Assignment #2

MSAN 593

***DUE**: Friday, July 29, 23:45*

Be sure to upload **both** an `*.Rmd` file as well as the compiled `pdf` to Canvas by the due date and time. Late submissions will receive a grade of zero.

Your `*.Rmd` file will be run on local machines by graders. If you file does not run, you will automatically lose 30% of the grade. If you resubmit your corrected homework by the last class of the module, it will be graded out of the remaining 70%. Failure to resubmit will result in a grade of zero.

Always use `echo = TRUE` so that I can see all your code, and include relevant results. You can also assume that the data file(s) being read into your `*.Rmd` file are in the current local directory, e.g., `read.csv('myFile.csv')` will work. Do **not** hard code a specific directory structure.

Question 1.1

1.1.1. Create 10,000,000 random variates $\sim \mathcal{U}\{4, 6\}$ and store the result in a vector called `myRunIfVec`. Create a histogram.

1.1.2. Sample randomly 100,000 times from `myRunIfVec` and plot the sample histogram. Describe the shape of the sampling distribution and note if it is different from the population distribution.

1.1.3. Sample two random elements of `myRunIfVec`, take the mean of those two elements, and store the value in `unifSampleMean_2`. Repeat this step 100,000 times, so that you will have sample 200,000 elements from `myRunIfVec` and created 100,000 2-sample means in `unifSampleMean_2`. Plot a histogram of `unifSampleMean_2`, describe the shape of the sampling distribution of the mean, and note if it is different from the population distribution.

1.1.4. Repeat (1.1.3), but this time sample five random elements, take the mean, and store the value in `unifSampleMean_5`. Repeat this step 100,000 times. Plot a histogram of `unifSampleMean_5`, describe the shape of the sampling distribution of the mean, and note if it is different from the population distribution.

1.1.5. Repeat (1.1.4), but this time sample ten random elements, take the mean, and store the value in `unifSampleMean_10`. Repeat this step 100,000 times. Plot a histogram of `unifSampleMean_10`, describe the shape of the sampling distribution of the mean, and note if it is different from the population distribution.

1.1.5. Repeat (1.1.4), but this time sample thirty random elements, take the mean, and store the value in `unifSampleMean_30`. Repeat this step 100,000 times. Plot a histogram of `unifSampleMean_30`, describe the shape of the sampling distribution of the mean, and note if it is different from the population distribution.

Question 1.2

Repeat **all** steps of Question #1, but this time initializing the process with a sample of 10,000,000 random variates from a negative exponential distribution with $\lambda = 0.5$.

Question 1.3

1.3.1 Create a **single** vector with 5,000,000 random variates from a $\sim \mathcal{N}\{-3, 1\}$, 5,000,000 random variates from a $\sim \mathcal{N}\{3, 1\}$ and store these values in the vector `myBdist`. Create a histogram and describe the distribution.

1.3.2 Sample five random elements of `myBdist`, take the mean of those five elements, and store the value in `myBdist_5`. Repeat this step 100,000 times, so that you will have sample 200,000 elements from `myBdist` and created 100,000 5-sample means in `myBdist_5`. Plot a histogram of `myBdist_5`, describe the shape of the sampling distribution of the mean, and note if it is different from the population distribution.

1.3.3 Repeat 1.3.2 with sample means of 10, 20 and thirty, creating histograms of each as you go along.

1.3.4 Write a short summary of what you have observed, and relate it to the theory you have learned in MSAN 504. What is this behavior called?

Question 2

Import `hw2.csv`. Your job is is of a janitorial nature. This is a data set is loosely based on real data, and needs to be validated. I am not telling you what these fields mean (this happens more often that you would think), although some are self-explanatory. I am only interested in those observations that were created on or after September 01, 2015. Be vigilant in your work. There is a lot of nonsense in this messy data. Clean it to the best of your abilities and generate a short report of a few pages (graphs and tables included) discussing your findings. Do not be fooled, making sense of messy and unknown data is a very difficult and time consuming task. There is not right or wrong answer. Grades for this question will be assigned on a competitive basis, i.e., the students who offers the best insight and report sets the bar for a grade of A, and everyone else will get an inferior (but scaled) grade.