The Exponential Distribution in R *versus* the Central Limit Theorem (CLT) — Part 1

Assignment: Statistical Inference Course Project

Simulation Exercise

Author: calaca

Course available at Coursera

Overview

This project consists in using simulation to explore inference and do some simple inferential data analysis. The project is divided into 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.

Simulations

1. Show the sample mean and compare it to the theoretical mean of the distribution

```r
# Seed
set.seed(2016)

# Lambda
lambda <- .2

# Sample
n <- 40

# Simulations
simulations <- 1000
```

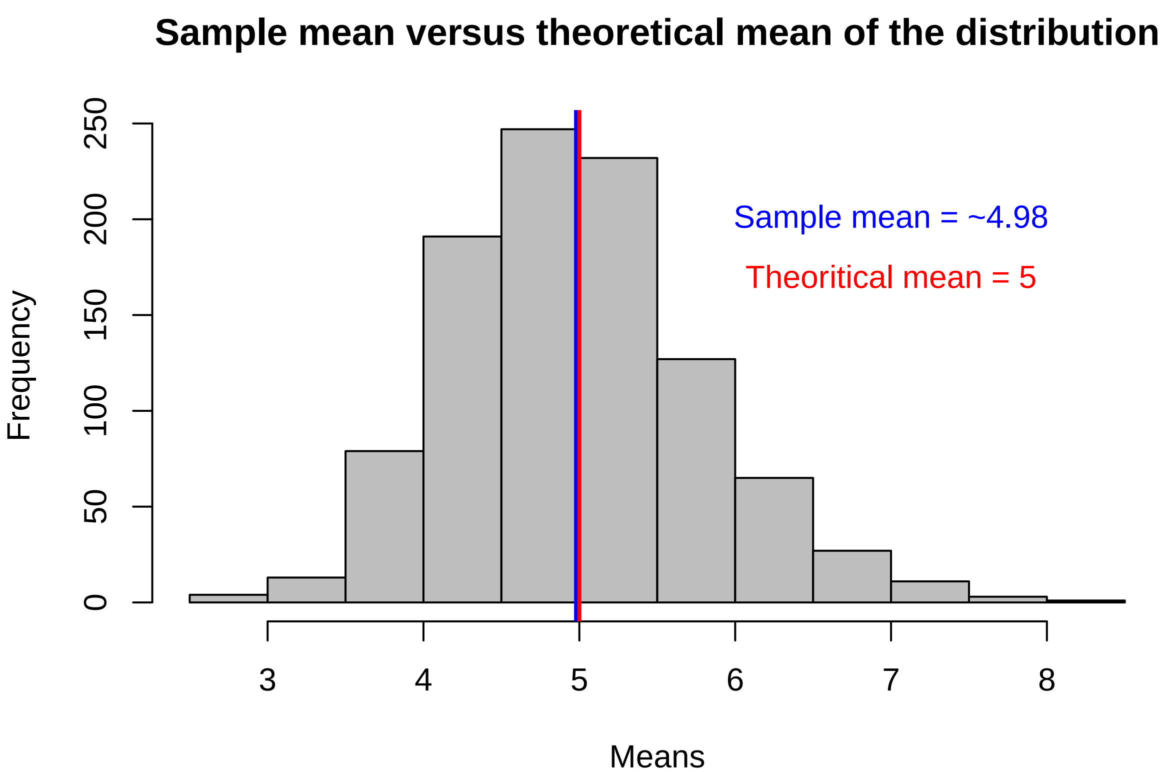
```r
# Replicating and simulating exponentials
simulated <- replicate(simulations, rexp(n, lambda))

# Mean of the simulated data
simulated_means <- apply(simulated, 2, mean)

# Mean of the simulated means
sample_mean <- mean(simulated_means)
sample_mean
```

```
## [1] 4.979186
```

```r
# Mean of the analytical expression
theoritical_mean <- 1/lambda
theoritical_mean
```

```
## [1] 5
```

```r
# Plot
hist(simulated_means, xlab = "Means", main = "Sample mean versus theoretical mean of the distribution",
abline(v = sample_mean, col = "blue", lwd = "2")
abline(v = theoritical_mean, col = "red", lwd = "2")
text(x = 7, y = 200, labels = "Sample mean = ~4.98", col = "blue")
text(x = 7, y = 170, labels = "Theoritical mean = 5", col = "red")
```



Observing the code and the plot we can see that, with a thousand simulations, the sample mean is ~4.98 and the theoritical mean of the distribution is 5.

2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution

```
# Standard Deviation of the sample means
sd_sample <- sd(simulated_means)
sd_sample
```

```
## [1] 0.7990522
```

```
# Theoritical Standard Deviation
sd_theoritical <- theoritical_mean/sqrt(n)
sd_theoritical
```

```
## [1] 0.7905694
```

```
# Variance of the sample means
var_sample <- sd_sample ^ 2
var_sample
```

```
## [1] 0.6384844
```

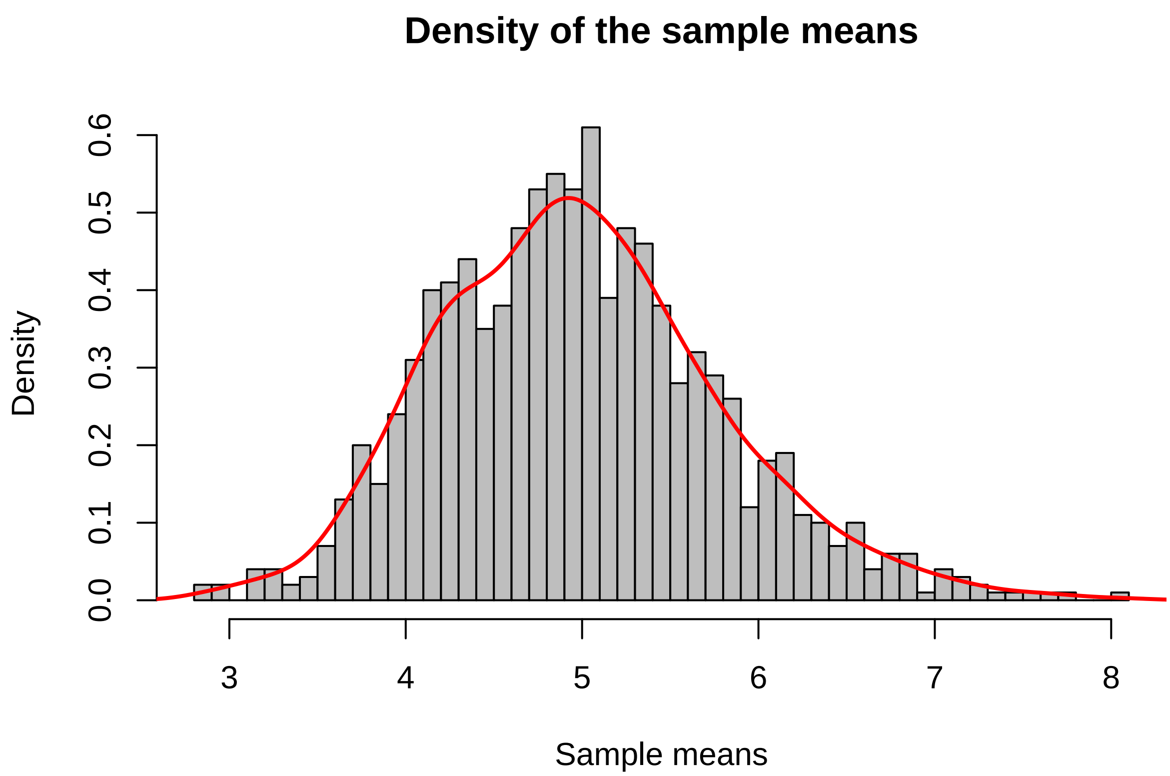
```
# Theoritical variance
var_theoritical <- sd_theoritical ^ 2
var_theoritical
```

```
## [1] 0.625
```

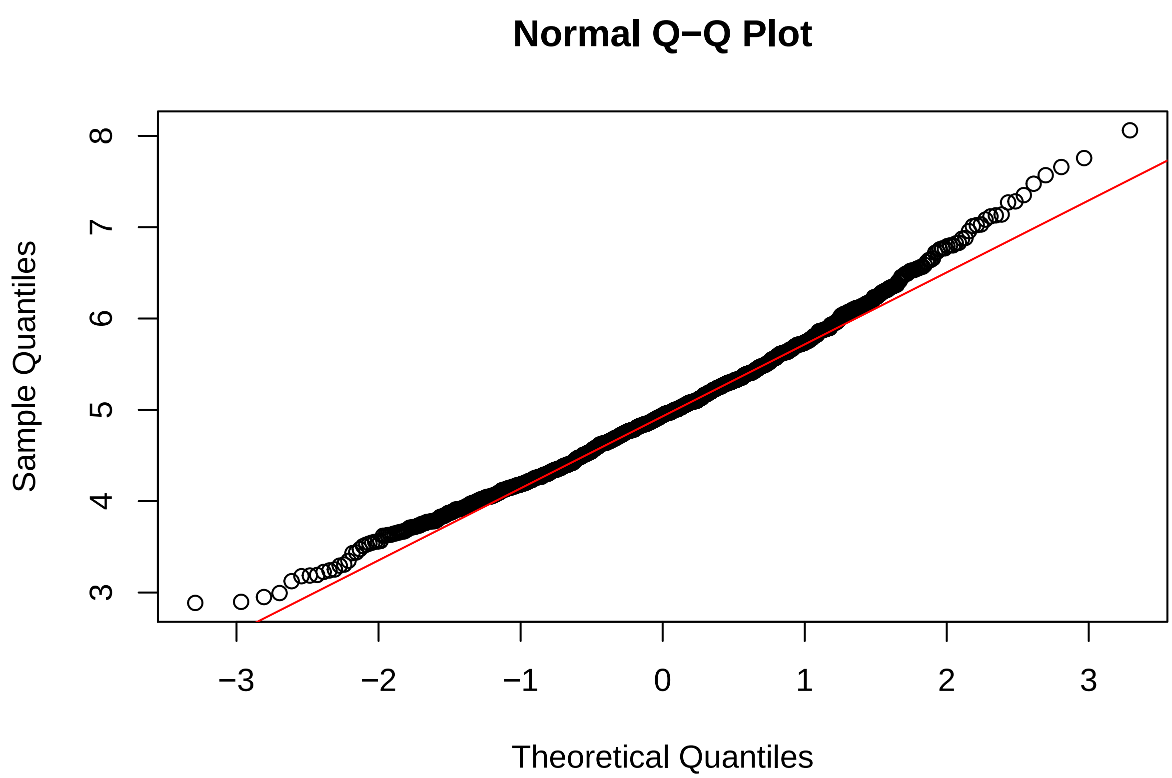
Therefore, it is possible to note that the sample variance (0.6384844) is very close to the theoritical variance (0.625).

3. Show that the distribution is approximately normal

```
hist(simulated_means, prob = TRUE, col = "grey", main = "Density of the sample means", ylab = "Density"
lines(density(simulated_means), lwd = 2, col = "red")
```




```r
qqnorm(simulated_means)
qqline(simulated_means, col = "red")
```



Observing both the density and the normal Q-Q plot, it is possible to see that the distribution of averages of

40 exponentials is close to the normal distribution. If we get more simulations, the density will be even closer to a bell-shaped form.