Statistical Inference - Coursera

Patricio Del Boca

21/09/2015

Overview:

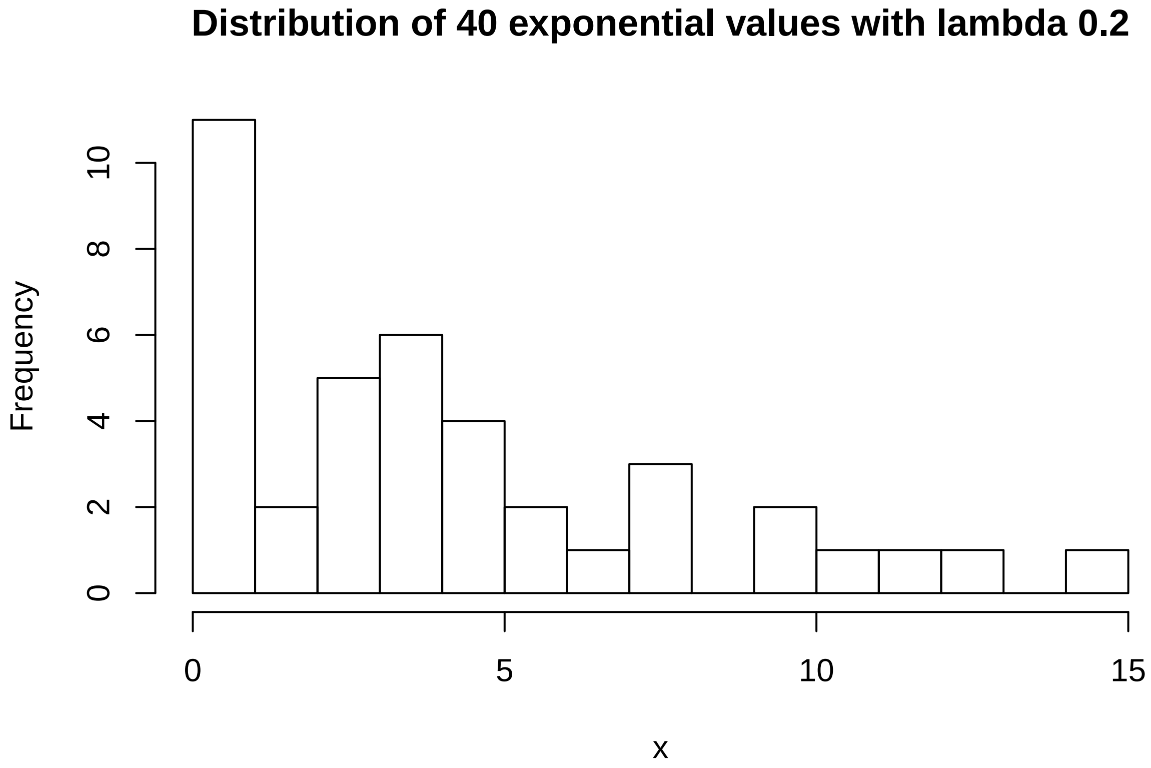
In this project I will investigate the exponential distribution in R and compare it with the Central Limit Theorem. **I will investigate the distribution of averages of 40 exponentials by doing 1000 simulations.**

Simulations:

The exponential distribution can be simulated in R with **rexp(n, lambda)** where lambda is the rate parameter. I will set lambda = 0.2 for all of the simulations.

```r
set.seed(31415)
lambda <- 0.2
n <- 40
simulations <- 1000
```

```r
x <- rexp(n, rate = lambda)
title <- sprintf("Distribution of %g exponential values with lambda %g", n, lambda)
hist(x, breaks = 20, main = title)
```



Sample Mean versus Theoretical Mean:

Theoretical Mean should be equal to Empirical Mean as the sample size trends to infinity. Empirical Mean should be more accurate as the size of the sample increases.

With sample size equals 40:

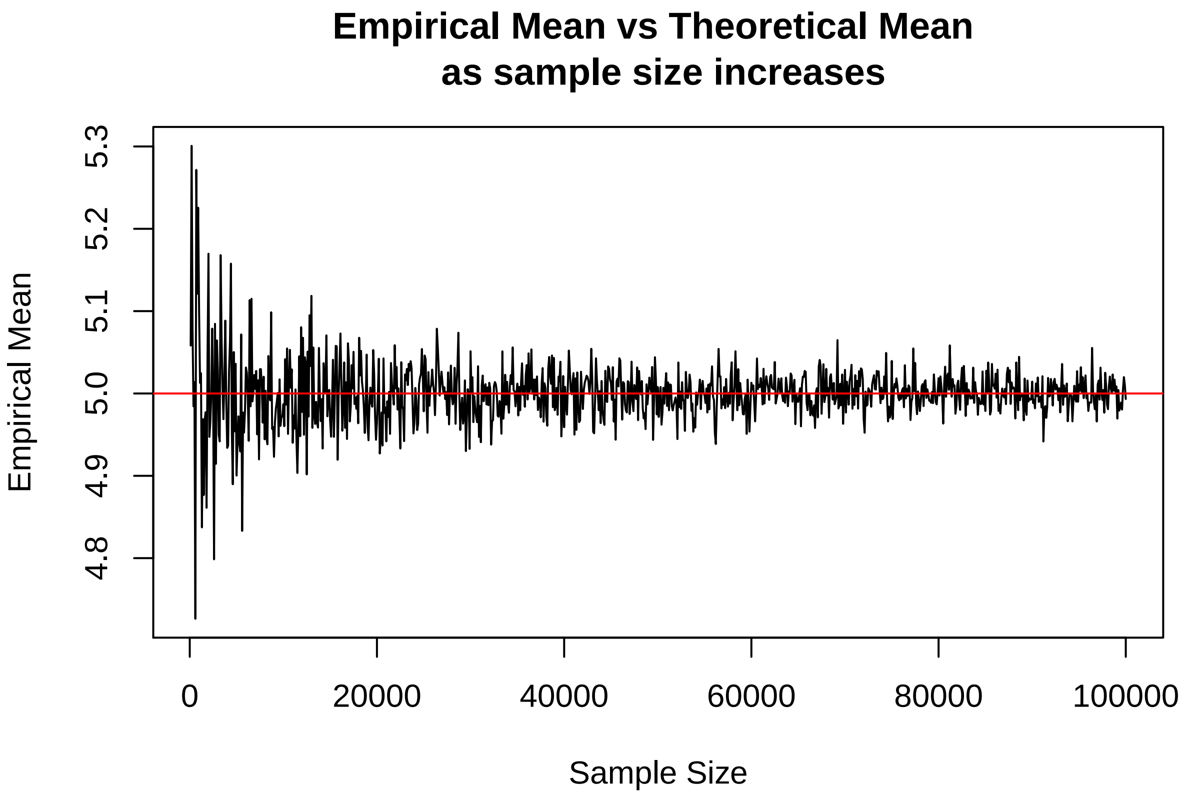
```r
theoreticalMean <- 1 / lambda
empiricalMean <- mean(x)
sprintf("Theoretical Mean: %g - Empirical Mean with n=40: %g", theoreticalMean, empiricalMean)
```

```
## [1] "Theoretical Mean: 5 - Empirical Mean with n=40: 4.20727"
```

How much varies the mean and the (Theoretical Mean - Empirical Mean) as the size of the sample increases?

```r
sampleSizes <- seq(100, 100000, 100)

means <- NULL
for (i in 1:length(sampleSizes)){
    thisMean <- mean(rexp(sampleSizes[i], rate = lambda))
    means <- c(means, thisMean)
}
title <- sprintf("Empirical Mean vs Theoretical Mean \n as sample size increases")
xlab = "Sample Size"
ylab = "Empirical Mean"
options(scipen = 20) # Remove scientific notation on x axis.
plot(x = sampleSizes, y = means, type='l', main=title, xlab = xlab, ylab = ylab)
abline(h=theoreticalMean, col = "red")
```



Sample Variance versus Theoretical Variance:

I will compare Standard Deviation since is most easy to understand due that it is in the same order of magnitude than de variable.

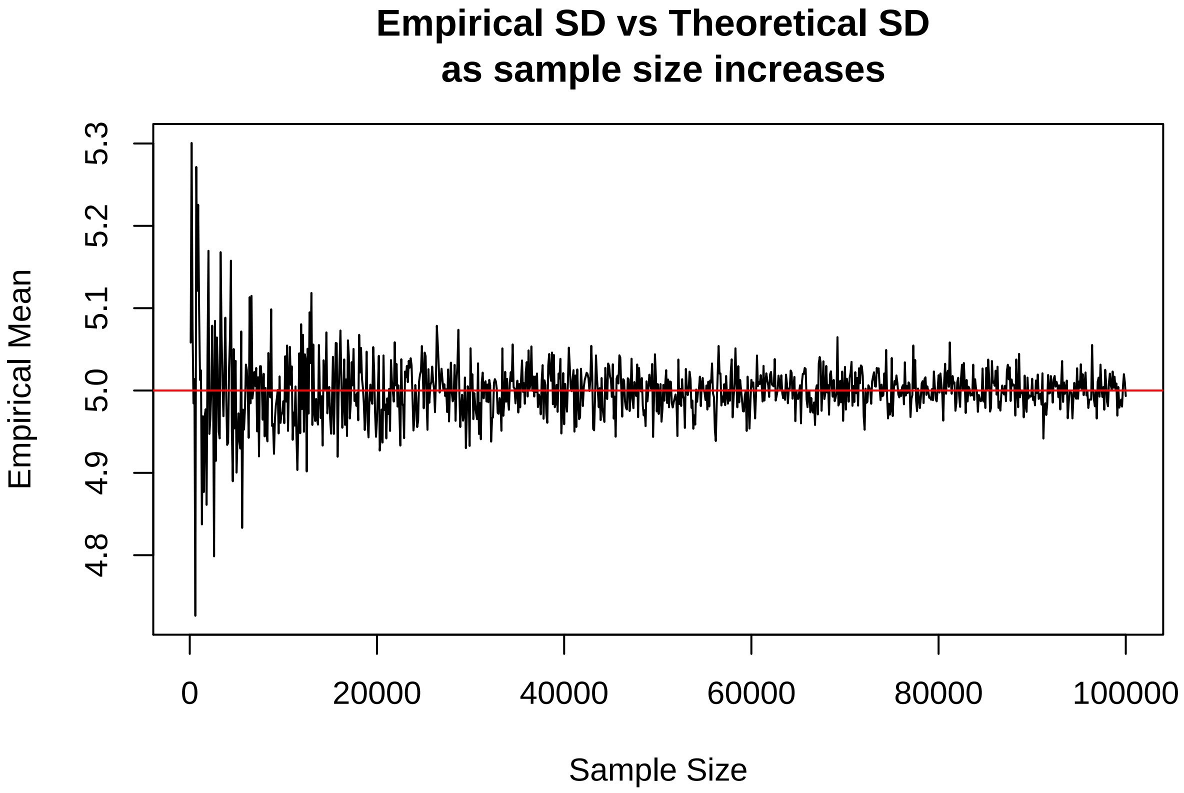
The estimate (sd) behaves like the mean, it increases its accuracy as n trends to infinity.

```
theoreticalSD <- 1 / lambda
empiricalSD <- sd(x)
sprintf("Theoretical SD: %g - Empirical SD with n=40: %g", theoreticalSD, empiricalSD)
```

```
## [1] "Theoretical SD: 5 - Empirical SD with n=40: 3.7531"
```

How much varies the Standard Deviation and the error (Theoretical SD - Empirical SD) as the size of the sample increases?

```
sds <- NULL
for (i in 1:length(sampleSizes)){
    thisSD <- sd(rexp(sampleSizes[i], rate = lambda))
    sd <- c(sds, thisSD)
}
title <- sprintf("Empirical SD vs Theoretical SD \n as sample size increases")
xlab = "Sample Size"
ylab = "Empirical Mean"
options(scipen = 20) # Remove scientific notation on x axis.
plot(x = sampleSizes, y = means, type='l', main=title, xlab = xlab, ylab = ylab)
abline(h=theoreticalSD, col = "red")
```

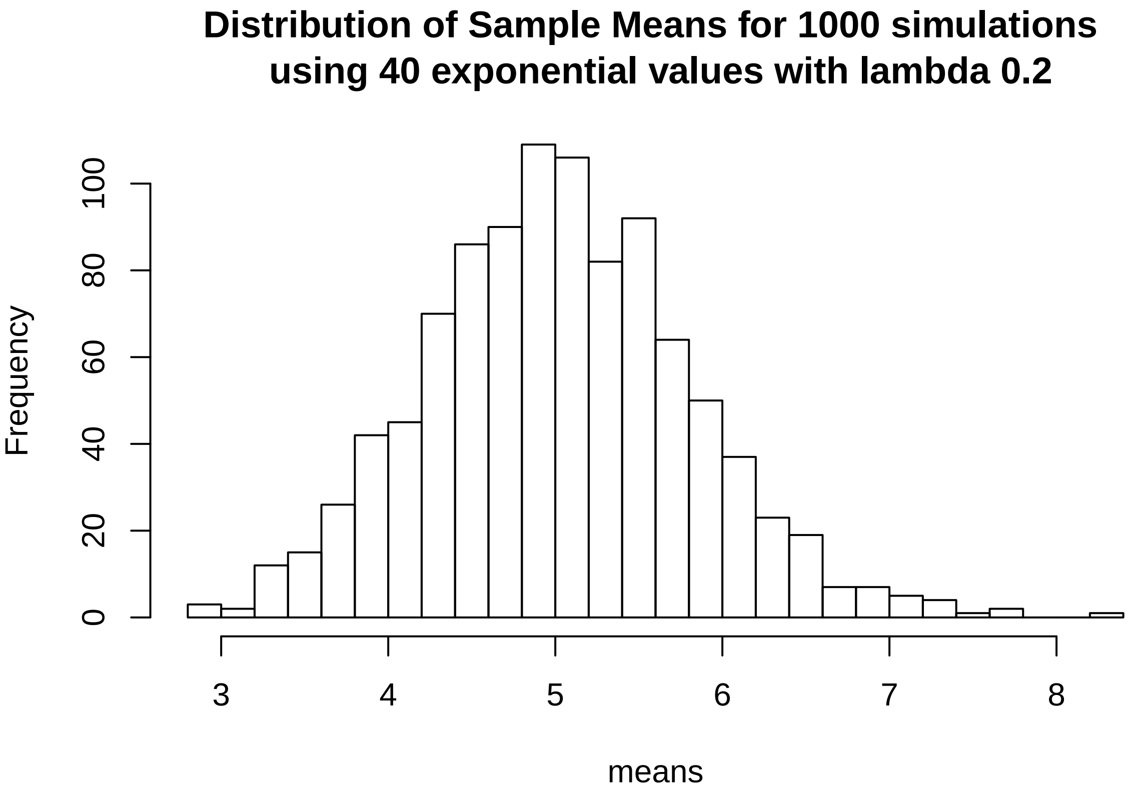


Distribution:

Now I will simulate a thousand times, and for each simulation extract the mean. The plot of the means distribution should be "gaussian" as the CLT exposes. Let's plot the distribution of a thousands means.

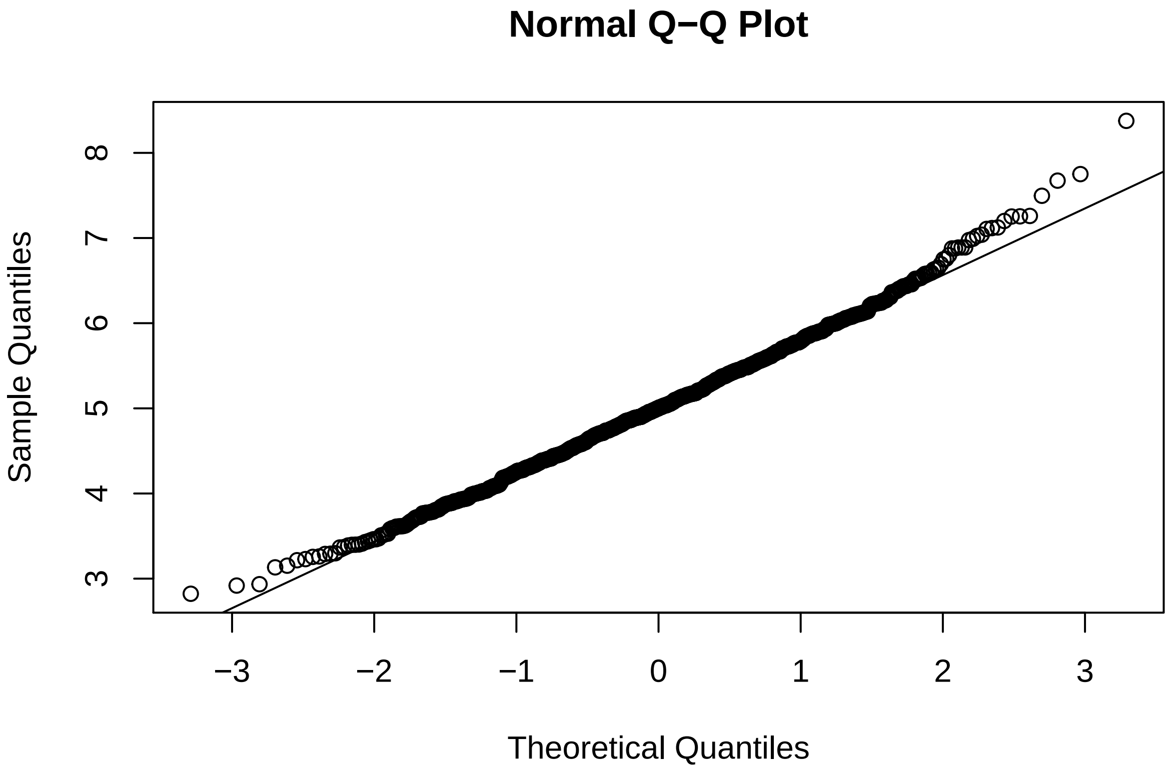
```
means <- NULL

for(i in 1:simulations){
    x <- rexp(n, rate = lambda)
    means <- (c(means, mean(x)))
}
title <- sprintf("Distribution of Sample Means for %g simulations \n using %g exponential values with la
hist(means, breaks = 20, main = title)
```



The distribution looks gaussian and it gets more gaussian as the sample size increases. This demonstrate in an empirical way the CLT. Also qqplot are used to compare sample distributions with a normal distribution. If the qqplot draws a straight line it can be said that the distribution of the sample is normal.

```
qqnorm(means)
qqline(means)
```



The normal Q-Q Plot also shows that the distribution of the samples means has a normal distribution. (CLT)