Statistical Inference Course Project Part 1

Sarthak

```
# loading the required libraries
library(ggplot2)
```

Simulation Exercise

1. Overview

The goal of this exercise is to investigate the exponential distribution and explore its adherence to the Central Limit Theorem.

2. Simulations

In this exercise, a distribution of means of 40 exponential distributions has to be simulated. Ideally, this would be accomplished by simulating 40 different exponential distributions and calculating their mean, and repeating this process infinite number of times. But, for practical purposes, we will perform 1000 such simulations, which will yield 1000 means.

The sample mean is the mean of the aforementioned distribution, and the sample variance is the variance of this distribution. These values will be calculated in the next 2 sections.

Now, this distribution will be simulated below

```
# setting some default values for this exercise

# setting a seed for reproducibility
set.seed(123)

# initializing a variable to store the lambda value of the exponential distribution
l <- 0.2

# initializing a variable to store the size for each simulation
n <- 40

# initializing a variable to store the number of repetitons of the simulation
repNum <- 1000
```

```
# generating the values for 40 exponential distributions, which is repeated 1000 times
expData <- matrix(rexp(repNum * n, rate = l), nrow = repNum, ncol = n)
# the matrix will having the values of 40 exponential distributions in each row

# calculating the mean for each simulation of 40 exponential distributions
meanSim <- apply(expData, 1, mean)
```

The variable `meanSim` contains the values of this distribution.

3. Sample Mean versus Theoretical Mean

For the exponential distribution, the mean is `1 / lamda`, where `lambda` is defined as the rate parameter. This will act as the theoretical mean of the distribution, since the theoretical mean of the distribution of means of `n` exponential distributions is equal to the population mean.

Now, a histogram of the distribution of means of 40 exponential distributions will be plotted

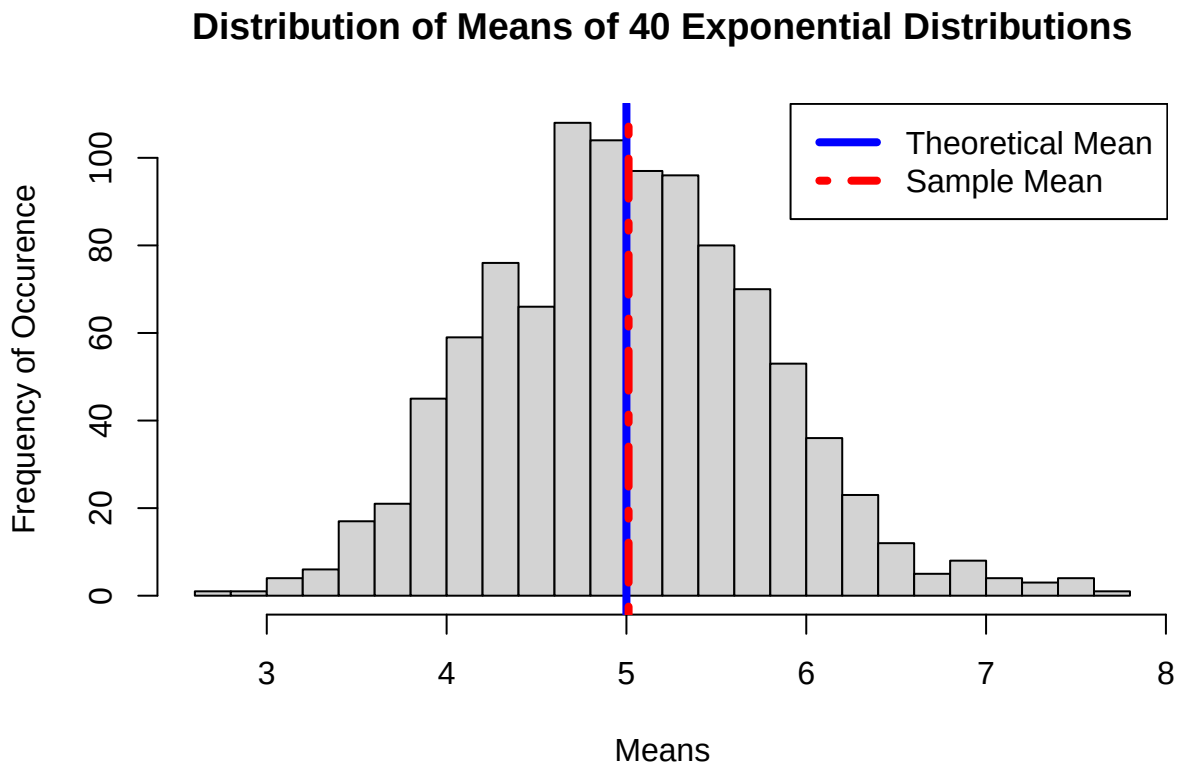
```
# calculating the theoretical mean
theoryMean <- 1 / l

# calculating the sample mean
sampleMean <- mean(meanSim)

# creating the histogram of the distribution of means
hist(meanSim, breaks = 30,
     main = "Distribution of Means of 40 Exponential Distributions",
     xlab = "Means", ylab = "Frequency of Occurence")

# showing the theoretical mean and sample mean in the histogram
abline(v = theoryMean, col = "blue", lwd = 4, lty = 1)
abline(v = sampleMean, col = "red", lwd = 4, lty = 6)

# highlighting the theoretical mean and sample mean
legend(x = "topright", lwd = 4, lty = c(1, 6), col = c("blue", "red"),
       legend = c("Theoretical Mean", "Sample Mean"))
```



From the histogram, the theoretical mean seems to be close to the sample mean. This can be verified from

their numerical values.

```
# displaying the sample mean
sampleMean
```

```
## [1] 5.011911
```

```
# displaying the theoretical mean
theoryMean
```

```
## [1] 5
```

Thus, the sample mean is almost equal to the theoretical mean.

4. Sample Variance versus Theoretical Variance

For this distribution, the theoretical variance will be `((1 / l)^2) / n`, where `1 / l` is the standard deviation of the exponential distribution. Thus, the theoretical standard deviation of the distribution is `(1 / l) / sqrt(n)`. This theoretical variance will be compared to the sample variance.

```
# calculating and displaying the sample variance
sampleVar <- var(meanSim)
sampleVar
```

```
## [1] 0.6088292
```

```
# calculating and displaying the theoretical standard
theoryVar <- ((1 / l)^2) / n
theoryVar
```

```
## [1] 0.625
```

The value of the sample variance is close to that of the theoretical variance.

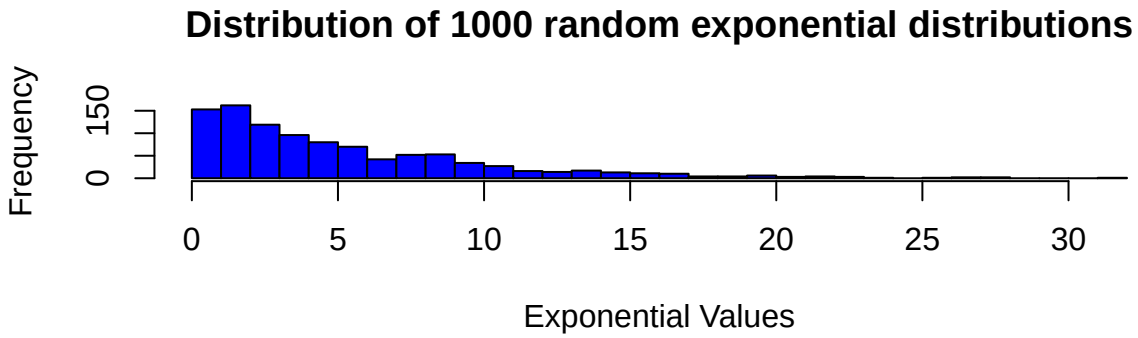
5. Normality of the Distribution

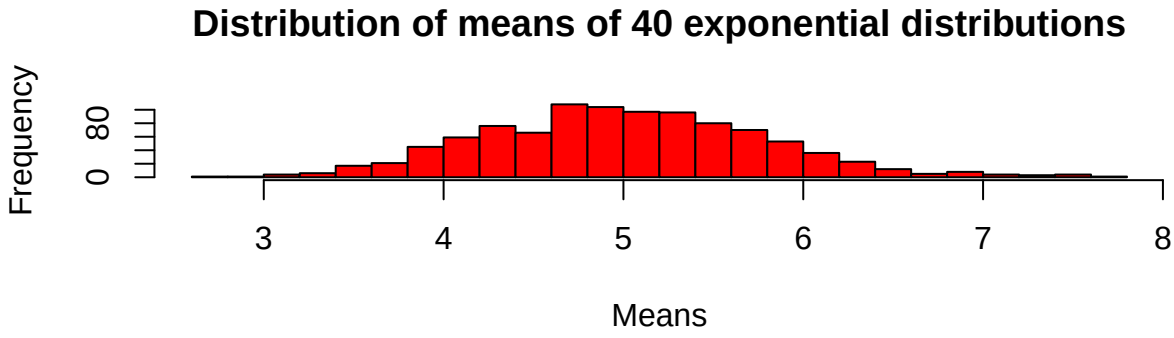
The distribution of means of 40 exponential distributions will be compared to the distribution of 1000 random exponential distributions to explore the normality of the former.

```
# creating a grid for 2 plots
par(mfrow = c(2, 1))

# creating a histogram of the distribution of 1000 random exponential
# distributions
hist(rexp(1000, rate = l), breaks = 30,
     main = "Distribution of 1000 random exponential distributions",
     col = "blue", xlab = "Exponential Values", ylab = "Frequency")

# creating a histogram of the distribution of means of 40 exponential
# distributions
hist(meanSim, breaks = 30, col = "red",
     main = "Distribution of means of 40 exponential distributions",
     xlab = "Means", ylab = "Frequency")
```





From the two histograms, it is clear that the distribution of 1000 random exponential distributions is not normal, rather it is skewed. On the other hand, the distribution of 1000 means of 40 exponential distributions is really close to the standard normal distribution, and is due to CLT, *since the distribution of means contains 1000 means*; and for a sufficiently large sample size, distribution of means of any distribution will approximate to the normal distribution.