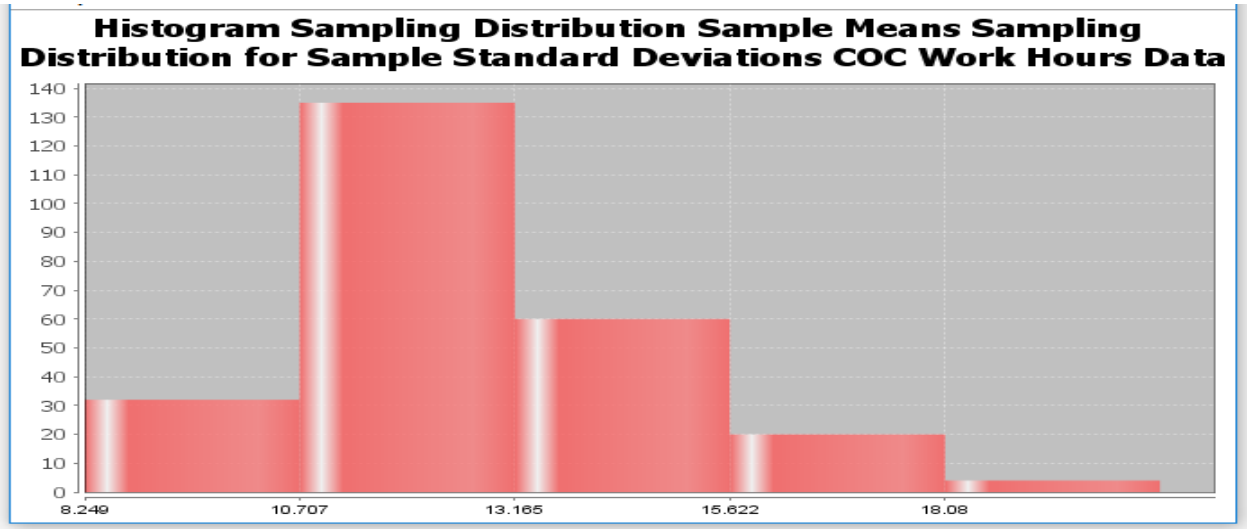
If the population was skewed, what is the shape of sampling distributions for sample standard deviations? Are there any sample size requirements for estimating sample standard deviations? What if the population was already normal?

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.

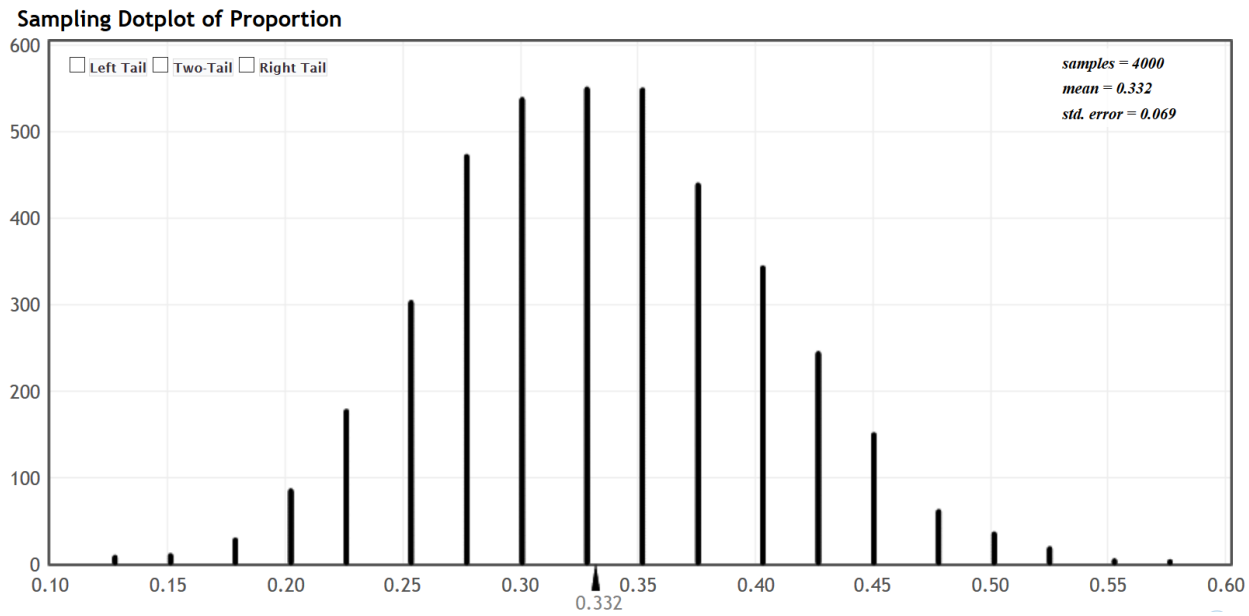
1. A sampling distributions of sample variance is usually skewed right. Later we will see that if the population is normal, the sampling distribution for sample variance will follow a skewed right Chi-Squared distribution. Requirements for traditional techniques for estimating population variance or 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

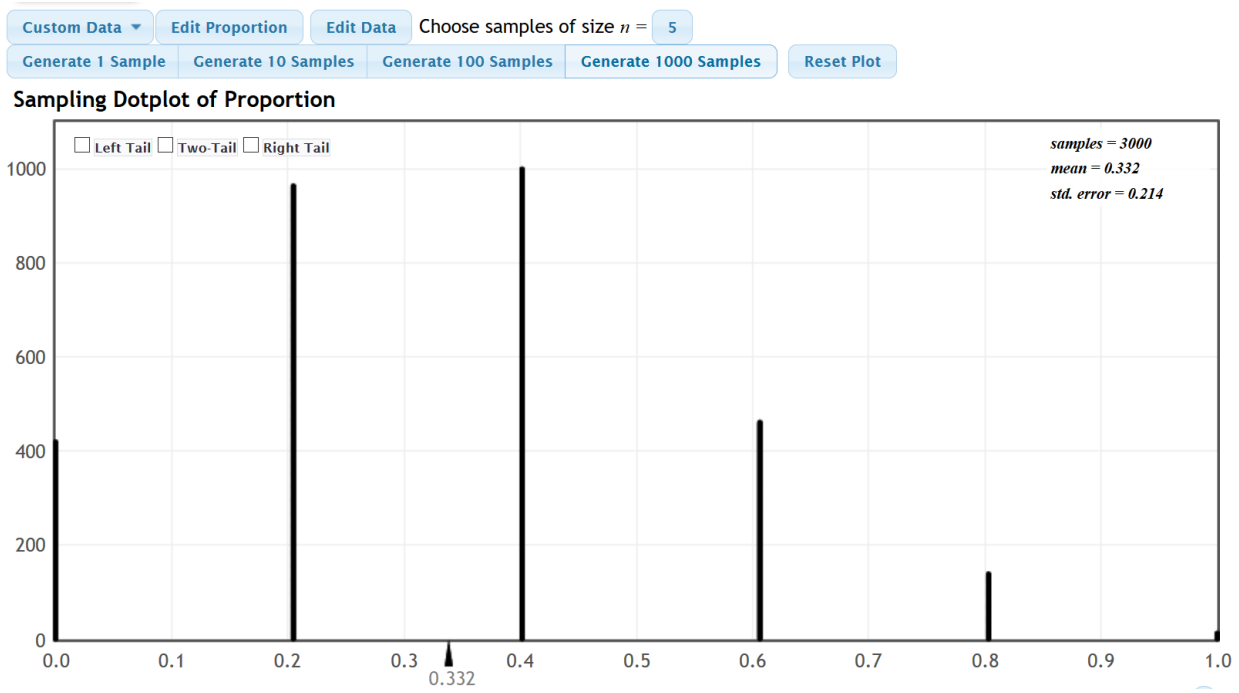
Proportion
0.332



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.
