Data 310: Stat. Methods for PSUniversity of Pennsylvania
Spring 2021 Professor Marc Meredith

Homework Week #3

1. Suppose you are finding the mean of random variables that have an expected value of 76 and variance of a variance of 35. Applying the Central Limit Theorem, use R to answer the following questions

   - (a) What is the probability that the average is between 73 and 78 in a random sample of 50 observations?
   - (b) Find a symmetric 99 percent confidence interval on the mean in a random sample of 75 observations?
   - (c) How large must my sample be in order for there to be a 95 percent probability that my sample average is between 75.5 and 76.5? (Note, you will likely have to round the answer you find to the nearest whole number. We can't sample a fraction of person. YET.)

2. One thing that we do when we make election projections is collect data from a sample of precincts and project something about what vote outcomes are likely to be in the entire state based on what we observe in the sample. We are going to do a simplified example of this using data from Ohio's 2016 presidential election. You will find on Canvas a dataset entitled "Ohio2016.xlsx." This dataset contains the number of ballots cast for Trump, Clinton, and all other candidates in this election by precinct).

   - (a) Load this dataset into R (I used the read_excel() function from the readxl library, another option is to use the import() function from the 'rio' package). Construct a variable that is equal to Trump's vote in a precinct divided by the total number of votes in the precinct. Plot a histogram of this variable. Would the pdf of this variable be well approximated by a normal distribution?
   - (b)Given that this is all of the precints in Ohio, we can treat this like a 'Census'. Calculate the population mean and variance for Trump's vote share in a precinct.
   - (c) Suppose we sampled Trump's vote share from 40 randomly selected precincts on Election Night and averaged them together. Combine your knowledge of the population mean and variance and the Central Limit Theorem to make a prediction about what the 99 percent confidence interval would be on this statistic.
   - (d) Calculate the bounds of this 99 percent confidence interval if 80 or 120 precincts were sampled instead.
   - (e) To confirm that the bounds calculated by the Central Limit Theorem are correct, run the following simulation
     - Draw a random sample of 120 precincts from the data (I use the 'sample()' function to do this).
     - Calculate the average Trump vote share in these precincts.

- Store this average in a vector
- Repeat those three steps 10,000 times using a for loop
- After running the 10,000 simulations, use the quantile() function to find 0.5 and 99.5 percentiles of the average Trump vote share over the 10,000 simulations. Compares these values to the bounds on the symmetric 99 percent confidence interval you estimated using the Central Limit Theorem.