# CS544 Module 5 Assignment



## Part1) Central Limit Theorem (30 points)

Initialize the city of Boston earnings dataset as shown below:

```
boston <- read.csv(
  "http://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 40000 to 400000 in steps of 20000 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 5000 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.
**c)** Draw 5000 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.
**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.

**a)** Generate 5000 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 1000 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 without replacement. Show the frequencies for the selected departments. Show the percentages of these with respect to sample size.

**b)** Show the sample drawn using systematic sampling. Show the frequencies for the selected departments. Show the percentages of these with respect to sample size.

**c)** 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.

**d)** 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.

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

# Submission:

- You must work on your assignments individually. You are not allowed to copy the answers from the others.
- Each assignment has a strict deadline. However, you are still allowed to submit your assignment within 2 days after the deadline with a penalty. 15% of the credit will be deducted unless you made previous arrangements with your facilitator and professor. Assignments submitted 2 days after the deadline will not be

Create a folder, CS544_HW5_lastName and place the following files 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/PDF document by pasting the output for each question including the plots and explanations.

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