Objectives:

- To continue to practice the distinction between ***parameters*** and ***statistic***
- To gain an understanding of the fundamental concept of ***sampling variability***
- To discover and understand the concept of ***sampling distributions*** as representing the long-term pattern of variation of a statistic under repeated sampling.
- To explore the sampling distribution of a sample proportion through actual experiments and calculator simulations.
- To discover the effect of sample size on the sampling distribution of a sample proportion
- To investigate the principle of ***statistical confidence*** as it relates to estimating a population parameter based on a sample statistic

parameter vs. statistic (Very important so I re-address): The purpose of sampling is to obtain a statistic that can predict the parameter of the population. ***Make sure you understand this!*** A parameter is rarely ever known. This is almost impossible in most cases. Therefore we need to carefully obtain a sample that we can reasonably assume represents the population, thus a the statistic will HOPEFULLY be close to the parameter.

An ***election poll*** is a ***statistic*** that we hope to be able to predict the outcome of an election which by necessity is a ***parameter***

"The primary goal of sampling is to estimate the value of the parameter based on the statistic"

Symbols for sampling and populations:

	parameter	Statistic
Proportion	π (pi)	$\hat{p}$ (p - hat)
Mean :	μ (mu)	$\bar{x}$ (x - bar)
Standard Deviation	σ (sigma)	s

Sampling distribution and Sampling Variability: The distribution of proportions of multiple samples. If you take many samples and find the sample proportion of each when you plot these you will find that the mean of these samples will closely match that of the parameter of what the proportion actually is. The more samples you obtain the closer the sample distributions average will be to that of the proportion.

Sampling Variability *is the idea that each sample is different, and you should not expect to obtain the same result every time you take a sample.*

Sampling distribution *is the pattern obtained when graphing the sample statistics due to the predictable manner in which samples vary.*

Remember we talked earlier about ***confidence*** in a statistic. The more samples we take the more ***confident*** we can be that the statistic is close to the parameter. We actually think of it being within a certain distance from the parameter. We can be more confident or we can shrink this distance as we take more samples.

If we take one sample of 25 *Reeses* Pieces, we are not very confident that we are close to what the actual is, but if we take numerous samples we become more confident, also if we enlarge our sample we become more confident.

If we take a bunch of samples and keep finding their proportions we know by natural law they will be *normally distributed* about that actual proportion. In fact we ***know*** that 95% of all samples taken will fall within two standard deviations of the actual proportion. THIS VERY IMPORTANT!

Problem: how do we know what the standard deviation of all samples is? We need to know this to find out how confident we are and/or what the distance from the actual may be. It turns out that when we take repeated samples of size ***n*** we may use the following formula for the standard deviation:

If π is the proportion we are aiming for, and n is the sample size
Then the sample proportion

$$\hat{p} = \pi$$

and the standard deviation

$$s = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Central Limit Theorem: The concept illustrated above is a property of the **Central Limit Theorem (CLT)** for a sample proportion. If the population proportion is known to be π, if numerous samples are taken and for each sample a sample proportion is calculated ($\hat{p}$) then the distribution of $\hat{p}$'s will be approximately normal with a standard deviation of $\sqrt{\frac{\pi(1-\pi)}{n}}$ and the center the distribution will be at π. This will be true as long as $n\pi \geq 10$ and $(1-n)\pi \geq 10$ and the samples are random.

Whenever invoking **CLT** be sure that:

(1) Samples are ***RANDOM*** from the population being studied (simply random is not enough, you need to make sure that all members of the population are equally likely to be chosen for the sample – If studying adult males, the sample cannot come from the male teachers at the High School – too many males are eliminated from the sampling frame)
(2) ***n*** (sample size) must be large enough. If *n* is not large enough then the sampling distribution will not follow the normal curve and therefore the normal calculations will not be accurate.

Some things to think about:

To increase your confidence or to tighten the distance of possibilities (thus keeping the same confidence but making it more accurate) you can increase sample size.

Because of the facts discussed - we only need to take one sample and we can make our predictions from this one sample. Election polls are a perfect example of this. We take one poll , we know what the standard deviation will be by knowing the approximate proportion parameter, and we can get a good guess as to what the rest of the population will be like.

Remember we MAY BE WRONG, but if we are that is because we obtained a very rare sample (assuming our sampling technique was reasonable.)

Statistical Significance: Since the distribution of the sample statistics is predictable we can look at how a sample fits with this expected distribution, if there is a "*large*" difference between what is expected and what is obtained it is said to be statistically significant. It is important to understand that "*large*" here is relative as indicated by the Z-score obtained from the sample.