Cousera: Statistical Inference

Project 1: simulation

March 2015. Marcos Gestal.

Overview

This project investigates 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. The project will investigate the distribution of averages of 40 exponentials and illustrates via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials.

Simulations

First of all, data for simulation should be generated. The simulation data will be stored in a matrix of dimensions (nSimulations x nDistributions), so each row correspond with an individual simulation, that is, nDistribution numbers from an exponential function that after will be averaged to check the normality of the distribution of the mean of n exponential distributions.

```
nDistributions <- 40;
nSimulations <- 5000;
lambda <- 0.2;

set.seed(1234)

simulationValues <- rexp( nSimulations * nDistributions, lambda )
simulationMatrix <- matrix(simulationValues, nSimulations, nDistributions)
```

Sample vs. theoretical mean of the distribution.

The theoretical mean is

$$\mu = \frac{1}{lambda}$$

and the sample mean is a random variable that should be centered around the same value.

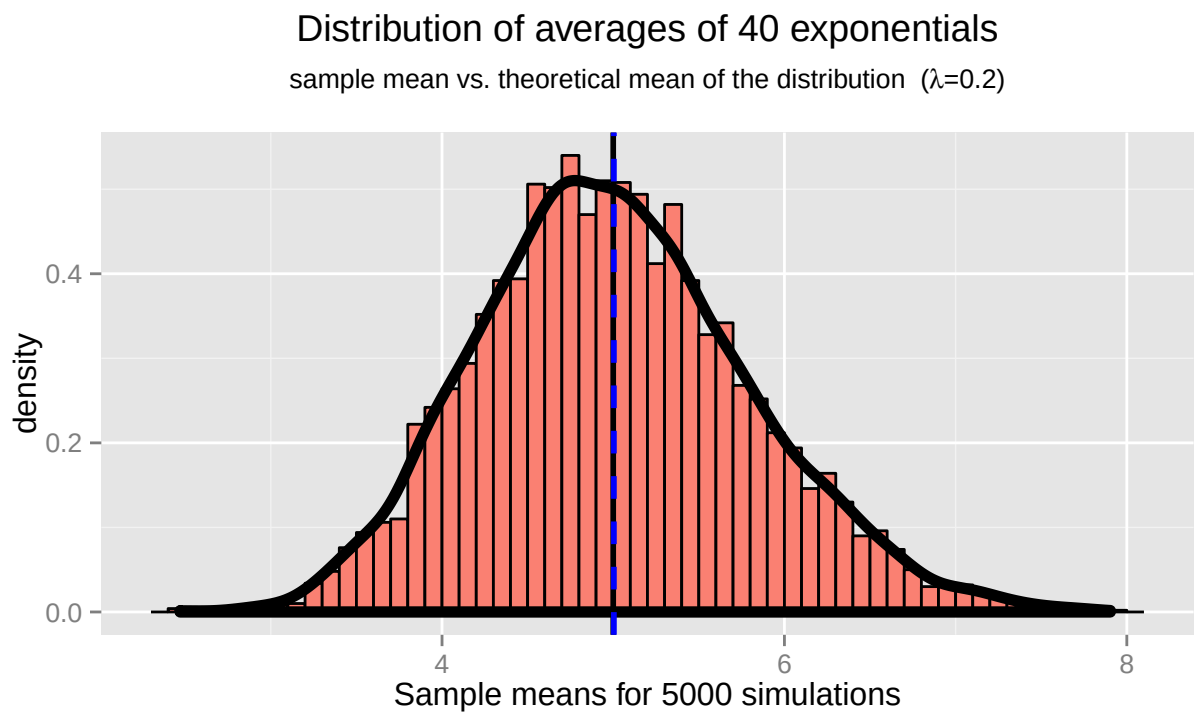
```
theoreticalMean <- 1/lambda

# Simulation mean: each position is the mean of nDistribution random values
#                  (from an exponential distribution)
#    dimensions: vector of nSimulatinons length
simulationMean <- rowMeans(simulationMatrix);

sampleMean <- mean(simulationMean)
```

```
## [1] "Theoretical mean: 5.000000"
```

```
## [1] "Sample mean: 5.003888"
```



As shown by the mean values and the previous graph the sample mean is very close to the theoretical mean. The values will be closer if we increase the number of simulations. Note how the sample mean is a random variable with a normal distribution around the population mean.

Sample vs. theoretical variance of the distribution

Standard deviation for an exponencial distribution is

$$S = \frac{1}{\lambda}$$

and talks about how variable the population is. So, the variance of the sample mean should be estimated with

$$\frac{S^2}{n}$$

where n is the number of distributions. This variance will show us how variable averages of random samples of size n from the population are.

```
theoreticalVariance <- ((1/lambda)^2)/nDistributions  # Theoretical Variance
sampleVariance<- var(simulationMean)
```

```
## [1] "Sample variance: 0.625956"
```

```
## [1] "Theoretical variance: 0.625000"
```

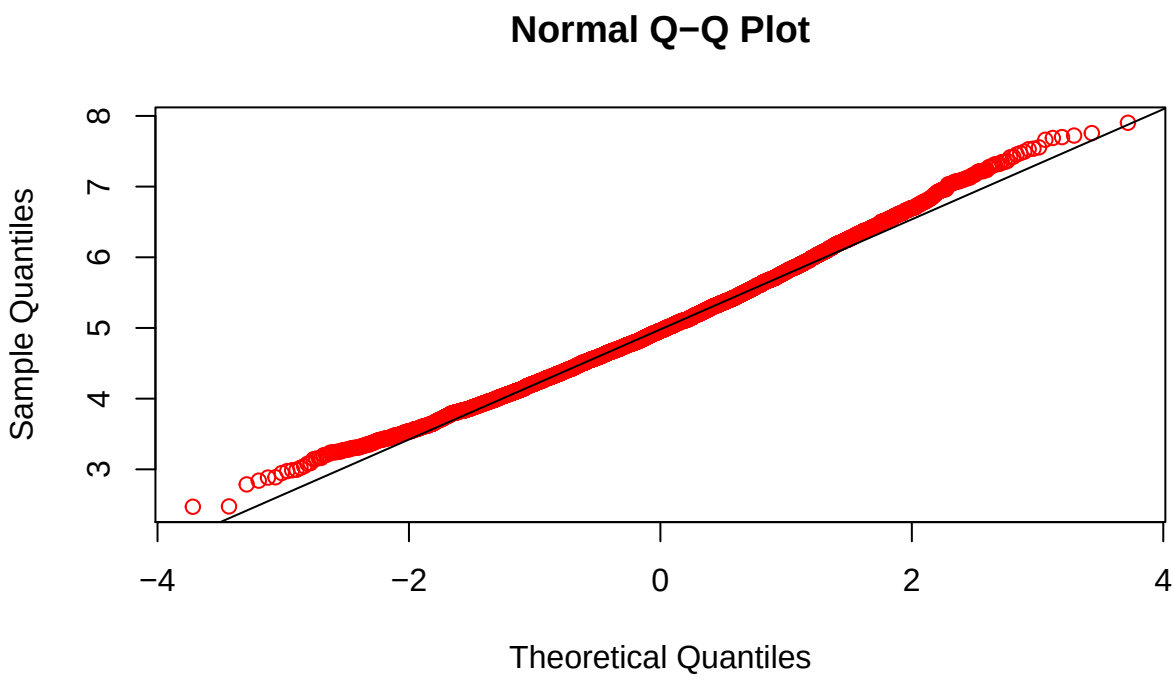
As in the previous point, the sample variance and the population variance are very similar. Note as if we increase the number of distributions, we would reduce the variance of the distribution of averages.

Comparing the distribution with a normal

At this point should be noted that the previous figure with the distribution of average is very similar to a Gaussian distribution (it is the distribution of the mean of 40 exponentials), compared with the distribution of an individual exponential distribution (see Appendix 2)

To check if a distribution is similar to a normal distribution, Q-Q plot can be used. In this kind of plot, each point represent the sample values vs the theoretical value, so if the distribution is normal all the points should be located over a 45 degrees line.

```
qqnorm(simulationMean, col="red")
qqline(simulationMean)
```



Following the previous graph, we can conclude that the distribution of averages of 40 exponentials is close to a normal distribution.

Conclusion

In probability theory, the Central Limit Theorem states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution (an exponential distribution in this experiment)

The performed simulation confirms this hypothesis because the distribution of the mean of 40 exponentials is really close to a normal distribution.

Appendix 1 - R code for figure 1

```
library(ggplot2)

g <- ggplot(data = data.frame(simulationMean), aes(x=simulationMean))

g <- g + geom_histogram(fill = "salmon", binwidth=0.1,
                        aes(y=..density..), color="black")
g <- g + geom_density(size = 2)

g <- g + geom_vline(xintercept = theoreticalMean, size = 1, color = "black",
                    linetype=1)

g <- g + geom_vline(xintercept = sampleMean, size = 1, color = "blue",
                    linetype=2)

g <- g + ggtitle(bquote(atop(paste("Distribution of averages of ", .(nDistributions),
                                   " exponentials"),
                             atop(paste("sample mean vs. theoretical mean of the distribution (",
                                   lambda, "=", .(lambda), ")")))))

g <- g + xlab(paste("Sample means for", nSimulations, "simulations"))
```

Appendix 2 - Exponential Distribution