STA-1380 – Elementary Statistics
Week #5

Hello! Welcome to the additional online **Weekly Resources** for the course of **STA-1380**. Following a traditional calendar semester, these will be some of the topics your professors will go over. If you do not see material your section is going over for the week, please look at the other resources listed for this course. In addition to these resources, there might be **Group Tutoring** for this course, please see our website for more details. These sessions will go over these materials in more detail as well as any questions about the material.

Any additional help or services can be found through the **Baylor Tutoring Website.** Visit to schedule a free 30-minute private tutoring session, drop-in times for your course, the Baylor Tutoring YouTube channel, or any additional tutoring resources.

Contacts: Sid Rich M-Th 9am-8pm (Fall and Spring class days) Office Phone: 254-710-4135

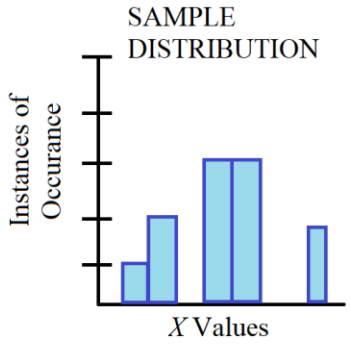
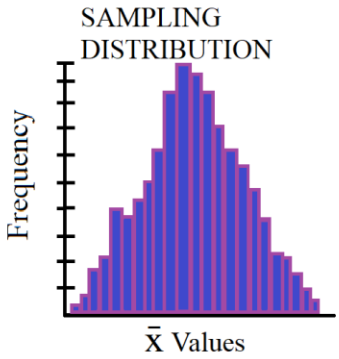
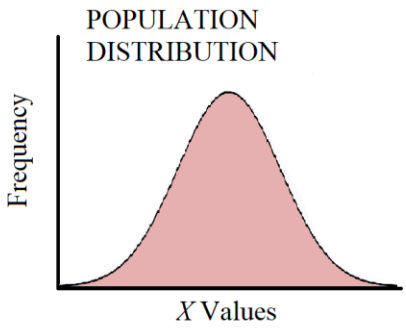
Topic of the Week:
"Sampling Distributions"

Key Points:

- Mean Distribution Parameters
- Central Limit Theorem

In previous chapters, you were taught how to take independent samples from the population and draw inference from them. Now, you will take multiple samples from the same population with the same requirements as before to create a Sampling Distribution.

A Sampling Distribution is a collection of 'Sample Distributions' from a certain population with a set *n* for every trial. These sampling distributions work to mimic the distribution of a population's random variable. It is important to recognize the difference between a sampling and a sample distribution. Within this chapter, there will be many symbols and formulas that you will have to learn. For now, the chart below will demonstrate the comparisons between a Sample, Sampling, and Population Distribution.

	SAMPLE DISTRIBUTION	SAMPLING DISTRIBUTION	POPULATION DISTRIBUTION
			
Mean	$\bar{x}$	$\mu_{\bar{x}}$	μ
Std. Div.	s	$\sigma_{\bar{x}}$	σ
Sample Size	n	'XX' amount of trials with n subjects	N
X-Axis	Potential X Values	$\bar{x}$ values from 'XX' trials	All Potential X Values
Y-Axis	# of times an $X = x$	How often the $\bar{x}$ occurs	How often $X = x$ occurs

Note the similarities between the notations and shapes of the graphs. This is not a coincidence. Every Sampling Distribution is made with the intent compiling the samples in a way that allows the use of distributions such as the Normal or other curves you will see in the future to gain valuable information. There are always going to be parallels between the distributions, so it is important to recognize what each distribution is attempting to display and how it correlates to the other two.

Starting with the means of each distribution, it is clear to see how each of the three are interconnected. The Sample Distribution's mean is denoted as 'X-bar', it is the mean derived from the average of a single sample. For the Sampling Distribution, the mean is denoted as 'Mu of X-bar,' describing that the averages of every 'X-bar' have been taken to form a sort of 'mean of means.' To demonstrate that this value comes from original sample data and not to be mistaken for the regular 'Mu,' the subscript is necessary. But why is the symbol μ used at all? It is because the Sampling Distribution mean is meant to be an unbiased, perfect point estimator, which is a lot of statistical words to say that "$\mu_{\bar{x}} = \mu$" The mean of the Sampling Distribution, if done correctly, should be exactly what the mean of the population is.

The standard deviation of the Sampling Distribution is treated the same way, except for the way it is calculated and determined. The standard deviation of the sample is not used to calculate

the Sampling Distribution, but rather, the idealized Population Distribution's std. dev is manipulated to better depict the spread of the samples. Therefore, it is important to remember that $\sigma_{\bar{x}}$ is based off the population parameter given, rather than the actual data from the sample.

Sample size is another important aspect of every distribution. With the Population Distribution, there is only ever 1 trial done with every individual inside of it. For Sample Distribution, it is a single trial with a set number of individuals randomly selected. This is why both graphs use *N* and *n*, the only difference being the smaller 'n' used to denote that a portion of the population was sampled rather than the whole.

With a Sampling Distribution, there is a set number of individuals sampled and data collected from, the only difference is the trials. Sampling Distributions are at their core, the condensed data of multiple repeated trials. Having a Sampling distribution with $n = 50$ and 400 trials conducted means that 400 samples from the same population took 50 individuals and collected the data. The Sampling Distribution shares the '*n*' with the Sample Distribution, but also has a secondary facet of the number of trials conducted.

The X-axis of the three graphs share a similar characteristic to the sample size. Both the Sample Distribution and the Population Distribution share the same units, the possible values that 'X' can take. The only difference being that the Population Distribution has every possible point, and the Sample Distribution's range being cut short to only the values collected during the trial.

The Sampling Distribution is where this changes. This no longer a matter of *X* values but of 'X-bar', with the Sampling Distribution being the only graph that deals with the sample means. The Y-Axis functions similarly to the means. The Population and Sampling Distribution share the same label of 'Frequency,' or 'Density' (these words can be used synonymously). The Sample Distribution is the only difference because it accounts for exactly how many times $X = x$.

As you work through this chapter, there will be times when you come across questions that seek to isolate and test your knowledge on these relationships. In addition to relationships, you will always want to learn and memorize the equations and formulas that provide each parameter and statistic.

Highlight #1
"Mean Distribution Parameters"

Definition: The set values gained from a sample of a larger population or derived from the population in order to provide inferences.

Notation: $\bar{x}$ = Sample Mean, s = Sample Std. Dev, s^2 = Sample Variance, n = Sample Size
$\mu_{\bar{x}}$ = Sampling Mean, $\sigma_{\bar{x}}$ = Sampling Std. Dev., $\sigma_{\bar{x}}^2$ = Sampling Variance,
μ = Population Mean, σ= Population Std. Dev., σ^2 = Pop. Variance N = Pop. Size

The Sample Distribution series of parameters comes solely from the data collected. This is why their notations are vastly different from everything else covered so far. Recall from your teacher's lectures that you can never adjust or calculate for the standard deviation on its own, you must always first find the variance of any set and manipulate it to fit your data. The standard deviation is the square root of the variance.

$\bar{x}$ is calculated by taking all the sample data and dividing it by the sample size: $\bar{x} = \frac{\Sigma (X = x)}{n}$

s^2 is calculated by taking the difference between the values and the mean, divided by n-1. $s^2 = \frac{\Sigma (x - \bar{x})^2}{(n-1)}$

Because the sample is not a perfect representation of the population, we use a smaller denominator to produce a slightly larger std. dev. for the sample. This larger parameter is used because it accounts for potential outliers and anomalies produced by human error.

The Sampling Distribution as stated previously is the distribution of sample means with a set number of trials. The parameters of this distribution seek to use the sample data to represent the population. As such, the mean can often be calculated or denoted in two ways. The std. dev. of the distribution does not base itself off the sample data, but rather the population parameters.

$\mu_{\bar{x}}$ is found by taking the mean of sample means, or directly equaling Mu. $\mu_{\bar{x}} = \frac{\Sigma (\bar{x})}{\#\text{ trials}}$ *or* $\mu_{\bar{x}} = \mu$

$\sigma_{\bar{x}}^2$ is found by dividing the variance of the population by the sample size. $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$

The standard deviation for the Sampling Distribution has another term, 'Standard Error.' The standard error is used to help find the margin of error, which will be used for confidence intervals in a later chapter.

Highlight #2
"Central Limit Theorem"

Definition: A Sampling Distribution will appear approximately normally distributed with a sufficiently large sample size regardless of the original population's shape and spread.

Rule: For STA-1380, the term 'sufficiently large' will be based on the Sample and Histograms, '30' is only a suggestion, n might need to be more or less depending on the original population.

The CLT is one of the 'checkmarks' required for statistical inference. In order to use the data collected, it must fit the parameters of the distributions it is compared to. For the Sampling Distribution, the parameters are that of the Normal Distribution. It would not make sense to try and compare the expected values of a bell-shaped curve with bi-modal sample data. As such, in order to use the properties of the Sampling Distribution, the Sampling Distribution must be bell shaped.

Due to the nature of collecting data, a larger sample size leads to a more uniform distribution. The term CLT refers to how as *n* approaches infinity, the data of the graph will collect around the center point or mean of the data. Look to see if the Sampling Distribution mirrors this theorem by collecting around the center of the data as *n* increases.

Check Your Learning

1. The HOA of Waco wants to figure out how often the residents in a neighborhood are watering their plants. The HOA has information from a previous city census that found the results on this question were normally distributed with an average watering of 9 hours aggregate a week with a variance of 4 hours. A sample of 40 households were polled and it was found that the mean watering time was 9.5 hours with a std. dev. of 3 hours.
    a. What is the standard deviation of the Sampling Distribution?
    b. Find the probability of finding results as extreme or more extreme than the sample given?

2. A sample is conducted on a population with a non-normal distribution with the mean of 145 and a standard deviation of 27. With a sample of 500 individuals, the mean collected is 147.8. The researchers are upset to find a non-significant result, but one intern stops them and begs them to reconsider.
    a. Without knowing the initial distribution's spread, how do you know that a Sampling Distribution can still be used?
    b. What is the probability of having this result or more extreme ($X > 150$)? Is this significant?

Things Students Struggle With

1. What Standard Deviation do I use? ($\sigma_{\bar{x}}^2$ vs s^2):

   a. One way to remember that the Sampling Distribution uses Sigma over 'S' is that the distribution is supposed to mimic the population distribution. Therefore, if a question ever asks you what the std. dev. is supposed to be, or what the mean should be, always relate it back to the population parameters.

2. Why do the samples and sampling distribution matter?:

   a. With the population known, it seems counter intuitive to use another distribution to find results. In reality, the population will almost never be known. The sample data will be used later down the line in the events that a population parameter is not known. This is why it is important to practice identifying and becoming familiar with it now. The Sampling Distribution works to fit the data into a form that statisticians can use known relationships to interpret data. This essentially means that since they know how a normal distribution acts and functions, it is worthwhile to manipulate data into this type of form by converting multiple samples into a Sampling Distribution.

Concluding Comments

That's it for this week! Please reach out if you have any questions and don't forget to visit the Tutoring Center website for further information at https://www.baylor.edu/tutoring

---

Answers to CYL

1. **a.** .316 or $(.1)^{½}$
   b. 1.58 std. dev or 5.7053%

2. **a.** 500 is sufficiently large for many distributions, it can be assumed this will normalize the data.
   b. 2.32 std. dev or 1.017%