MATH 140 **Name** ________________________ **Date:**

R Activity - CLT for Proportions

Instructions

In this activity we use R to investigate the Central Limit Theorem for proportions. Your code in R will practice the following tasks:

- Create a population in which a given proportion of the elements have a particular feature.
- Simulate the drawing of random samples from this population, and determining sample proportions, $\hat{p}$.
- Approximate the sampling distribution for $\hat{p}$
- Compare your approximate sampling distribution with the results of the Central Limit Theorem
- Use the CLT to estimate some probabilities.

Based on your code you will record answers to various questions on this worksheet. The completed worksheet is due in class on Tuesday, March 31.

1. **Create a Population**

   (a) Think of a scenario (real or fantastic) in which you want to investigate the proportion of a population having a certain feature. Maybe it's the proportion of all skittles that are green, or the proportion of all dragons having 6-toed feet, or . . . . Record your population below, and the feature you are interested in studying.

   Population:

   Feature of Interest:

   (b) Decide on a (reasonable or fantastic) value for p, the proportion of your population having your feature of interest. Be sure your value is between 0 and 1, preferably between .2 and .8. Record this value below

   $p =$

2. **Draw a sample**

   In your script, you created your `population`, and were asked to draw a sample of size $n = 100$ from it, storing the results in the vector `draw`. You then calculated $\hat{p}$ for this sample. Record that sample proportion below. Is it close to the population proportion p? Exactly the same?

   $\hat{p} =$

3. **The sampling distribution for $\hat{p}$.** In the script, you approximated the sampling distribution for $\hat{p}$ by generating a vector called `results_100` that stores the values for 100,000 different sample proportions, from 100,000 independent samples from your population. You then made a visual of the distribution of these $\hat{p}$ values. Describe the shape of the distribution. Is it bell-shaped? Based on this distribution, would you describe the sample proportion you obtained in Q2 as typical, fairly common but on the high side of typical, fairly common but on the low side of typical, or a very rare occurrence?

4. **The Central Limit Theorem.** According to the central limit theorem, the sampling distribution for $\hat{p}$ is nearly normal, so it should be bell-shaped.

   (a) In the case of your population, and your sample size of $n = 100$, according to the Central Limit Theorem, what is the mean and what is the standard deviation (also called standard error in this context) of this sampling distribution?

   mean of the theoretical sampling distribution ($n = 100$) =

   standard error =

   (b) In Q3, we constructed `results_100` to approximate the sampling distribution for $\hat{p}$. Determine the mean and standard deviation of the `results_100` vector.

   `mean(results_100)` =

   `sd(results_100)` =

   How do they compare to your answers to part (a) of this question?

5. **Change the sample size.** Now we essentially repeat questions 3 and 4 except with samples of size $n = 400$. You ran code to generate the vector `results_400`, containing 100000 different sample proportions from drawing 100000 different samples of size $n = 400$ from the population.

   (a) Consulting the barplot of the `results_400` vector, does this visual resemble a bell curve? How does this visual compare to the visual of the `results_100` vector in terms of both center and spread?

   (b) According to the CLT, what should be the mean and standard error of the sampling distribution for $\hat{p}$, when $n = 400$?

   mean of the theoretical sampling distribution ($n = 400$) =

   standard error =

(c) Determine the mean and standard deviation of the `results_400` vector. How do they compare to your answers to part (b) of this question?

6. **Estimating Likelihoods** (Show your work. Calculate z-scores and use pnorm() to find probabilities.)

   (a) Use the CLT to estimate the probability that more than half the voters in a sample of size 100 likely voters are in favor of Candidate A, assuming that 46% of the population of likely voters actually favors Candidate A.

   (b) Use the CLT to estimate the probability that more than half the voters in a sample of size 500 likely voters are in favor of Candidate A, assuming that 46% of the population of likely voters actually favors Candidate A.