# Statistical Inference - Simulation Exercise

*Silva, RAFAEL*

*November 30, 2017*

## Setup

This section presents all the libraries used during the project.

```
library(ggplot2)
library(gridExtra)
```

## Introduction

In this project the exponential distribution is investigated in R and compared with the Central Limit Theorem.

It 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.

The objective of this part is to answer the following questions:

1. How does the sample mean compare to the theoretical mean of the distribution?
2. How variable the sample is (via variance) and compare it to the theoretical variance of the distribution?
3. Is the distribution approximately normal?

The lambda was set as 0.2 for all of the simulations. The distribution of averages of 40 exponentials will be investigated. To do that a thousand simulations are done in the chunk below.

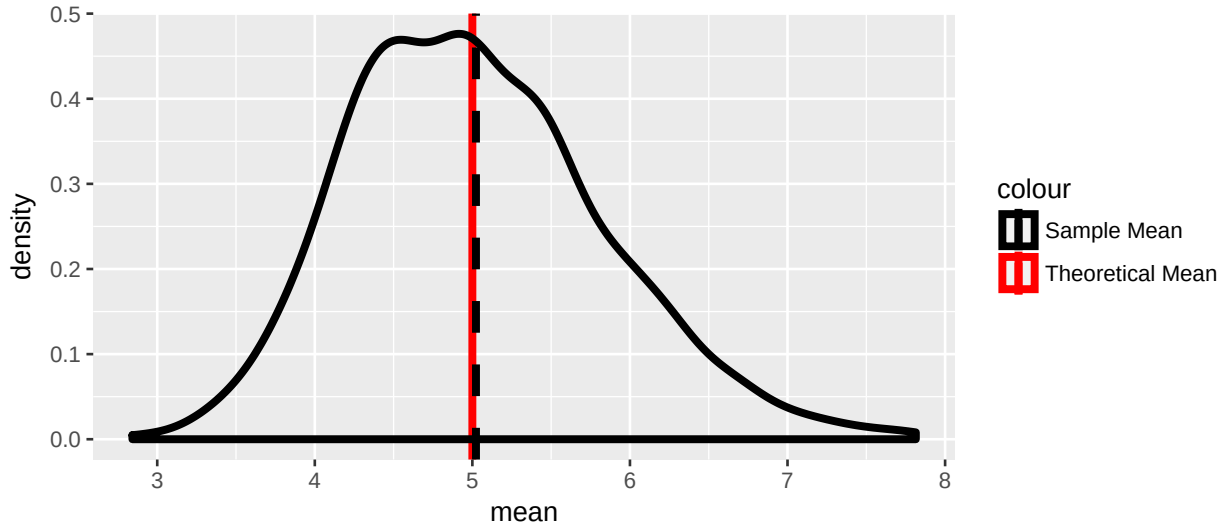
```
## Setting simulation parameters
lambda = 0.2            # Lambda
tmu = 1/lambda          # Theoretical Mean
ts = 1/lambda           # Theoretical Standard Deviation
tvar = ts^2             # Theoretical Variance
n = 40                  # Number of rolls per simulation
ns = 1000               # Number of simulations

## Simulating data
set.seed(123456)
samples <- matrix(rexp(ns*n, rate = lambda), nrow = ns, ncol = n)
```

## 1. Sample Mean

To answer the first question the mean of each simulation was calculated and the distribution of the results are compared to the theoretical mean in the resulting graph.

The code used to generate the plot is available in the appendix.



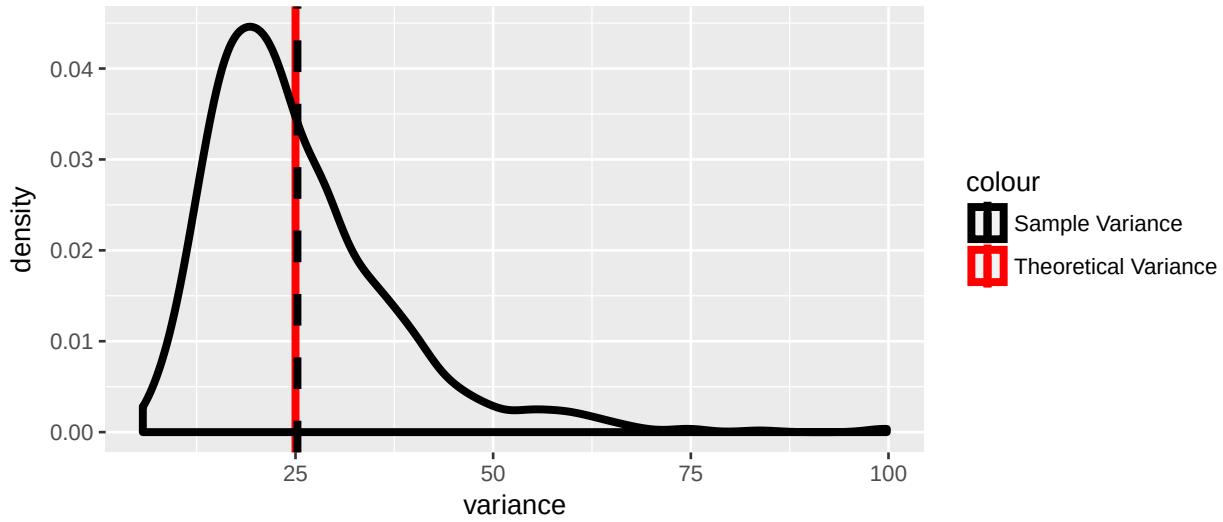

As the plot shows, that the center of the **distribution of sample means** is `5.02`, and is fairly close to the **theoretical mean** of the distribution which is `5`.

Another interesting fact is that the **standard error of the sample mean** and the **theoretical calculated standard error of the sample mean** for 40 samples are, respectively, `0.81` and `0.79`, which is an excellent approximation.

## 2. Sample Variance

For the second question, the same strategy is used, however, instead of extracting the mean of each simulation, the variance is take and its distribution is compared to the theoretical variance in the resulting graph.

The code used to generate the plot is available in the appendix.

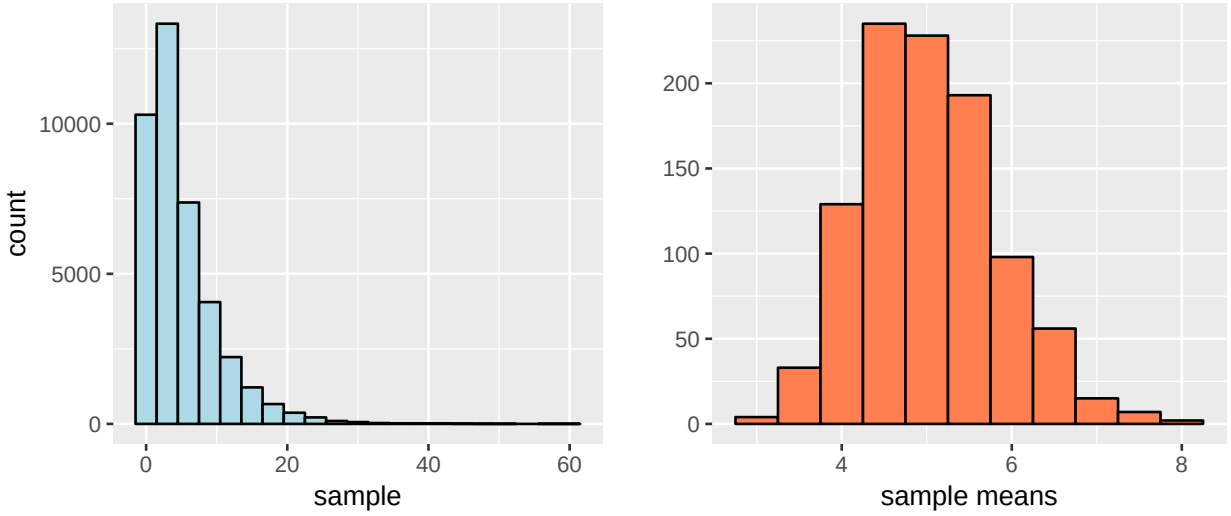



Once again, as the plot shows, the center of the **distribution of the sample variances** (`25.24`) lies very close to the **theoretical variance** (`25`).

## Distribution comparisson

Finally, the third question is answered by the double panel plot bellow.

The code used to generate the plot is available in the appendix.




The first plot (right) shows that the distribution of the `40,000` samples are **exponential**, regardless of that, the plot on the right shows that the distribution of `1,000` means of samples of size `40` is **approximately normally distributed**. This reinforces the central limit theorem that states that the sample mean is approximately normally distributed.

## Conclusion

This exercise studies and confirms some of the aspects of the CLT.

By the answer of the first question, it is clear that the distribution of the sample mean is approximately normal, and its mean approximate the population mean, and its standard error is the standard deviation of the population divided by the square root of the number of samples.

In the second question, it was observed that the distribution of the sample variance is also approximately normal and centered at the population variance.

In the third question, it is possible to observe that, regardless of the population distribution, the distribution of the sample means will always be approximately normally distributed.

## Appendix

Below it is possible to see the code used to generate the plots, in order from the first to the third.

```
## Calculating sample means
smus <- apply(samples, 1, mean)

## Calculating the center of the distribution
smu <- mean(smus)

## Ploting
g1 <- ggplot(aes(x = smus), data = as.data.frame(smus)) +
      geom_vline(aes(xintercept = tmu, col = "Theoretical Mean"), size = 1.5) +
      geom_density(aes(col = "Sample Mean"), size = 1.5) +
      geom_vline(aes(xintercept = smu, col = "Sample Mean"),
                 linetype = 2, size = 1.5) +
      scale_color_manual(labels = c("Sample Mean", "Theoretical Mean"),
                         values = c("black", "red")) +
      xlab("mean")

print(g1)
```

```
## Calculating sample variances
svars <- apply(samples, 1, var)

## Calculating the center of the distribution
svar <- mean(svars)

## Ploting
g2 <- ggplot(aes(x = svars), data = as.data.frame(svars)) +
      geom_vline(aes(xintercept = tvar, col = "Theoretical Variance"), size = 1.5) +
      geom_density(aes(col = "Sample Variance"), size = 1.5) +
      geom_vline(aes(xintercept = svar, col = "Sample Variance"),
                 linetype = 2, size = 1.5) +
      scale_color_manual(labels = c("Sample Variance", "Theoretical Variance"),
                         values = c("black", "red")) +
      xlab("variance")

print(g2)
```

```
## Linearizing samples
samp1 <- as.vector(samples)

## Ploting
g3 <- ggplot(aes(x = samp1), data = as.data.frame(samp1)) +
      geom_histogram(col = "black", fill = "lightblue", binwidth = 3) +
      xlab("sample")
g4 <- ggplot(aes(x = smus), data = as.data.frame(smus)) +
      geom_histogram(col = "black", fill = "coral", binwidth = 0.5) +
      xlab("sample means") + ylab("")
grid.arrange(g3, g4, ncol = 2)
```