Simulation Exercise - Statistical Inference Course

Pedro A. Alonso Baigorri

9 de octubre de 2017

Synopsis

In this project I will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution will 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.

I will set lambda = 0.2 for all of the simulations. I will investigate the distribution of averages of 40 exponentials.

To do this exercise I will run three set of simulations with sizes of 100, 1.000, and 10.000 to demonstrate how the main values of mean and variance is moving to the theoretical value when the number of simulations is increased.

Data preparation

First of all, I will set the main attributes and calculate the theoretical values for mean of variance.

In addition I will create some dataframes to store the results of the simulations.

```
# Load required libraries
library(ggplot2)

# setting the main attributes
lambda <- 0.2
n <- 40

# calculating the theorical mean and variance
th_mean <- 1/lambda
th_var <- (1/(lambda^2))/n

# creating a data frame to store the data of distribution, mean and variance for
# the different simulations
df_mean <- data.frame(th_mean, "Theoretical")
colnames(df_mean) <- c("mean", "size")

df_var <- data.frame(th_var, "Theoretical")
colnames(df_var) <- c("variance", "size")

df_dist <- data.frame(x = numeric(), size = character())
```

According to this, the theoretical value for the MEAN is 5 and for the VARIANCE is 0.625

Run of simulations

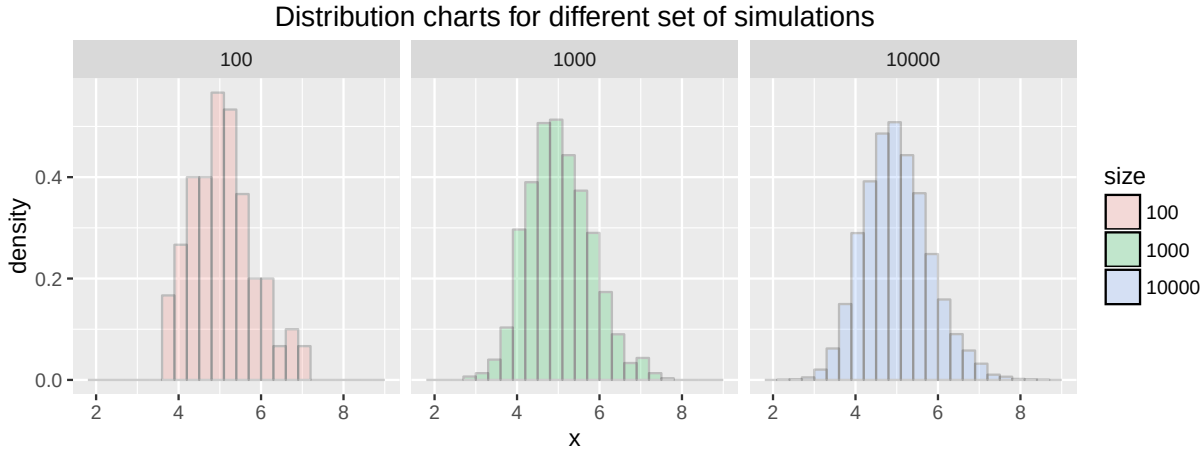
Now I will run the simulations for a size of 100, 1000 and 10.000 and I will store the results of each simulation in some dataframes in order to do a further analysis and plot the results.

```r
#
# loop for simulations of 100, 1000, 10000
#
simulations <- c(100, 1000, 10000)


for (j in simulations){

    mns = NULL

    for (i in 1 : j) mns = c(mns, mean(rexp(n, lambda)))

    # setting the data of distribution
    df_dist_t <- data.frame(mns, as.factor(j))
    colnames(df_dist_t) <- c("x", "size")
    df_dist <- rbind (df_dist, df_dist_t)

    #setting the data of means
    df_mean_t <- data.frame(mean(mns), as.factor(j))
    colnames(df_mean_t) <- c("mean", "size")
    df_mean <- rbind(df_mean, df_mean_t)

    #setting the data of variances
    df_var_t <- data.frame(var(mns), as.factor(j))
    colnames(df_var_t) <- c("variance", "size")
    df_var <- rbind(df_var, df_var_t)

}
```

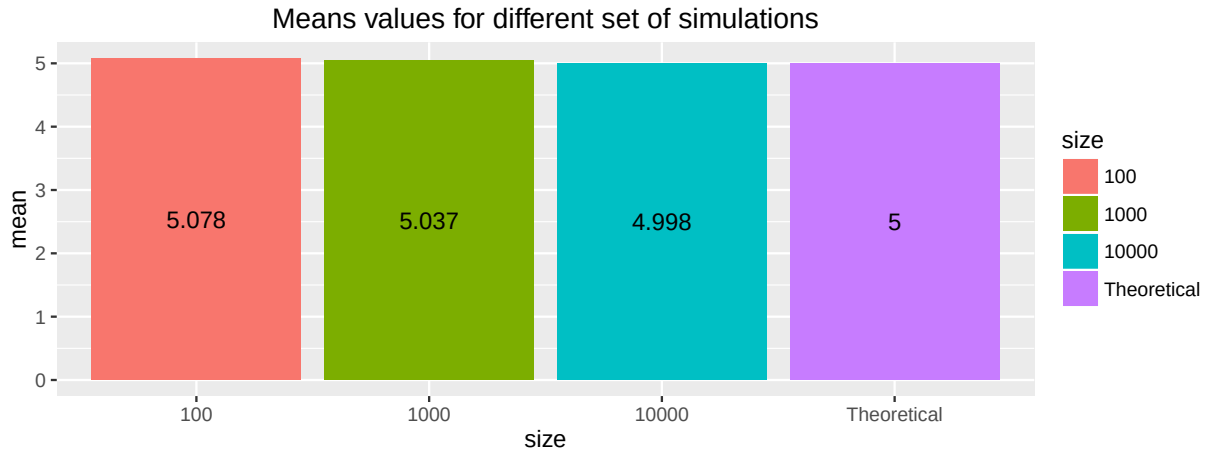
Results

The results obtained on these simulations can be showed in the following plots.

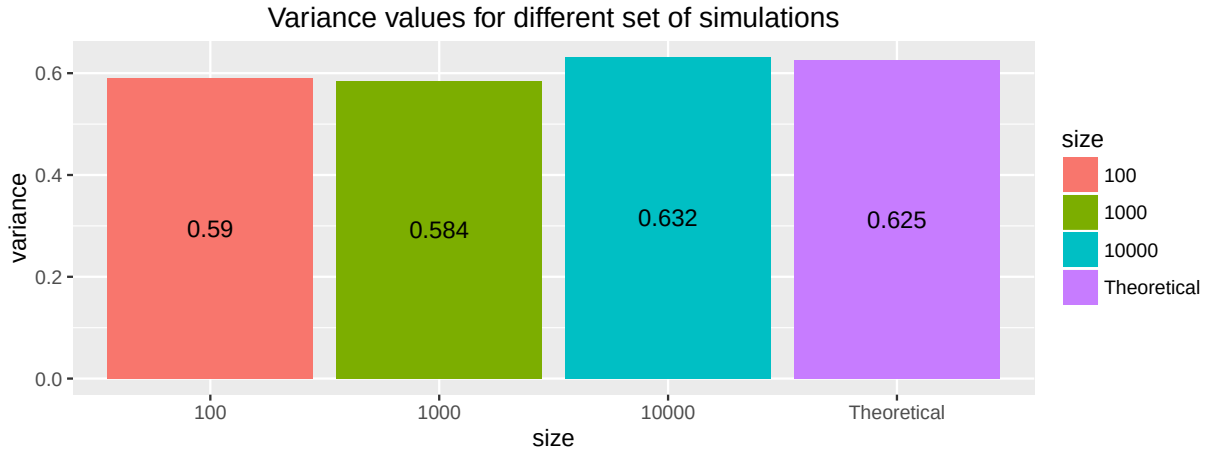
In the first plot we can see the distribution and how it's centered in the theoretical MEAN. And we can also see that the distribution is approaching to normal when the size of the simulation is increasing.



In the next plot we can see the mean of the simulations comparing with the theoretical mean and how this mean is approaching to the theoretical mean when the size of simulations is increasing.



And the last plot shows the same phenomenon with the variance of the simulations.



Conclusions

After this exercise we have demonstrated that the Central Limit Theorem is true for the Exponential distribution with the following main conclusions:

- The distribution of the simulation approaches to a normal distributions when the size of the simulations is getting large
- The mean of the simulations approaches to the Theoretical mean (5) when the size of the simulations is getting large
- The variance of the simulations approaches to the Theoretical variance (0.625) when the size of the simulations is getting large.

Annexes

R Code for plotting the figures

```
# plotting distribution charts
g <- ggplot(df_dist, aes(x = x, fill = size)) +
    geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..density..))

g + facet_grid(. ~ size) + ggtitle("Distribution charts for different set of simulations")
```

```
# plotting mean charts
df_mean$size <- factor(df_mean$size, levels = c("100", "1000", "10000", "Theoretical"))

g <- ggplot(df_mean, aes(size, mean)) + geom_bar(stat = "identity", aes(fill = size)) +
    ggtitle ("Means values for different set of simulations")

g +  geom_text(aes(label = round(mean, 3), y = mean - 0.5* mean))
```

```
# plotting variance charts
df_var$size <- factor(df_var$size, levels = c("100", "1000", "10000", "Theoretical"))

g <- ggplot(df_var, aes(size, variance)) + geom_bar(stat = "identity", aes(fill = size)) +
    ggtitle ("Variance values for different set of simulations")

g +  geom_text(aes(label = round(variance, 3), y = variance - 0.5* variance))
```