Statistical Inference

Rishi

28 August 2016

Overview

The purpose of this data analysis is to investigate the exponential distribution and compare it to the Central Limit Theorem. For this analysis, the lambda will be set to 0.2 for all of the simulations. This investigation will compare the distribution of averages of 40 exponentials over 1000 simulations.

Simulations

Set the simulation variables lambda, exponentials, and seed.

```
set.seed(3)
lambda <- 0.2
numSim <- 1000
sampleSize <- 40
sim <- matrix(rexp(numSim*sampleSize, rate=lambda), numSim, sampleSize)
rowMeans <- rowMeans(sim)
```

The distribution of sample means is as follows.

Plot histogram of averages

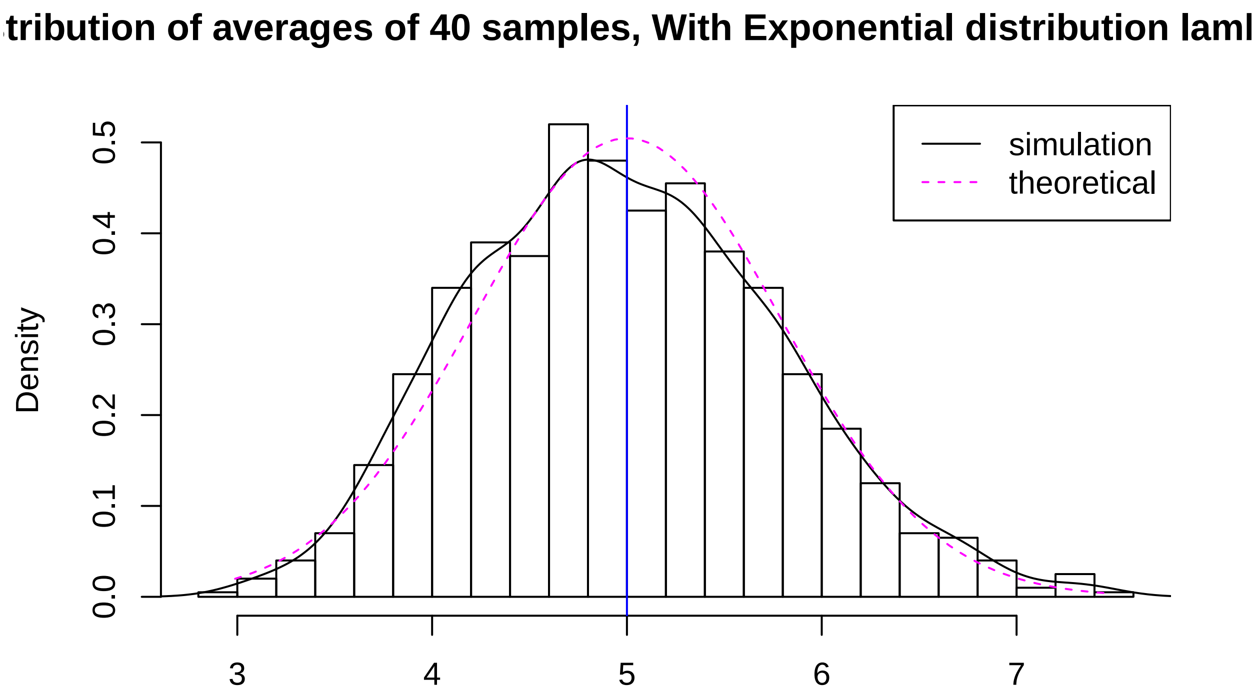
```
hist(rowMeans, breaks=20,prob=T, main="Distribution of averages of 40 samples, With Exponential distribu

#Density of Averages of Samples
lines(density(rowMeans))

# Theoratical Center of distribution
abline(v=1/lambda, col="blue")

# Theoretical density of the averages of samples
xfit <- seq(min(rowMeans), max(rowMeans), length=100)
yfit <- dnorm(xfit, mean=1/lambda, sd=(1/lambda/sqrt(sampleSize)))
lines(xfit, yfit, pch=22, col="magenta", lty=2)

# add legend
legend('topright', c("simulation", "theoretical"), lty=c(1,2), col=c("black", "magenta"))
```

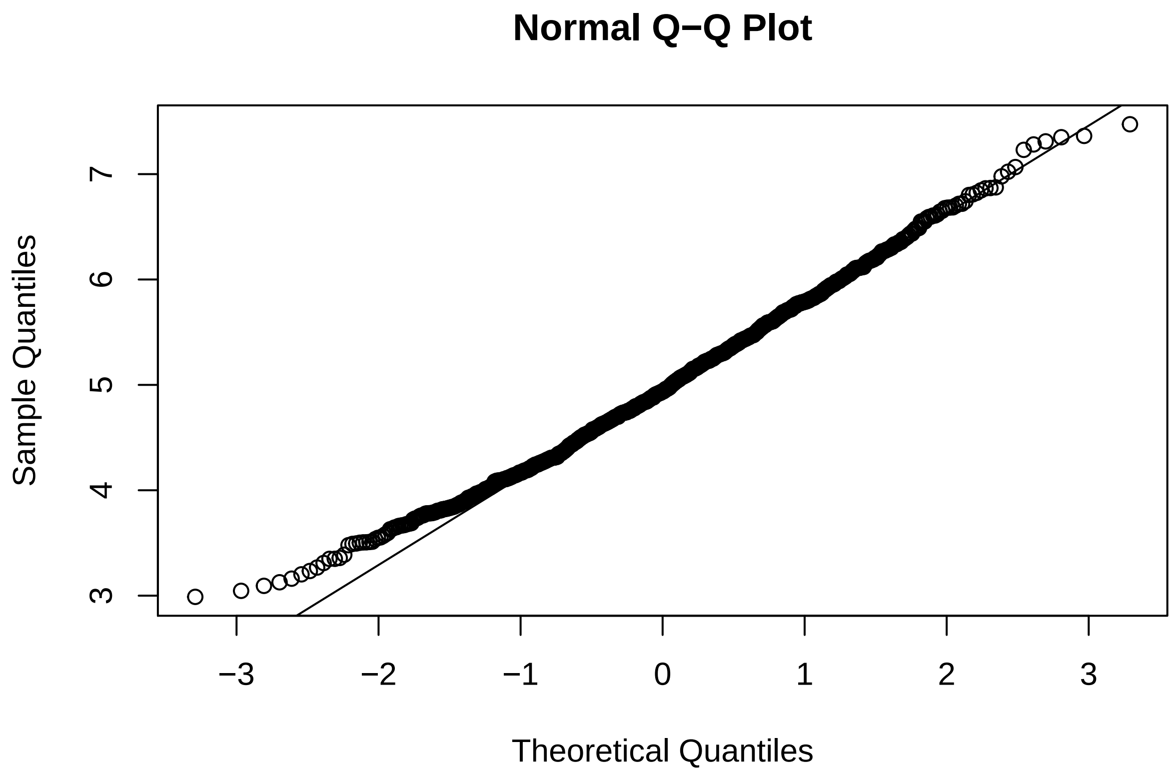


Due to the central limit theorem, the averages of samples follow normal distribution. The figure above also shows the density computed using the histogram and the normal density plotted with theoretical mean and variance values. Also, the q-q plot below suggests the normality.

The distribution of sample means is centered at 4.9866197 and the theoretical center of the distribution is 5. The variance of sample means is 0.6257575 where the theoretical variance of the distribution is = 0.625

Due to the central limit theorem, the averages of samples follow normal distribution. The figure above also shows the density computed using the histogram and the normal density plotted with theoretical mean and variance values. Also, the q-q plot below suggests the normality.

```r
qqnorm(rowMeans); qqline(rowMeans)
```

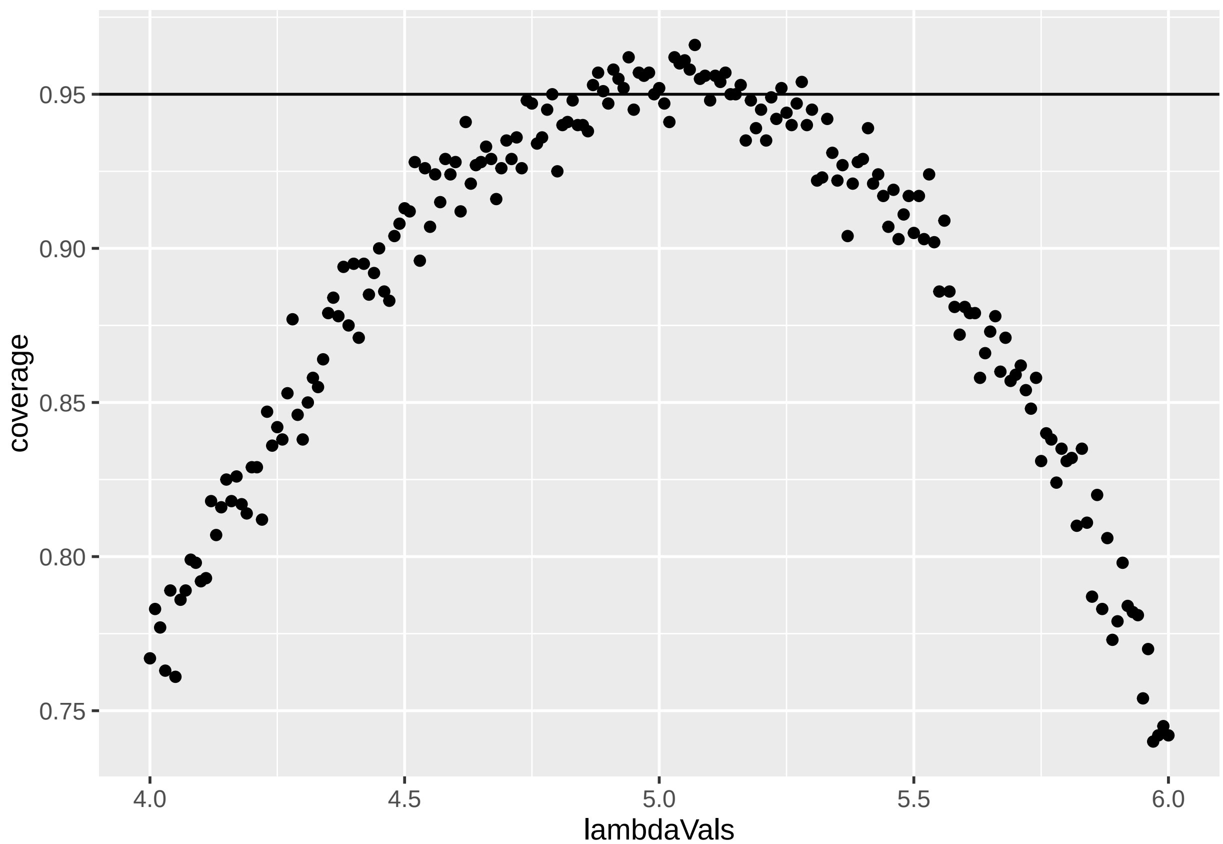


Finally, let's evaluate the coverage of the confidence interval for

$1/\lambda = \bar{X} \pm 1.96\frac{S}{\sqrt{n}}$

```
lambdaVals <- seq(4, 6, by=0.01)
coverage <- sapply(lambdaVals, function(calLambda) {
    muConf <- rowMeans(matrix(rexp(sampleSize*numSim, rate=0.2),
                               numSim, sampleSize))
    ll <- muConf - qnorm(0.975) * sqrt(1/lambda**2/sampleSize)
    ul <- muConf + qnorm(0.975) * sqrt(1/lambda**2/sampleSize)
    mean(ll < calLambda & ul > calLambda)
})

library(ggplot2)
qplot(lambdaVals, coverage) + geom_hline(yintercept=0.95)
```



The 95% confidence intervals for the rate parameter (λ) to be estimated ($\hat{\lambda}$) are $\hat{\lambda}low = \hat{\lambda}(1-\frac{1.96}{\sqrt{n}})$ and $\hat{\lambda}upper = \hat{\lambda}(1+\frac{1.96}{\sqrt{n}})$.

For selection of $\hat{\lambda}$ around 5, Average of the sample mean falls within the confidence interval at least 95% of the time.