**CS544 Module 5 Assignment**

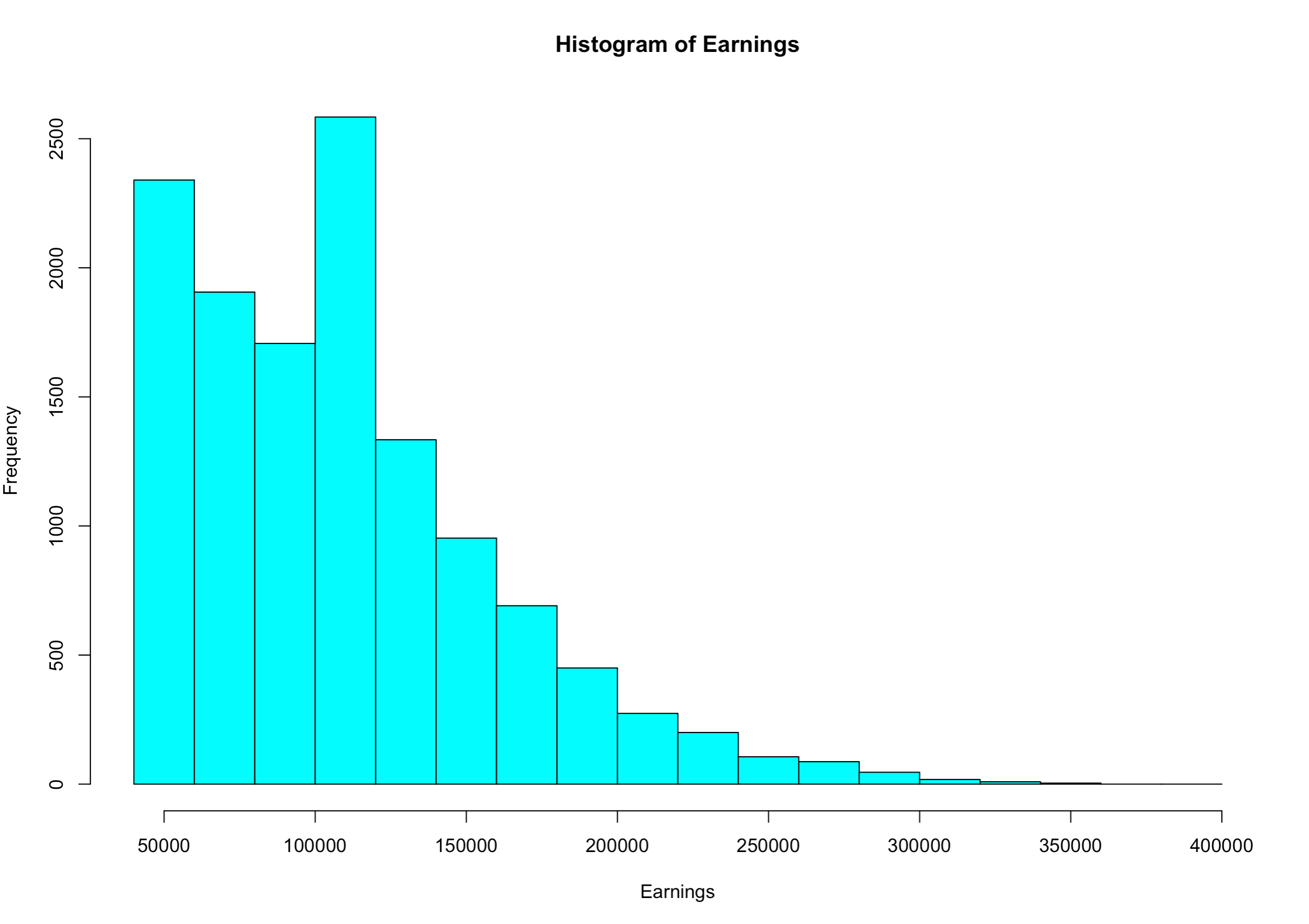
Abraham Cepeda Oseguera

U75425818

**Part 1 – Central Limit Theorem**

Suppose a student has 40% chance of scoring a perfect score in an exam with randomly selected questions. Each student will be provided 5 attempts

1. 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?



Mean = 108,680.9

Standard Deviation = 50,474.7

Description: The data is completely right skewed with a peak between 100,000 and 120,000.

1. 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. Set the start seed for random numbers as the last 4 digits of your BU id.

Chart, histogram

Description automatically generated

Mean = 109,116

Standard Deviation = 15,883

1. 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. Set the start seed for random numbers as the last 4 digits of your BU id.

Chart, histogram

Description automatically generated

Mean = 108,625

Standard Deviation = 7,847

1. Compare of means and standard deviations of the above three distributions.

The means were very similar throughout the three distributions with 108,680.9, 109,116 and 108,625 respectively. On the other hand, there are major differences in the standard deviation between the three distributions with 50,474.7, 15,883, and 7,847 respectively. Even though the means remained very similar, the standard deviation varied massively.

**Part 2 – Central Limit Theorem – Negative Binomial Distribution**

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.

1. Generate 5000 random values from this distribution. Show the barplot with the proportions of the distinct values of this distribution.

Chart, histogram

Description automatically generated

1. 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.

Chart, histogram

Description automatically generated

1. Compare of means and standard deviations of the data from a) with the four sequences generated in b).

|  |  |  |
| --- | --- | --- |
|  | Mean | Standard Deviation |
| Initial Data | 3.0064 | 2.475315 |
| Sample (10) | 2.9972 | 0.7907702 |
| Sample (20) | 2.98135 | 0.5501703 |
| Sample (30) | 3.000633 | 0.4535928 |
| Sample (40) | 3.00795 | 0.4013014 |

The means are very similar for all the datasets since all move around 3. The standard deviation decreases as the sample size increases.

**Part 3 – Sampling**

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.

1. 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.

Text

Description automatically generated with medium confidence

Text

Description automatically generated with medium confidence

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

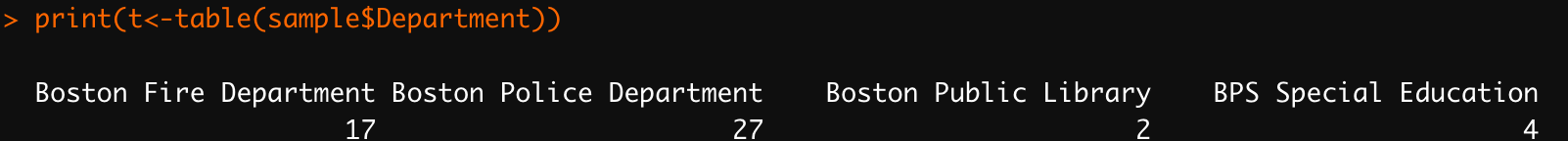
Text

Description automatically generated with medium confidence

Graphical user interface

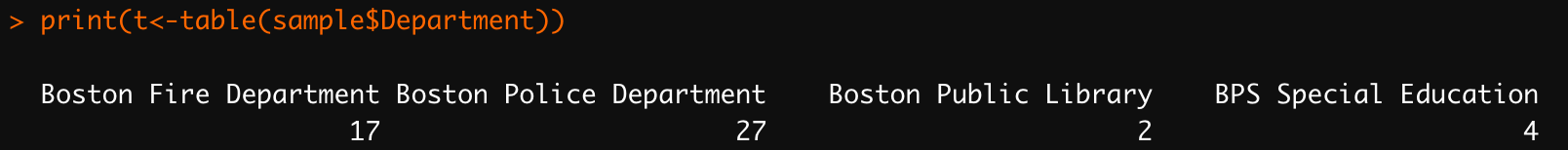
Description automatically generated with low confidence

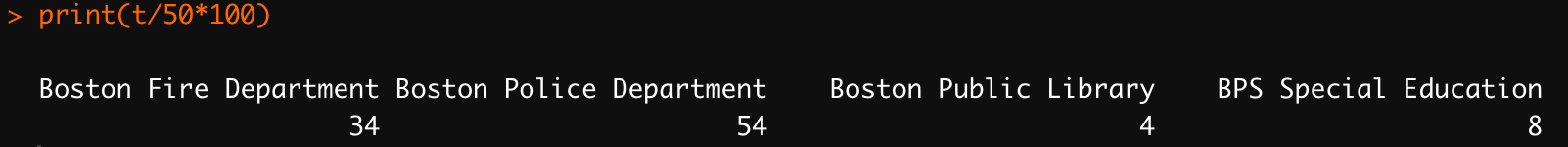
1. Calculate the inclusion probabilities using the Earnings variable. Using these values, show the sample drawn using systematic sampling with unequal probabilities. Show the frequencies for the selected departments. Show the percentages of these with respect to sample size.





1. 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.





1. Compare the means of Earnings variable for these four samples against the mean for the data.

**Extra Credit Questions**

1. If 70% of the population has had their first COVID vaccination, what is the probability that in a random sample of 40 people, more than 35% have not had their first COVID vaccination?
2. If scores on a particular test are normally distributed with an average score of 50 and a standard deviation of 4, what is the probability that the average score of a group of 35 randomly selected people will be less than 49?