Week 6
Elementary Statistics: STA 1380

The concepts this resource covers are the topics typically covered during the first few weeks of the semester. If you do not see the topics your particular section of class is learning this week, please take a look at the weekly resources listed on our website for additional topics throughout the semester.

We also invite you to look at the group tutoring chart on our website to see if this course has a group tutoring session offered this semester.

If you have any questions about these study guides, group tutoring sessions, private thirty minute tutoring appointments, the Baylor Tutoring YouTube channel or any tutoring services we offer, please visit our website www.baylor.edu/tutoring or call our front desk during open business hours (M-Th 9 AM-8PM on class days) at 254-710-4135.

KEYWORD SECTION GOES HERE

==

Topics of the Week:

-

Short Review - Statistics

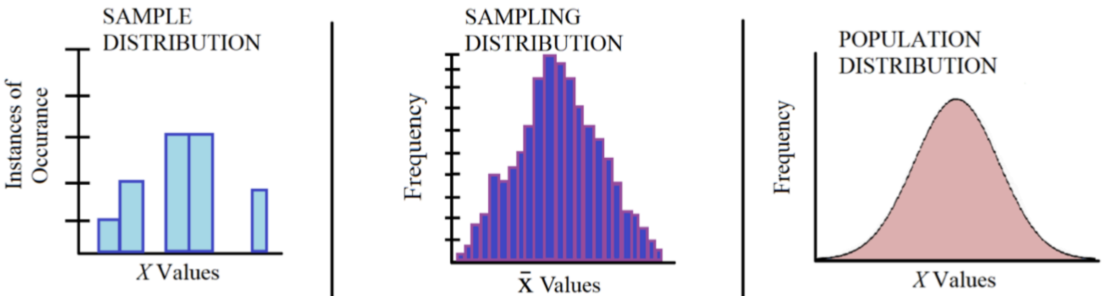
No, not the name of the class. Statistics are ways to measure and describe sample data. Think about what the term means, what does that denote? It denotes the "center" of the data, same with terms like median and mode. Before we move into Sampling Distributions think about how the 'moment' or statistics of a dataset are described. What happens when we have a series of the same statistic from data sets that are from the same population and similar? We get a series of statistics, like means that we can compare in themselves. Review your notation before continuing on. It would be to be comfortable with the following statements before approaching Sampling Distributions:

- What is a statistic? What makes it different from a parameter?
- What is a population? What are subsets of a population?
- What symbols represent sample mean vs population mean, sample std dev vs population std dev, etc?

Concept #1 : Sampling Distributions

If we think about the sample statistics of a sample data set we can think of the mean, the variance, and standard deviation these are observed for one set of data. Let's call the first mean of

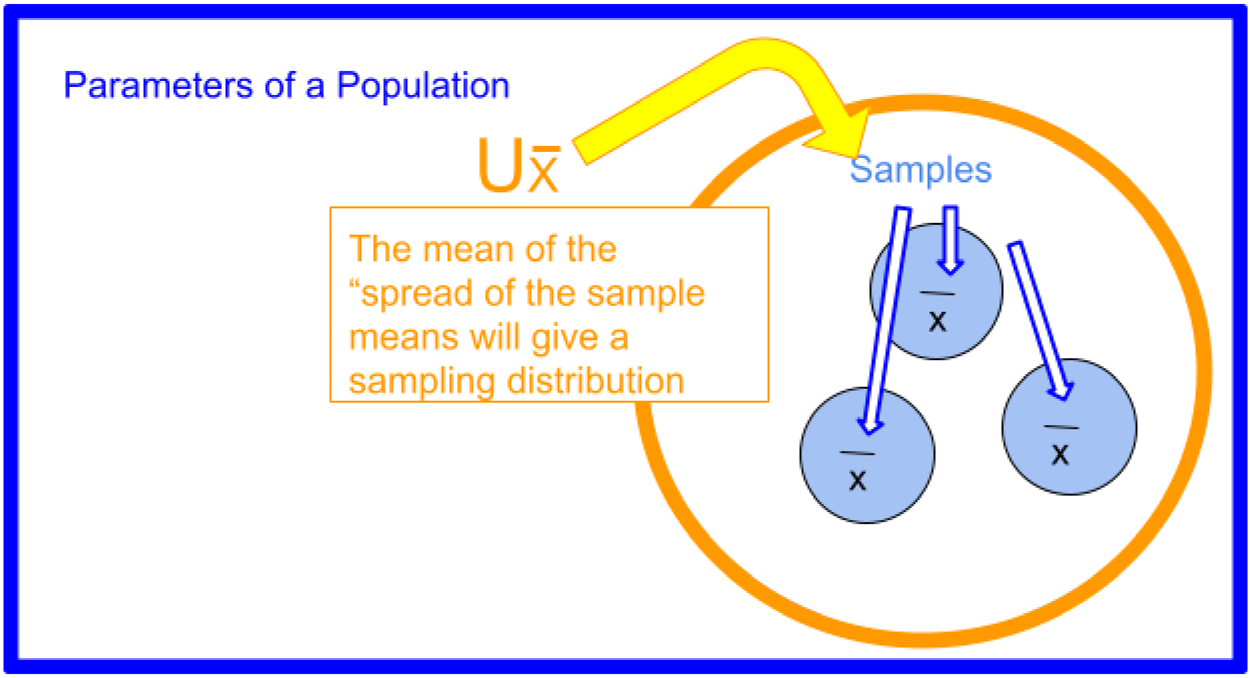
the first data set “A”. If we took 25 more sets of data from the same population and found 20 more means (“B” through “Z”), then we would have a set of means all from this larger population. But what if we took the average of the averages? We would get a new ‘higher’ average that involves more data. Imagine you did this experiment 100’s of times. The graph that denotes where the average of the averages fall is called the **Sampling Distribution.**
See the image below for an example of what this looks like:



Mean	$\bar{x}$	$\mu_{\bar{x}}$	μ
Std. Div.	s	$\sigma_{\bar{x}}$	σ

The first graph in blue is the graph of **one sample set of data** using x-bar as its average and s as its standard deviation. The next graph in purple is the Sampling Distribution or **the collection of means from multiple sets of sample data**. The final graph in pink is the **population graph** which shows the entire population. Sampling Distributions are useful estimates of population Distributions due to **how as the number of sample means grows the average of these sample averages will resemble the population’s true average.** 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 σ is manipulated to better depict the spread of the samples. Therefore, it is important to remember that $\sigma\bar{x}$ is based on the population parameter given, rather than the actual data from the sample.

Sample size is another important aspect of every distribution regardless of what type it is. 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.



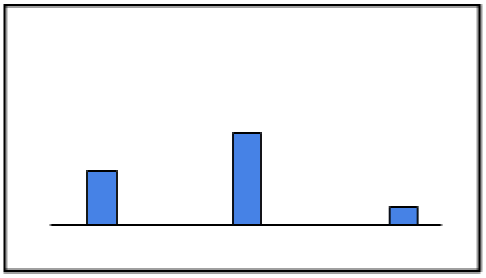
The Sample Distribution series of parameters comes solely from the data collected. This is why their notations are vastly different from everything else covered so far. Recall that you can never adjust or calculate for the standard deviation on its own, you must always first find the variance of any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

<u>**Concept Check:**</u> See the notation below each graph. Why does the notation change? An easy way to remember this is that **x-bar** is a sample mean, and **μ** is the population mean, but the sampling mean or **μx rewarding bar** is the intermediary between a single sample and its entire population.

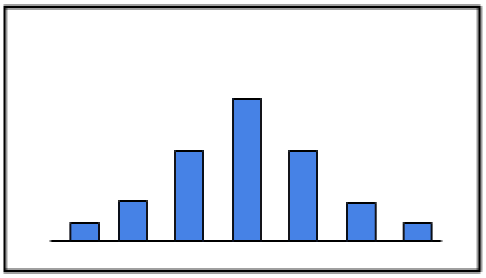
<u>**Concept #2 : Central Limit Theorem**</u>

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

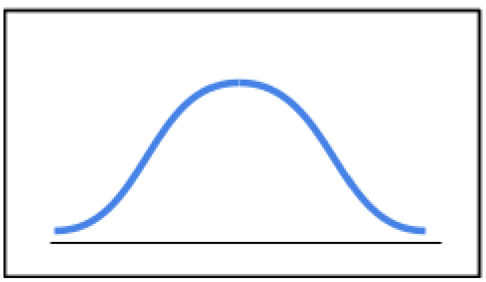
Rule: For STA-1380, the term 'sufficiently large' will be based on the Sample and Histograms, **'30' is only a suggestion**, n might need to be more or less depending on the original population. The CLT is one of the 'check marks' required for statistical inference. In order to use the data collected, 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 *(two bump)* 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. Basically as the number of samples increases we get closer and closer to a true parameter.



A few data points will seem to be random, but we can add more to expand into a pattern.



More data points will create a central movement, or generalized area where most points should fall about the middle



Once you have many points, due to how these points will all fall arounds a centralized point or mean, it will create a normal distribution

<u>Example Questions:</u>

1. 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 **sample mean 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

Can Sampling Distribution still be used?

b. What is the probability of having this result or more extreme ($X > 150$)? Is this significant?

c. What notation will all of these sample values given be associated with?

2. 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 **sample mean watering time was 9.5** hours with a **sample 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?