Statistical Inference Project 1

Overview

In this project I investigated the exponential distribution in R and compared it with the Central Limit Theorem.

Simulations

The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. I set lambda = 0.2 for all of the simulations and investigated the distribution of averages of 40 exponentials. The number of simulations conducted was 1000.

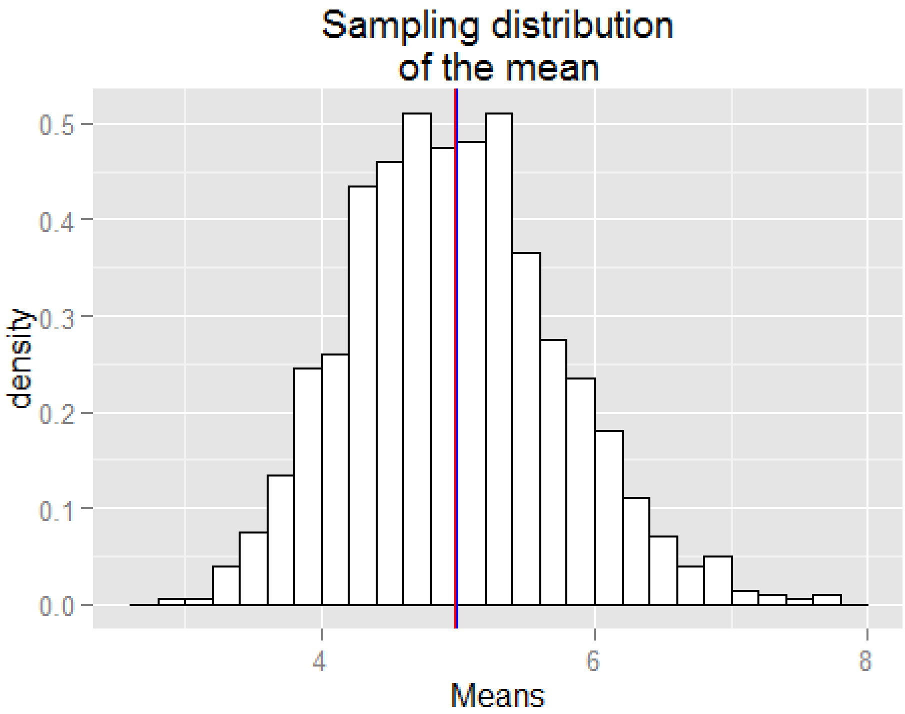
1. Show the sample mean and compare it to the theoretical mean of the distribution.

```r
## Load required libraries
library(ggplot2)
library(grid)
library(gridExtra)
library(modeest)

# Set initial variables
set.seed(1234)
nosim <- 1000
n <- 40
lambda <- 0.2

# Generate simuations and calculate sample means and sds
  simData =  matrix(rexp(n*nosim, lambda), nrow = nosim, ncol = n)
  meanData = data.frame(Means = apply(simData, 1, mean))
  theoreticalMean <- 1/lambda
  sampleMean <- mean(meanData$Means)

# Generate Plot
  ggplot(meanData,aes(x=Means)) +
  geom_histogram(aes(y=..density..),binwidth = lambda,colour="black", fill="white") +
  geom_vline(aes(xintercept = theoreticalMean, color="red")) +
  geom_vline(aes(xintercept = sampleMean, color="blue")) +
  labs(title="Sampling distribution\nof the mean") +
  scale_colour_manual(values = c("red","blue"), name = "Density",
          labels = c("Simulation", "Theoretical"))
```



The graph shown above is the sampling distribution of the mean. The theoretical mean is 5 and is shown as the red line in the graph above. This value is very close to the mean of the sampling distribution of 4.9742388 shown as the blue line in the graph above. These values show that the theoretical mean is almost equal to the mean of the sampling distriution.

This illustrates the law of large numbers which states that if you take samples of larger and larger size from any population, then the mean of the sample tends to get closer and closer to the population mean which is also the theoretical mean.

2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

```
populationSd <- 1/lambda
theoreticalSd <- (1/lambda)/sqrt(n)
theoreticalVariance <- theoreticalSd^2
sampleStandardDeviation<-  sd(meanData$Means)
sampleVariance <- var(meanData$Means)
```

The variance of the sampling distriubtion shows the degree to which the means from the different samples differ from each other and from the population mean. It gives you a sense of how close the particular sample mean is likely to be to the population mean

The theoretical variance is 0.625 which is similar to the variance of the sampling distribution of 0.5949702. The variance of the sampling distribution is small meaning that the sample means did not vary very much from each other and were very close to the population mean.

According to the Central LimitTheorem the higher the sample size used the narrower the spread of the sample distribution will become.

3. Show that the distribution is approximately normal.

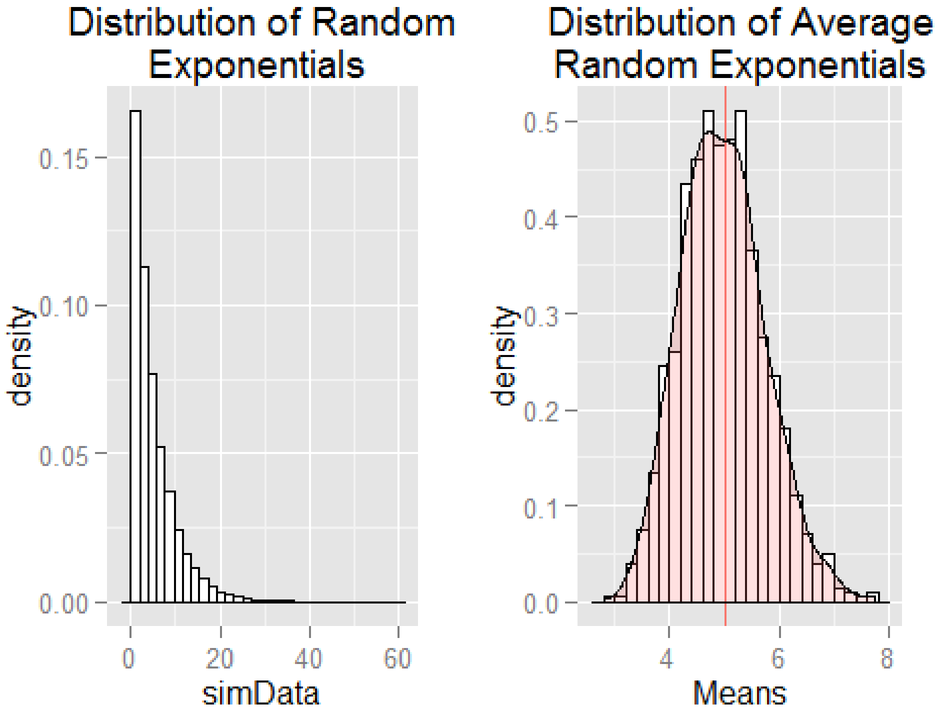
```
## Generate variable and data
sampleData <- data.frame(simData = rexp(n*nosim, lambda))
sampleMode <- mlv(meanData$Means, method ="mfv")
sampleMedian <- median(meanData$Means)

## Generate Plot
```

```
p1 <- ggplot(sampleData,aes(x=simData)) +
     geom_histogram(aes(y=..density..),colour="black", fill="white") +
     labs(title="Distribution of Random\nExponentials ")

p2 <- ggplot(meanData,aes(x=Means)) +
      geom_histogram(aes(y=..density..),binwidth = lambda,colour="black",
fill="white") +
      geom_vline(aes(xintercept = theoreticalMean, color="red")) +
      labs(title="Distribution of Average\nRandom Exponentials") +
     geom_density(alpha=.2, fill="#FF6666")  # Overlay with transparent density plot

grid.arrange(p1,p2, ncol=2)
```



As you can see from the graphs shown above the distribution of 40 random exponential exhibits a exponential distribution. However when the averages of 1000 trials of 40 random exponentials is shown the distribution exhibits a normal distribution.

The sampling distribution of the mean is a normal distribution because it demonstrates the following characteristics.

1. Its mean, median and mode are equal.
   - Mean is 4.9742388
   - Median is 4.931579
   - Mode is 4.9742388
2. It is symmetrical which means that if the distribution is cut in half, each side would be the mirror of the other.

This demonstrates the Cental Limit Theorem which states that the distribution of the sample means will become normally distributed regardless of the shape of the parent population provided the sample size is large enough. The sample size of 40 appears large enough to prove the central limit theorem.