Statistical Inference Course Project

Part 1: Simulation Exercises

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 . For this simulation, we set lambda=0.2 .

Through this simulation we will be able to investigate the distribution of the averages of 40 numbers sampled from an exponential distribution with a lambda value equal to 0.2.

First, we need to do a thousand simulated averages of 40 exponentials, so that we may have enough sampled data to work.

Simulating the data sample

```
#  Load libraries
library(ggplot2)
#  Set a seed so that you may reproduce the exact same random sample numbers
set.seed(1)
#  Set values for lambda, the number of simulations we'll do, and the sample size
lambda <- 0.2
simulations <- 1000
sample <- 40
#  Create a matrix that stores 1000 random simulations 40 times.
data <- matrix(rexp(simulations*sample, rate=lambda), simulations, sample)
averages <- rowMeans(data)
```

Comparing the sample distribution to the theoretical distribution

```
# Compare the sample mean to the theoretical mean.
samplemean <- mean(averages)
theorymean <- 1/lambda
```

```
samplemean
## 4.990025
theorymean
## 5
```

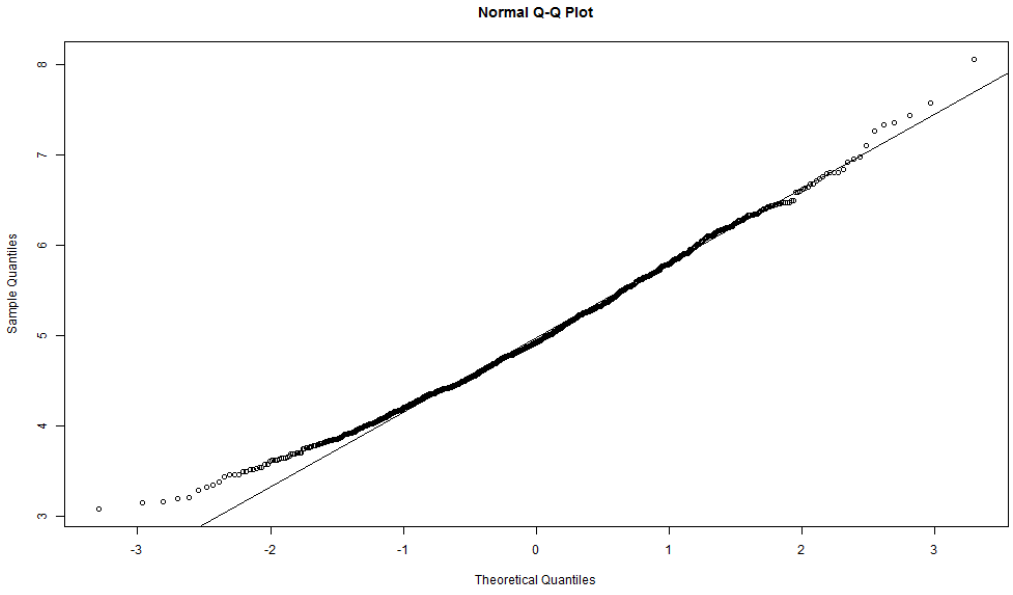
The distribution of sample means is centered at 4.990025 and the theoretical center of the distribution is lambda^-1 = 5 .

```
# Compare the sample variance with the theoretical variance
samplevariance <- var(averages)
theoryvariance <- (1/lambda)^2/n
samplevariance
##  0. 6177072
theoryvariance
##   0.625
```

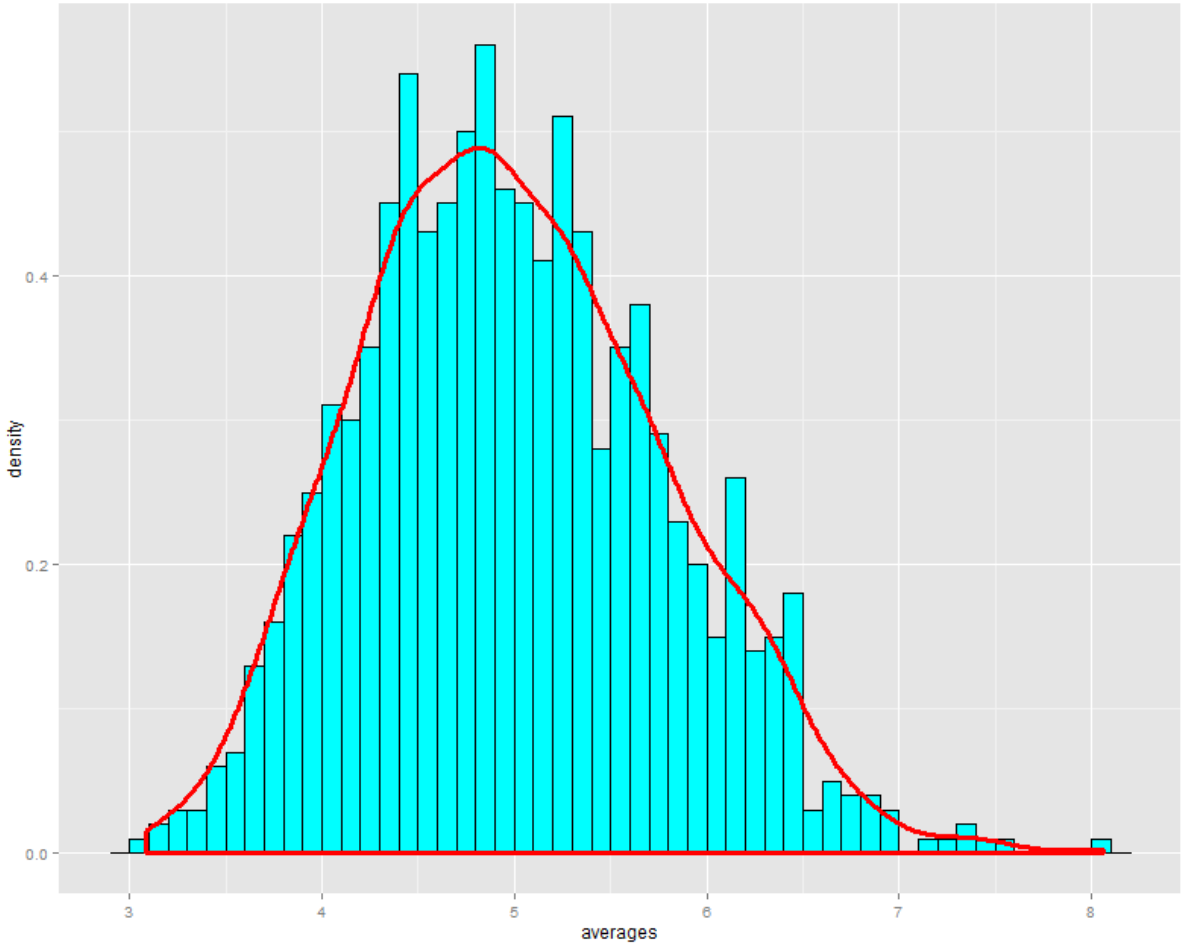
The variance of sample means is 0. 6177072 where the theoretical variance of the distribution is sigma^2 / n = 1/(lambda^2 n) = 1/(0.04 times 40) = 0.625 .

According to the central limit theorem, the averages of samples follow normal distribution. Also, this q-q plot below suggests the normality.

```
qqnorm(averages); qqline(averages)
```



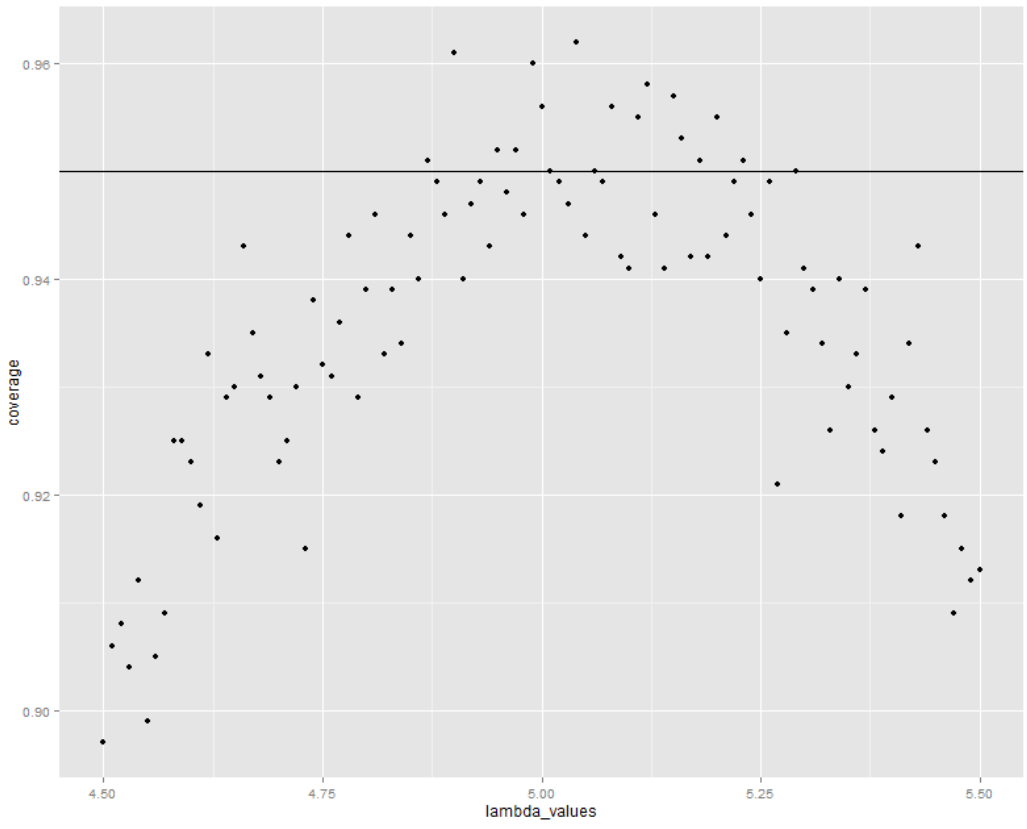
```
# Plot the histogram of averages to compare the sample means to the theoretical distribution.
plotavgs <- data.frame(averages)
graf <- ggplot(plotavgs, aes(x=averages))
graf <- graf + geom_histogram(aes(y="density"), colour="black", fill="cyan", binwidth=1/10)
graf + geom_density(colour="blue", size=1)
```



Comparing the sample distribution (cyan) to the theoretical distribution (red), we can clearly see that the distribution appears fairly close to a normal one, as both distributions appear to follow the same pattern.

We also need to calculate and compare the confidence intervals for the sample distribution and for the theoretical distribution.

```
realconfint <- round(mean(averages) + c(-1,1)*1.96*sd(averages)/sqrt(sample),3)
theoryconfint <- theorymean + c(-1,1)*1.96*sqrt(theoryvariance)/sqrt(sample)
realconfint
## 4.746  5.234
theoryconfint
## 4.755  5.245
# Plot the intercept for the confidence interval
coverage <- sapply(lambda_values, function(lamb){
mus <- rowMeans(matrix(rexp(sample*simulations, rate=0.2), simulations, sample))
lowerlimit <- mus -qnorm(0.975)*sqrt(1/lambda^2/sample)
upperlimit <- mus +qnorm(0.975)*sqrt(1/lambda^2/sample)
mean(lowerlimit < lamb & upperlimit > lamb)} )
qplot(lambda_values, coverage) + geom_hline(yintercept=0.95)
```



The 95% confidence intervals for the rate parameter (lambda) seem fairly similar, wich means at least 95% of the values will fall within those values for both the sample distribution as in the theoretical distribution.

Conclusions

We may conclude that the sample distribution approaches the normal distribution as the central limit theorem implies, meaning we can run random simulations of data, for data sets with normal distributions in R, and get fairly close approximations.