Using the Central Limit Theorem to study some properties in the exponential distribution

Juan Sebastián Beleño Díaz

Overview

This project is an investigation about the exponential distribution in R and its comparation 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. For experimental purpouses it'll be set lambda = 0.2 for all of the simulations and this investigation will be focused on the distribution of averages of 40 exponentials. Note that it's necessary to do a thousand simulations.

Simulations

First step to start the project is adding the libraries that will be used during the development of the project.

```
# Libraries needed
library(ggplot2)
```

In this point, It's ready to start the heavy part of the project, so the simulation can be done using the following code:

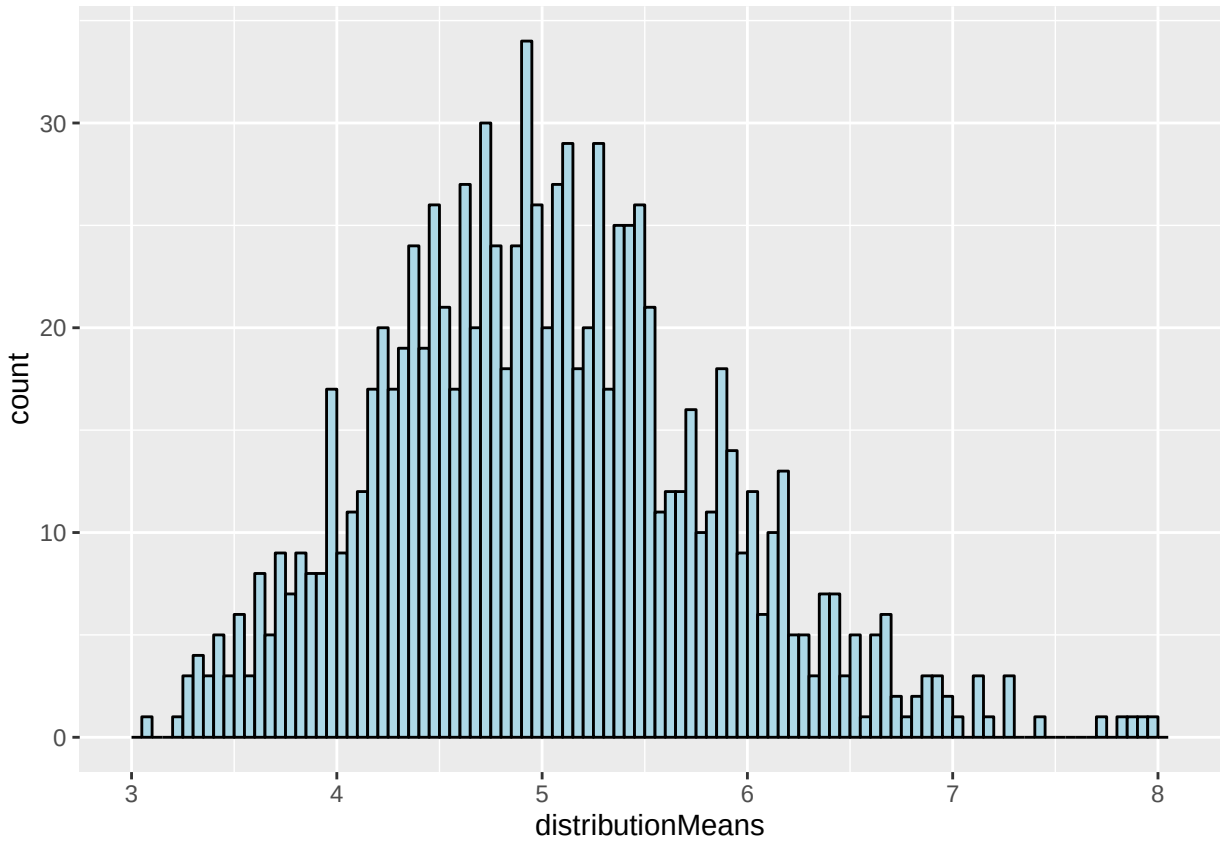
```
# Define constants
nosim <- 1000       # Number of simulations
n <- 40             # Sample size
lambda <- 0.2

# As guarantee of reproductibility its mandatory to set a seed
set.seed(19920328)

# The simulation similar to bootstrap (Boom Goes the Dynamite)
distribution <- matrix(rexp(n * nosim, lambda) , nosim, n)
distributionMeans <- apply(distribution, 1, mean)
```

To see the result of the simulation is necessary to plot a graphic with them.

```
g <- ggplot(
            data.frame(mean = distributionMeans),
            aes(x = distributionMeans)
         )
g <- g + geom_histogram(color = "black", fill = "lightblue", binwidth = 0.05)
g
```



Sample Mean versus Theoretical Mean

It's shown the differences between the theoretical mean and the sample mean

```
# Calculate theoretical mean and sample mean
tmean <- 1/lambda
smean <- mean(distributionMeans)
c(tmean, smean)
```

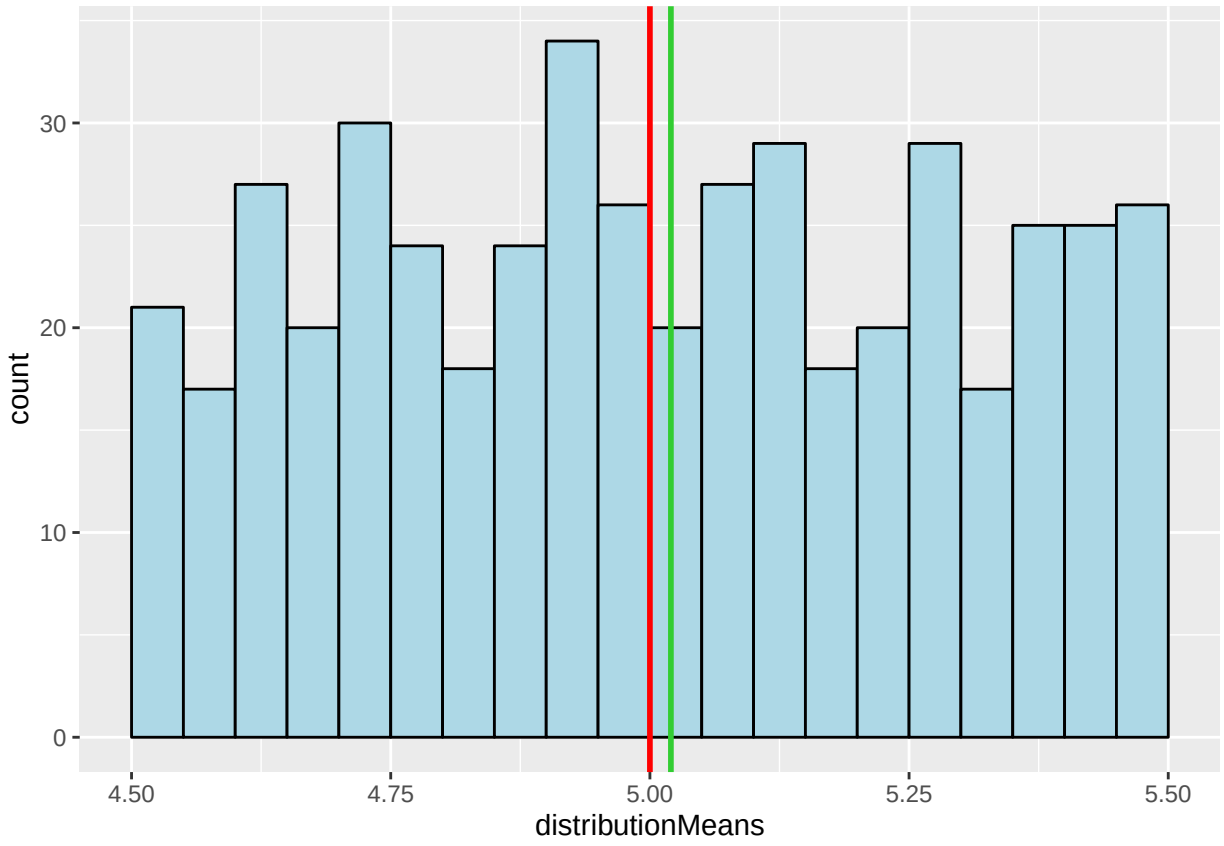
```
## [1] 5.000000 5.020261
```

To see better the difference is necessary to plot it, the red line represents the theoretical mean and the green one represents the sample mean, as we can see they are pretty close and for that it's applied a zoom over the original plot.

```
g <- ggplot(
            data.frame(mean = distributionMeans),
            aes(x = distributionMeans)
        )
g <- g + geom_histogram(color = "black", fill = "lightblue", binwidth = 0.05)
g <- g + geom_vline(xintercept=tmean, size = 1, color = "red")
g <- g + geom_vline(xintercept=smean, size = 1, color = "limegreen")
g <- g + xlim(4.5, 5.5)
g
```

```
## Warning: Removed 523 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Sample Variance versus Theoretical Variance

As we can see the variances also are pretty close one to the other.

```
# Calculate theoretical variance and sample variance
tvar <- (1/lambda)^2/n
svar <- var(distributionMeans)
c(tvar, svar)
```

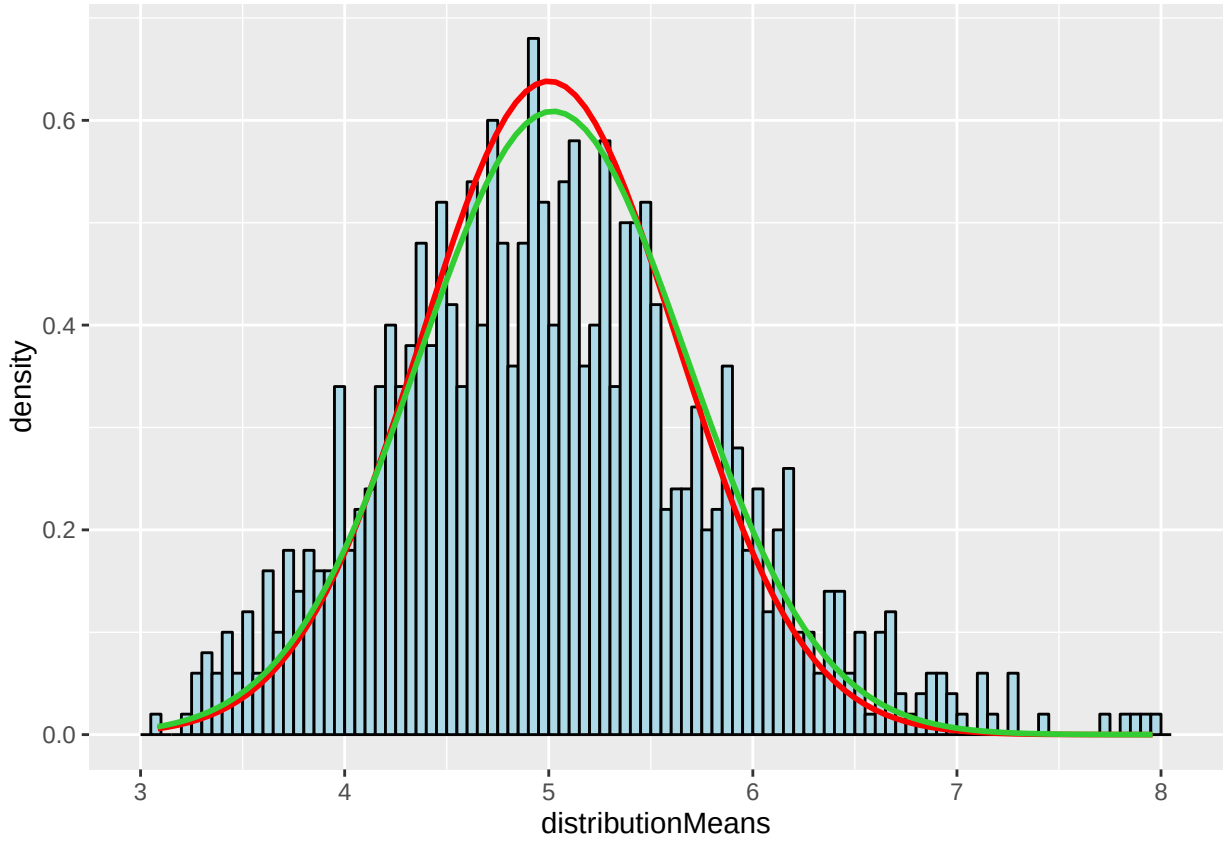
```
## [1] 0.6250000 0.6551741
```

Distribution

In this case it's compared the data shown in the histogram with a posible normal distribution, but before doing this, the histogram is normalized.

```
g <- ggplot(
            data.frame(mean = distributionMeans),
            aes(x = distributionMeans)
        )
```

```
g <- g + geom_histogram(
            color = "black", fill = "lightblue",
            binwidth = 0.05, aes(y=..density..)
            )
g <- g + stat_function(
            fun = dnorm, arg = list(mean = tmean , sd = tvar),
            colour = "red", size=1
            )
g <- g + stat_function(
            fun = dnorm, arg = list(mean = smean , sd = svar),
            colour = "limegreen", size=1
            )
g
```



The graphic above shows how it's the behaivor of the data compared with both possible normal distributions, one the theoretical and the other from the sample and as we can appreciate, our data match very well with a normal distribution.