Statistical Inference Course Project

Patihe Suip

November 21, 2015

This is the project for the statistical inference class. In it, you will use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. 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. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should 1. Show the sample mean and compare it to the theoretical mean of the distribution. 2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. 3. Show that the distribution is approximately normal.

In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials. Below are simulation to explore inference;-

Setting the environment

```
setwd("D:\\Data Scientist\\Statiscal Inference\\Project")
```

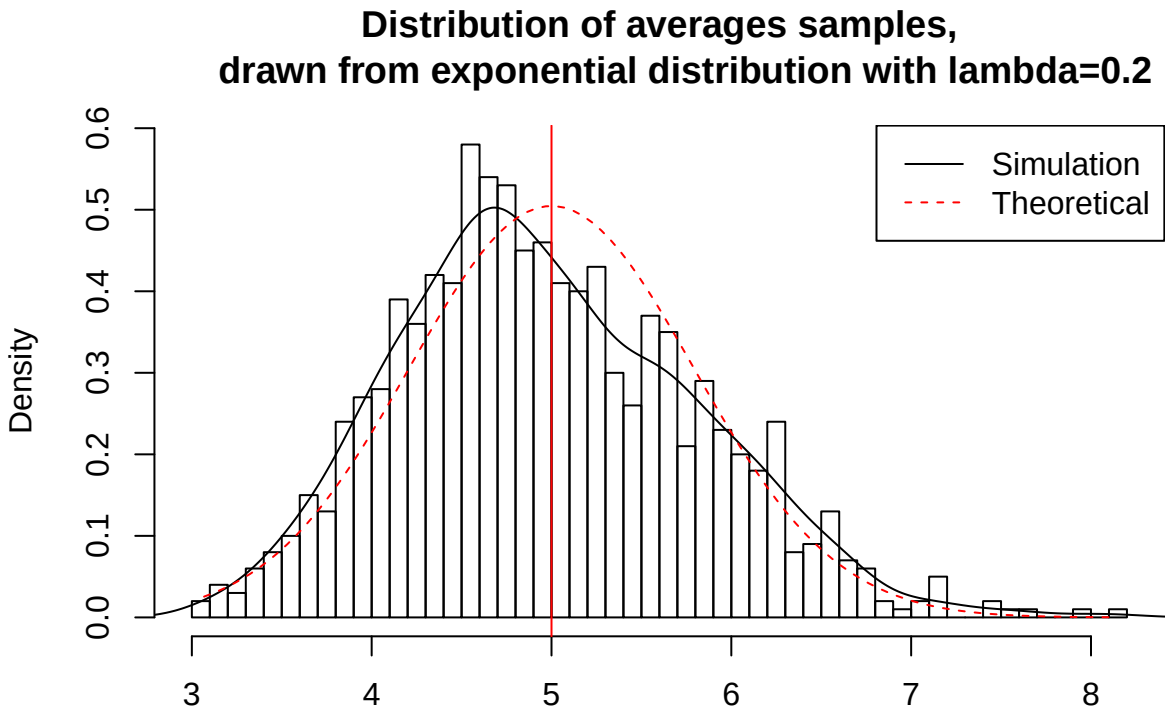
By doing a thousand simulated averages of 40 exponentials.

```
set.seed(13)
lambda <- 0.2
num_sim <- 1000
sample_size <- 40
sim <- matrix(rexp(num_sim*sample_size, rate=lambda), num_sim, sample_size)
row_means <- rowMeans(sim)
```

Sample means are plotted in histogram along with theoretical mean and theoretical density.

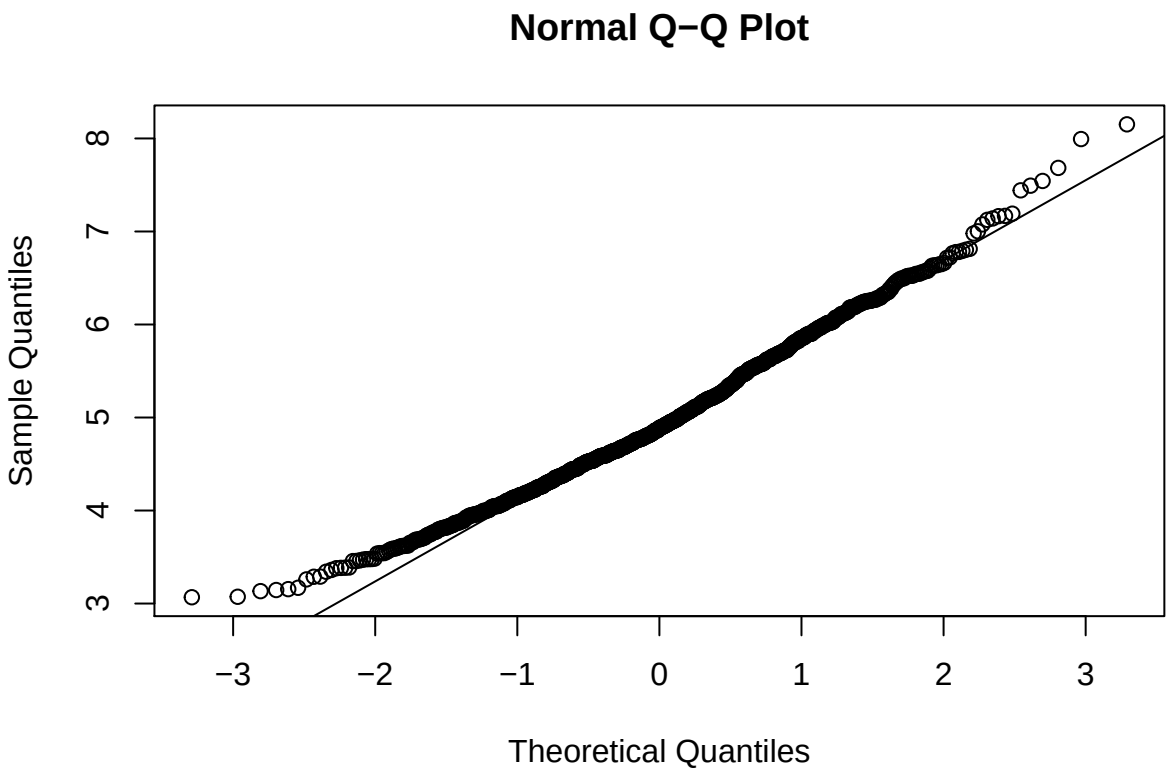
```
# Ploting the histogram of averages
hist(row_means, breaks=50, prob=TRUE,
     main="Distribution of averages samples,
     drawn from exponential distribution with lambda=0.2",
     xlab="")

# Density of the averages samples
lines(density(row_means))
```

```r
# Theoretical center of distribution
abline(v=1/lambda, col="red")

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

# Adding legend
legend('topright', c("Simulation", "Theoretical"), lty=c(1,2), col=c("black", "red"))
```



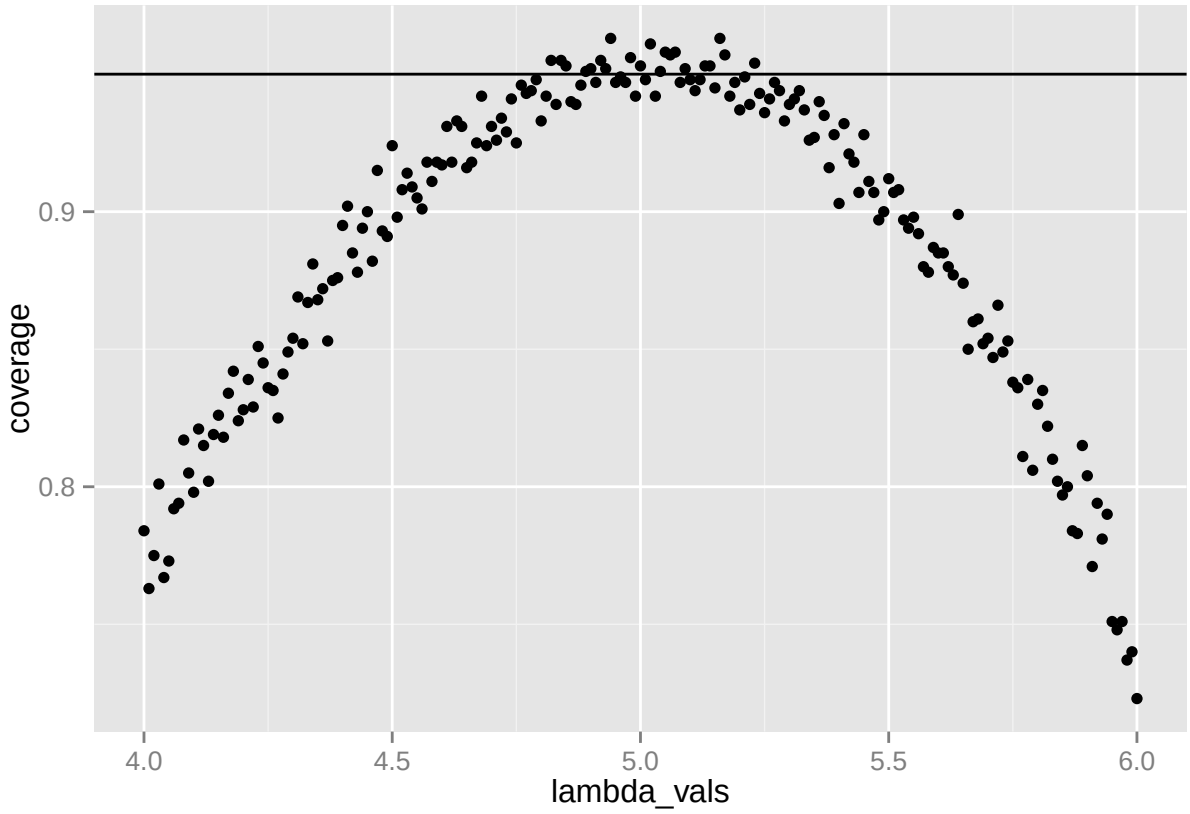
Distribution of sample means is centered at 4.9725119 and the theoretical center of the distribution is $\lambda^{-1} = 5$. The variance of sample means is 0.6849965 where the theoretical variance of the distribution is $\sigma^2/n = 1/(\lambda^2 n) = 1/(0.04 \times 40) = 0.625$.

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.



Final, evaluating the coverage of the confidence interval for $1/\lambda = \bar{X} \pm 1.96\frac{S}{\sqrt{n}}$

```
## Warning: package 'ggplot2' was built under R version 3.2.1
```



The 95% confidence intervals for the rate parameter (λ) to be estimated ($\hat{\lambda}$) are $\hat{\lambda}_{low} = \hat{\lambda}(1 - \frac{1.96}{\sqrt{n}})$ agnd $\hat{\lambda}_{upp} = \hat{\lambda}(1 + \frac{1.96}{\sqrt{n}})$. As can be seen from the plot above, for selection of $\hat{\lambda}$ around 5, the average of the sample mean falls within the confidence interval at least 95% of the time. Note that the true rate, λ is 5.