Weilin Lu

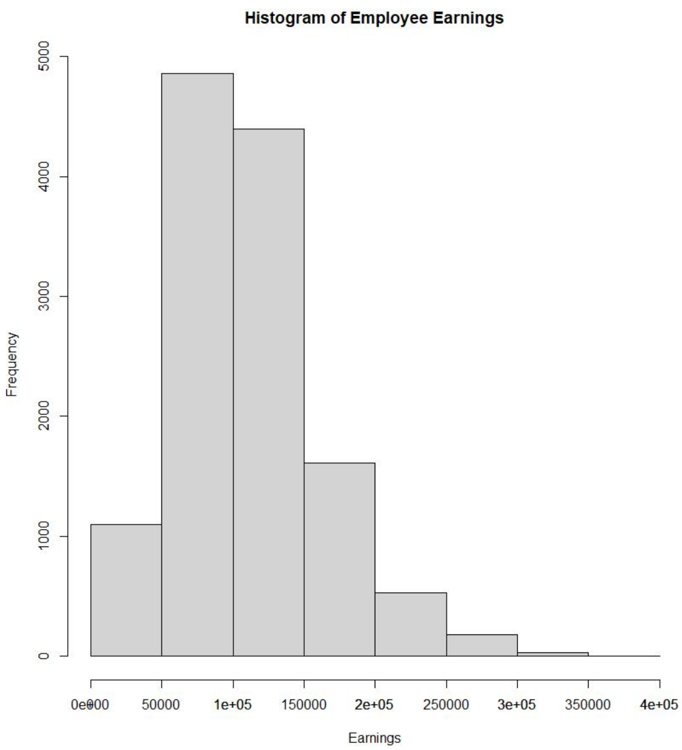
CS-544

Assignment 5

Part1

A) :

```
#part 1
print('Part 1')
#a
print('A')
boston <- read.csv("https://people.bu.edu/kalathur/datasets/bostonCityEarnings.csv", colClasses = c("character", "character", "character", "integ
hist(boston$Earnings, breaks=seq(0, 400000, by=50000), xlab="Earnings", ylab="Frequency", main="Histogram of Employee Earnings")
axis(side=1, at=seq(0, 400000, by=50000), labels=seq(0, 400000, by=50000))
mean_earnings <- mean(boston$Earnings)
sd_earnings <- sd(boston$Earnings)
cat("Mean of Employee Earnings: ", mean_earnings, "\n")
cat("Standard Deviation of Employee Earnings: ", sd_earnings, "\n")
```

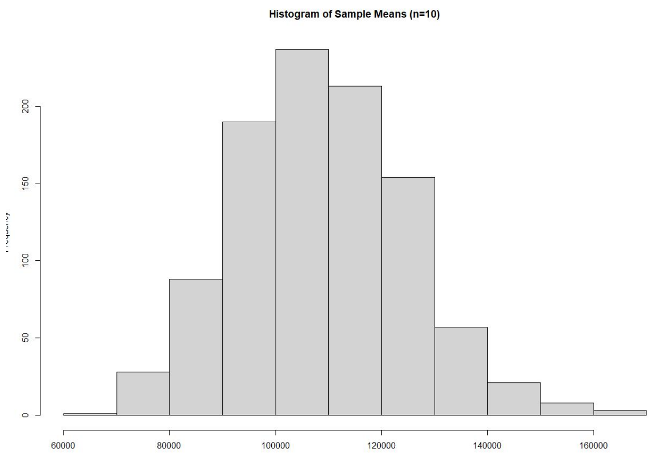



```
> boston <- read.csv("https://people.bu.edu/kalathur/datasets/bostonCityEarnings.csv", colClasses = c("character", "character", "character", "integer", 
haracter"))
> hist(boston$Earnings, breaks=seq(0, 400000, by=50000), xlab="Earnings", ylab="Frequency", main="Histogram of Employee Earnings")
> axis(side=1, at=seq(0, 400000, by=50000), labels=seq(0, 400000, by=50000))
> mean_earnings <- mean(boston$Earnings)
> sd_earnings <- sd(boston$Earnings)
> cat("Mean of Employee Earnings: ", mean_earnings, "\n")
Mean of Employee Earnings:  108680.9
> cat("Standard Deviation of Employee Earnings: ", sd_earnings, "\n")
Standard Deviation of Employee Earnings:  50474.7
```

From the shape of the histogram, we can infer that the employee earnings are positively skewed, meaning that most of the employees earn lower salaries and only a few earn higher salaries.

B) :

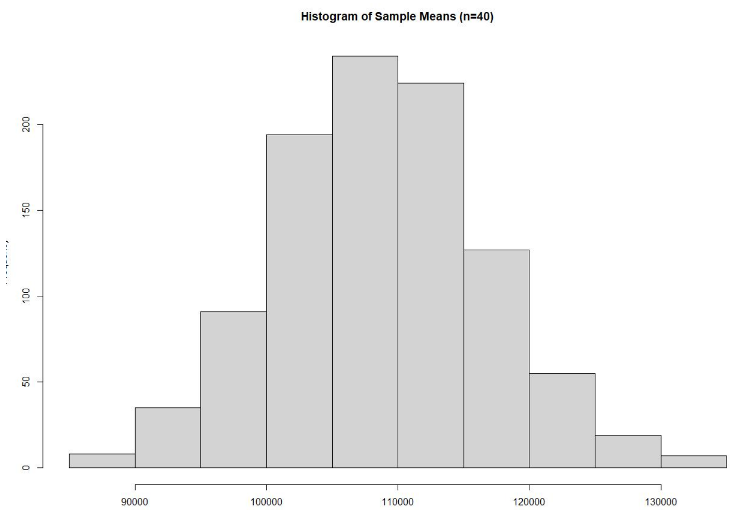
```
#b
print('B')
set.seed(7308)
n_samples <- 1000
n <- 10
sample_means <- replicate(n_samples, mean(sample(boston$Earnings, size=n, replace=FALSE)))
hist(sample_means, main="Histogram of Sample Means (n=10)", xlab="Sample Means", ylab="Frequency")
mean_sample_means <- mean(sample_means)
sd_sample_means <- sd(sample_means)
cat("Mean of Sample Means (n=10): ", mean_sample_means, "\n")
cat("Standard Deviation of Sample Means (n=10): ", sd_sample_means, "\n")
```




```
> cat("Mean of Sample Means (n=10): ", mean_sample_means, "\n")
Mean of Sample Means (n=10):  109120.6
> cat("Standard Deviation of Sample Means (n=10): ", sd_sample_means, "\n")
Standard Deviation of Sample Means (n=10):  16626.66
```

C) ：

```
print('C')
set.seed(7308)
n_samples <- 1000
n <- 40
sample_means <- replicate(n_samples, mean(sample(boston$Earnings, size=n, replace=FALSE)))
hist(sample_means, main="Histogram of Sample Means (n=40)", xlab="Sample Means", ylab="Frequency")
mean_sample_means <- mean(sample_means)
sd_sample_means <- sd(sample_means)
cat("Mean of Sample Means (n=40): ", mean_sample_means, "\n")
cat("Standard Deviation of Sample Means (n=40): ", sd_sample_means, "\n")
#d
```




```
> cat("Mean of Sample Means (n=40): ", mean_sample_means, "\n")
Mean of Sample Means (n=40):  108772
> cat("Standard Deviation of Sample Means (n=40): ", sd_sample_means, "\n")
Standard Deviation of Sample Means (n=40):  8075.851
```

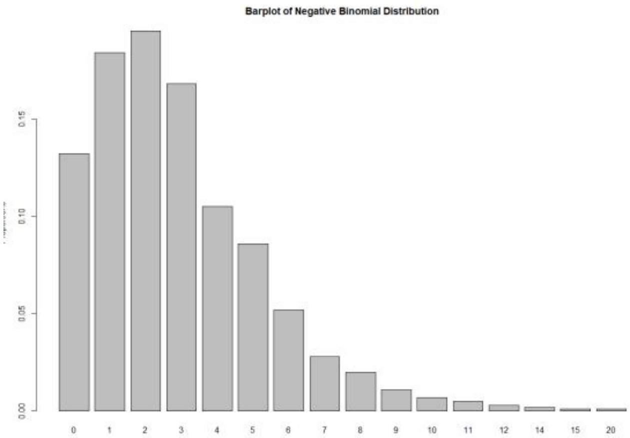
D) ：

Comparing the means and standard deviations of the above three distributions, we can observe that the mean of the population (employee income) is higher than the mean of the sample mean and the standard deviation of the population is higher than the standard deviation of the sample means. As the sample size increases, the distribution of the sample mean becomes more normal and its standard deviation becomes smaller. This is consistent with the central limit theorem.

Part2

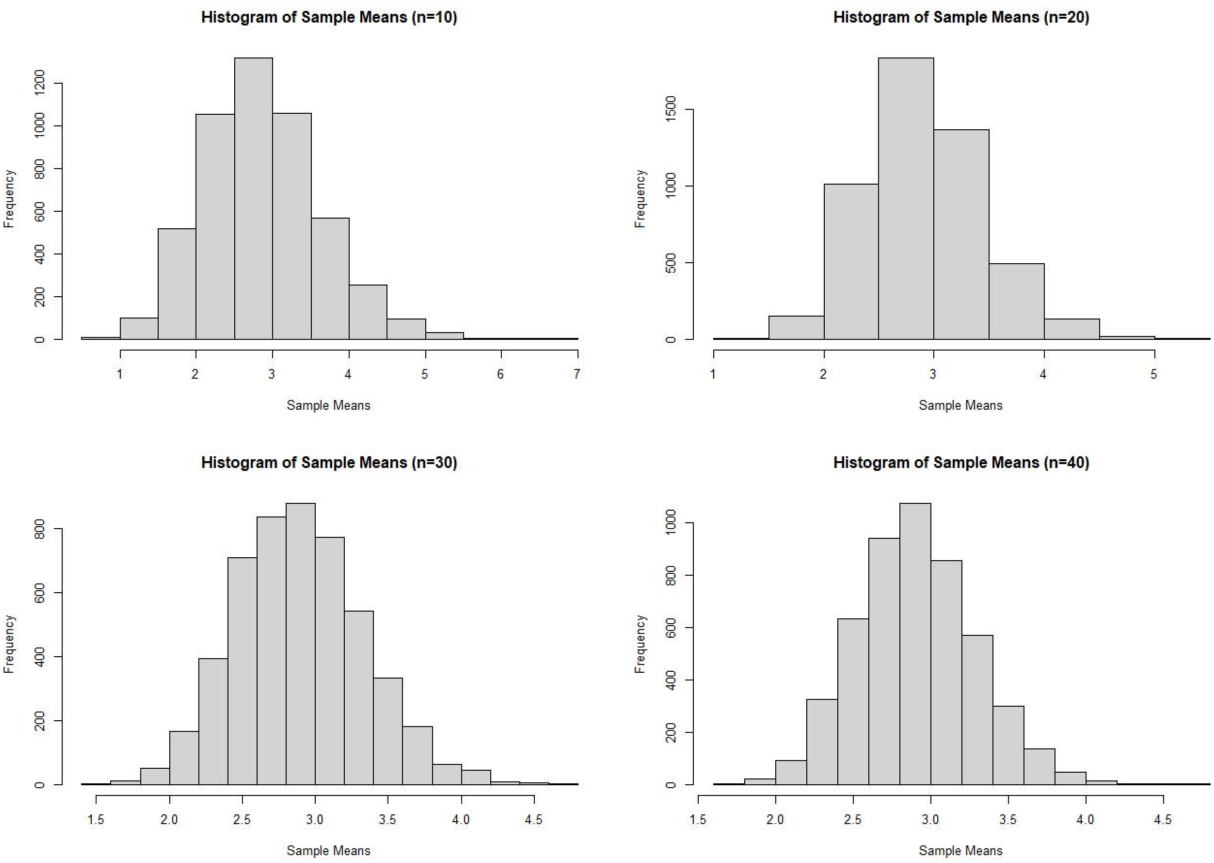
A) :

```
#a
print('A')
set.seed(7308)
n <- 1000
negbin <- rnbinom(n, size=3, prob=0.5)
proportions <- table(negbin) / n
barplot(proportions, xlab="Distinct Values", ylab="Proportions", main="Barplot of Negative Binomial Distribution")
```



B) :

```
#b
print('B')
set.seed(7308)
n_samples <- 5000
sample_sizes <- c(10, 20, 30, 40)
par(mfrow=c(2, 2))
for (i in 1:length(sample_sizes)) {
  n <- sample_sizes[i]
  sample_means <- replicate(n_samples, mean(sample(negbin, size=n, replace=FALSE)))
  hist(sample_means, main=paste("Histogram of Sample Means (n=", n, ")", sep=""), xlab="Sample Means", ylab="Frequency")
}
```



C) :

```r
#c
print('C')
#c
set.seed(7308) # Set the seed for reproducibility
data <- rnbinom(1000, size = 3, prob = 0.5) # Generate data from negative binomial distribution

# Create a matrix to store the means and standard deviations of the different sequences
results <- matrix(0, nrow = 2, ncol = 5)
colnames(results) <- c("Data", "Sample_Size_10", "Sample_Size_20", "Sample_Size_30", "Sample_Size_40")

# Fill in the first column of the matrix with the mean and standard deviation of the data
results[1, 1] <- mean(data)
results[2, 1] <- sd(data)

# Loop over different sample sizes and generate 5000 samples for each sample size
for (i in 1:4) {
  sample_size <- 10*i
  samples <- replicate(5000, sample(data, size = sample_size, replace = FALSE))
  means <- apply(samples, 2, mean) # Calculate the means of each sample
  results[1, i+1] <- mean(means) # Fill in the mean of the means in the matrix
  results[2, i+1] <- sd(means) # Fill in the standard deviation of the means in the matrix
}

# Print the matrix of results
results
```

```
> results
         Data Sample_Size_10 Sample_Size_20 Sample_Size_30 Sample_Size_40
[1,] 2.919000      2.9112600      2.9194000       2.909107      2.9203650
[2,] 2.441393      0.7669295      0.5330014       0.439340      0.3783101
```

We can see that the means of the sample sequences are very close to the mean of the original data, and the standard deviations of the sample sequences are smaller than the standard deviation of the original data. This is in line with the central limit theorem, which states that the sample means tend to be normally distributed around the population mean, with a smaller variance as the sample size increases.

Part3

A) :

```r
#part 3
print('Part 3')
boston <- read.csv(
  "https://people.bu.edu/kalathur/datasets/bostonCityEarnings.csv",
  colClasses = c("character", "character", "character", "integer", "character")
)
top_departments <- names(sort(table(boston$Department), decreasing = TRUE)[1:5])
subset_data <- boston[boston$Department %in% top_departments,]
set.seed(7308)
#a
print('A')
sample_data <- subset_data[sample(nrow(subset_data), 50, replace = TRUE),]
table(sample_data$Department)
prop.table(table(sample_data$Department)) * 100
```

```
> table(sample_data$Department)

  Boston Fire Department Boston Police Department     Boston Public Library  BPS Facility Management     BPS Special Education
                      18                       20                         2                        2                         8
> prop.table(table(sample_data$Department)) * 100

  Boston Fire Department Boston Police Department     Boston Public Library  BPS Facility Management     BPS Special Education
                      36                       40                         4                        4                        16
> |
```

B) :

```r
#b
print('B')
set.seed(7308)
earnings_range <- range(subset_data$Earnings)
step <- diff(earnings_range) / length(subset_data$Earnings)
inclusion_probs <- (subset_data$Earnings - earnings_range[1]) / step
sample_indices <- seq(1, nrow(subset_data), length.out = 50, along.with = inclusion_probs)
sample_data <- subset_data[sample_indices, ]
table(sample_data$Department)
prop.table(table(sample_data$Department)) * 100
```

```
> table(sample_data$Department)

  Boston Fire Department Boston Police Department     Boston Public Library  BPS Facility Management     BPS Special Education
                    1672                     2732                       384                      415                       611
> prop.table(table(sample_data$Department)) * 100

  Boston Fire Department Boston Police Department     Boston Public Library  BPS Facility Management     BPS Special Education
               28.758170                46.990024                  6.604747                 7.137943                 10.509116
```

C) :

```
#c
print('C')
set.seed(7308)
strata_sizes <- table(subset_data$Department)
stratum_sample_sizes <- floor(strata_sizes / sum(strata_sizes) * 50)
stratum_sample <- lapply(split(subset_data, subset_data$Department), function(x) x[sample(nrow(x), stratum_sample_sizes[unique(x$Department)]),
sample_data <- do.call(rbind, stratum_sample)
table(sample_data$Department)
prop.table(table(sample_data$Department)) * 100
```

```
> table(sample_data$Department)

  Boston Fire Department Boston Police Department     Boston Public Library  BPS Facility Management     BPS Special Education
                      14                       23                         3                        3                         5
> prop.table(table(sample_data$Department)) * 100

  Boston Fire Department Boston Police Department     Boston Public Library  BPS Facility Management     BPS Special Education
                29.16667                 47.91667                   6.25000                  6.25000                  10.41667
>
```

D) :

```
#d
print('D')
mean(subset_data$Earnings)
mean(sample_data$Earnings)
set.seed(7308)
simple_random_sample <- subset_data[sample(nrow(subset_data), 50, replace = TRUE),]
mean(simple_random_sample$Earnings)
earnings_range <- range(subset_data$Earnings)
step <- diff(earnings_range) / length(subset_data$Earnings)
inclusion_probs <- (subset_data$Earnings - earnings_range[1]) / step
sample_indices <- seq(1, nrow(subset_data), length.out = 50, along.with = inclusion_probs)
systematic_sample <- subset_data[sample_indices, ]
mean(systematic_sample$Earnings)
strata_sizes <- table(subset_data$Department)
stratum_sample_sizes <- floor(strata_sizes / sum(strata_sizes) * 50)
stratum_sample <- lapply(split(subset_data, subset_data$Department), function(x) x[sample(nrow(x), stratum_sample_sizes[unique(x$Department)]),
stratified_sample <- do.call(rbind, stratum_sample)
mean(stratified_sample$Earnings)
```

```
> mean(subset_data$Earnings)
[1] 133921.4
```

```
> mean(sample_data$Earnings)
[1] 140977.1
```

```
> mean(simple_random_sample$Earnings)
[1] 135844.9
```

```
> mean(systematic_sample$Earnings)
[1] 133921.4
> strata_sizes <- table(subset_data$Department)
> stratum_sample_sizes <- floor(strata_sizes / sum(strata_sizes) * 50)
> stratum_sample <- lapply(split(subset_data, subset_data$Department), function(x) x[sample(nrow(x), stratum_sample_sizes[unique(x$Department)]),])
> stratified_sample <- do.call(rbind, stratum_sample)
> mean(stratified_sample$Earnings)
[1] 132240.6
```

As we can see, the mean earnings for the simple random sample,the stratified sample and systematic sample are relatively close to the mean of the full data-set. This suggests that the systematic sampling method may have captured the full range of earnings levels in the data-set.