CS544 Module 5 Assignment

Part1) Central Limit Theorem (30 points)

Initialize the city of Boston earnings dataset as shown below:

```
boston <- read.csv(
  "https://people.bu.edu/kalathur/datasets/bostonCityEarnings.csv",
  colClasses = c("character", "character", "character", "integer", "character"))
```

The data in the file contains the total earnings of the employees of city of Boston.

a) Show the histogram of the employee earnings. Use breaks from 0 to 400000 in steps of 50000 and show the corresponding tick labels on the x-axis. Compute the mean and standard deviation of this data. What do you infer from the shape of the histogram?
b) Draw 1000 samples of this data of size 10, show the histogram of the sample means. Compute the mean of the sample means and the standard deviation of the sample means. Use sample() function with replace as FALSE for drawing the samples. Set the start seed for random numbers as the last 4 digits of your BU id.
c) Draw 1000 samples of this data of size 40, show the histogram of the sample means. Compute the mean of the sample means and the standard deviation of the sample means. Use sample() function with replace as FALSE for drawing the samples. Set the start seed for random numbers as the last 4 digits of your BU id.
d) Compare of means and standard deviations of the above three distributions.

Part2) Central Limit Theorem – Negative Binomial distribution (30 points)

Suppose the input data follows the negative binomial distribution with the parameters size = 3 and prob = 0.5. Set the start seed for random numbers as the last 4 digits of your BU id.

a) Generate 1000 random values 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, draw 5000 samples from the data generated in a). Use sample() function with replace as FALSE. 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).

Part3) Sampling (40 points)

Create a subset of the dataset from Part1 with only the top 5 departments based on the number of employees working in that department. The top 5 departments should be computed using R code. Then, use %in% operator to create the required subset.

Use a sample size of 50 for each of the following.
Set the start seed for random numbers as the last 4 digits of your BU id.

a) Show the sample drawn using simple random sampling with replacement. Show the frequencies for the selected departments. Show the percentages of these with respect to sample size.

b) Calculate the inclusion probabilities using the *Earnings* variable. Using these values, show the sample drawn using systematic sampling with unequal probabilities. Show the frequencies for the selected departments. Show the percentages of these with respect to sample size.

c) Order the data using the *Department* variable. Draw a stratified sample using proportional sizes based on the *Department* variable. Show the frequencies for the selected departments. Show the percentages of these with respect to sample size.

d) Compare the means of *Earnings* variable for these four samples against the mean for the data.

Submission:

When the term *lastName* is referenced, please replace it with your last name.

Provide all R code in a single file, **CS544_HW5_LastName.R**. Clearly mark each subpart of each question.

Provide the corresponding outputs from the R console in a single PDF document, **CS544_HW5_LastName.pdf**

Upload the two files to the Assignments section of Blackboard.

Note: Only ONE submission is allowed. Please be sure that what you are submitting is your final submission.