Stat 123 Spring 2022 Homework Assignment 3
Due date will be posted on Brightspace

Using R Markdown, please complete the following assignment. Your assignment should be submitted as a pdf (whether you knit directly to PDF, or knit to HTML or Word and then convert the file to a pdf).

1. Download and save the *homework3Data.csv* dataset and read it into R. This data set contains 6 numerical variables A, B, C, X, Y, Z.

   (a) If you use the function hist() to plot a histogram followed by the function abline(v=3,col='red'), this will add a red vertical line at $x = 3$.

   Using these functions:
   - plot a histogram for each of the variables.
   - add vertical lines for the sample mean and the sample median of those variables. Make the sample mean lines red and the sample median lines blue.
   - add a green density curve to each plot.
   - make sure your histogram has a main title.

   (b) One of the variables is normally distributed. Determine which variable it is and justify why you think it is that variable.

   (c) For the normally distributed variable you identified in part (b), use the $68 - 95 - 99.7$ rule to determine the intervals such that approximately 68% of the data, 95% of the data, and 99.7% of the data lie within those intervals.

   (d) Use the quantile() function to approximate those same intervals. Are the intervals the same?

   (e) Use the qnorm() function (with the sample mean and sample standard deviation) to approximate those same intervals. Are these intervals the same as the intervals in either part (c) or part (d)?

   (f) Suppose you wish to estimate the population mean for the normally distributed variable you identified in part (b). Compute the following:

   - an estimate of the population mean.
   - the estimated standard error of the statistic.
   - the critical value for an 88% confidence interval.
   - a 88% confidence interval for the population mean.

2. For this question, you will need to install the package 'dplyr' into R by typing in the command install.packages('dplyr'). Then you need to load dplyr into R by typing in the command library(dplyr). We will be using the *starwars* data set that is built into the dplyr package.

   (a) Create a vector called *names* which contains the names of starwars characters that are included in the data set.

   (b) The function nchar() determines the number of characters in a string. How many characters are in the 5th, 20th, and 34th elements of the *names* vector?

   (c) Create an empty numeric vector called *num_char*. Write a loop which calculates the number of characters in each element of the *names* vector, and puts the corresponding number in the *num_char* vector.

   (d) Now do the same thing that you did in part (c) using the lapply() or sapply() function in R. Be careful that your output is a vector.

3. Consider again the *homework3Data.csv* dataset and the variable X.

   (a) Write a bootstrap computing the median on $10,000$ samples (with replacement) of size 600 of the variable X.

   (b) Plot the resulting sampling distribution for the median of X.

   (c) Determine an estimate for the median of X.

   (d) Compute a 95% confidence interval for the median of X.