CS544 Module 5 Assignment

Part1) Central Limit Theorem (20 points)

The input data consists of the sequence from 1 to 20 (1:20). Show the following three plots in a single row.

a) Show the histogram of the densities of this distribution.
b) Using all samples of this data of size 2, show the histogram of the densities of the sample means.
c) Using all samples of this data of size 5, show the histogram of the densities of the sample means.
d) Compare of means and standard deviations of the above three distributions.

Part2) Central Limit Theorem (20 points)

The data in the file queries.csv contains the number of queries Google has had each day for a one year period (365 days). The data file is also available at http://kalathur.com/cs544/data/queries.csv. Use this link to read the data using read.csv function when submitting the homework.

a) Show the histogram of the distribution of the number of queries. Compute the mean and standard deviation of the number of queries Google has had per day.
b) Draw 1000 samples of this data of size 5, show the histogram of the densities of the sample means. Compute the mean of the sample means and the standard deviation of the sample means.
c) Draw 1000 samples of this data of size 20, show the histogram of the densities of the sample means. Compute the mean of the sample means and the standard deviation of the sample means.
d) Compare of means and standard deviations of the above three distributions.

Part3) Central Limit Theorem – Negative Binomial distribution (20 points)

Suppose the input data follows the negative binomial distribution with the parameters size = 5 and prob = 0.5.

a) Generate 1000 random numbers from this distribution. Show the barplot with the proportions of the distinct values of this distribution.
b) With samples sizes of 10, 20, 30, and 40, generate the data for 5000 samples using the same distribution. Show the histograms of the densities of the sample means. Use a 2 x 2 layout.
c) Compare of means and standard deviations of the data from a) with the four sequences generated in b).

Part4) Sampling (40 points)

Use the MU284 dataset from the *sampling* package. Use a sample size of 20 for each of the following.

a) Show the sample drawn using simple random sampling without replacement. Show the frequencies for each region (REG). Show the percentages of these with respect to the entire dataset.

b) Show the sample drawn using systematic sampling. Show the frequencies for each region (REG). Show the percentages of these with respect to the entire dataset.
c) Calculate the inclusion probabilities using the S82 variable. Using these values, show the sample drawn using systematic sampling. Show the frequencies for each region (REG). Show the percentages of these with respect to the entire dataset.
d) Order the data using the REG variable. Draw a stratified sample using proportional sizes based on the REG variable. Show the frequencies for each region (REG). Show the percentages of these with respect to the entire dataset.
e) Compare the means of RMT85 variable for these four samples with the entire data.

Submission:

Create a folder, CS544_HW5_lastName and place the following file in this folder.

Provide the R code, **HW5_lastName.R**, with each portion of the code clearly identified by the corresponding question. Prepare a corresponding word document by pasting the output for each question (**HW5_lastName.docx)**

Archive the folder (CS544_HW5_lastName.zip). Upload the zip file to the Assignments section of Blackboard.