Statistical Inference Course Project - Part 1

Carolina San Martín

Part 1: Simulation exercise

Overview

This is the first of two-parts final project of the Statistical Inference Course from Coursera's Data Science by Johns Hopkins University. This first part aims to investigate the exponential distribution in R and compare it with the Central Limit Theorem: given that lambda = 0.2 for all of the simulations, this part illustrates the properties of the distribution of the mean of 40 exponentials over a thousand simulations.

I made a pdf report to answer the questions presented in the project rubric using 'knitr'.

Simulations

First, it is needed setting parameters to define the distribution:

```
# Set the seed of R's random number generator to start simulation
set.seed(10000)

# Lambda provided
lambda <- 0.2

# Number of exponentials
n <- 40

# Number of simulations
n_simulation <- 1000
```

Then, we simulate the 40 exponentials and estimate the mean of the simulated data.

```
# Simulate the exponentials
mns <- NULL
for (i in 1:n_simulation) mns <- c(mns, mean(rexp(n,lambda)))
```

1. Sample mean and theoretical mean of the distribution.

Calculate and compare the sample mean to the theoretical mean. The mean of the exponential distribution is equal to $1/\lambda$. In our case, λ is equal to 0.2. Thus, the actual mean is not as far of value from the theoretical mean, because values are 5.01 and `1/lambda`, respectively.

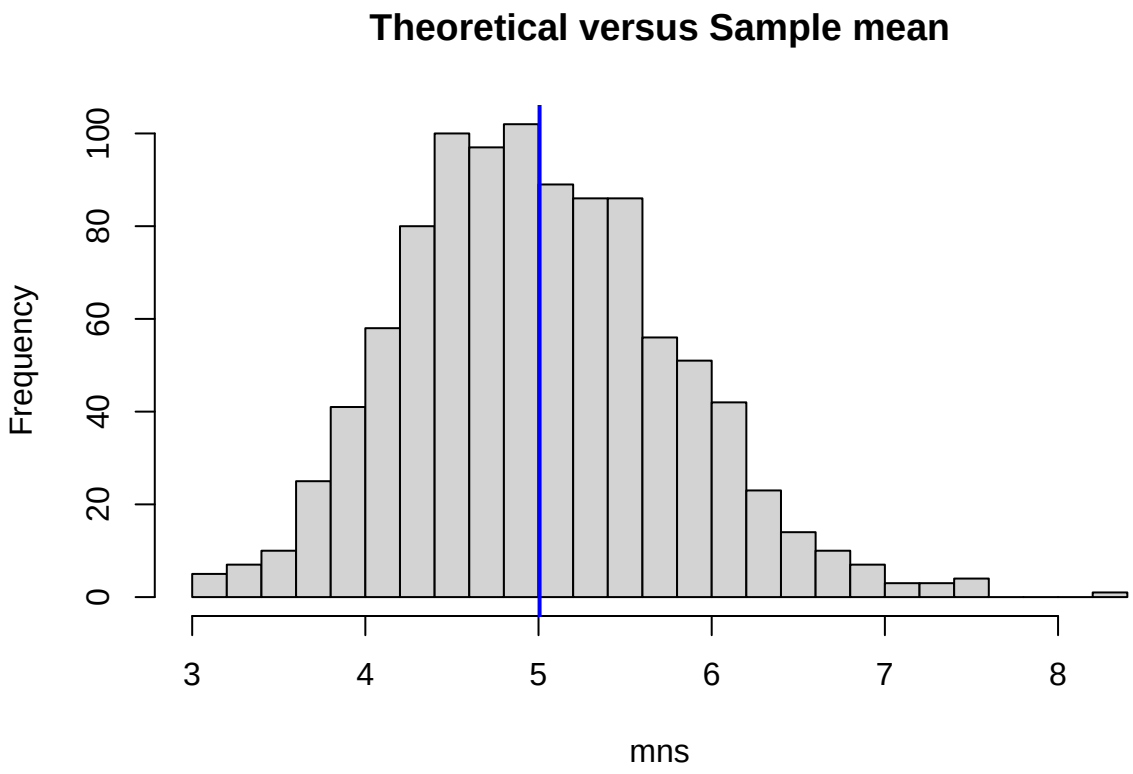
```
# Sample mean
mean(mns)
```

```
## [1] 5.00599
```

```
# Theoric mean of exponential distribution
1/lambda
```

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

```
# Creating the histogram
hist(mns,
     main="Theoretical versus Sample mean",
     breaks=30)
abline(v = mean(mns), lwd="2", col="blue")
```



2. Variance of the distribution.

The following code chunk relates the theoretrical standard deviation to the standard deviation value calculated from the simulations.

```
(1/lambda)/sqrt(n)
```

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

```
sd(mns)
```

```
## [1] 0.7832666
```

Now, the same for variance

```
((1/lambda)/sqrt(n))^2
```

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

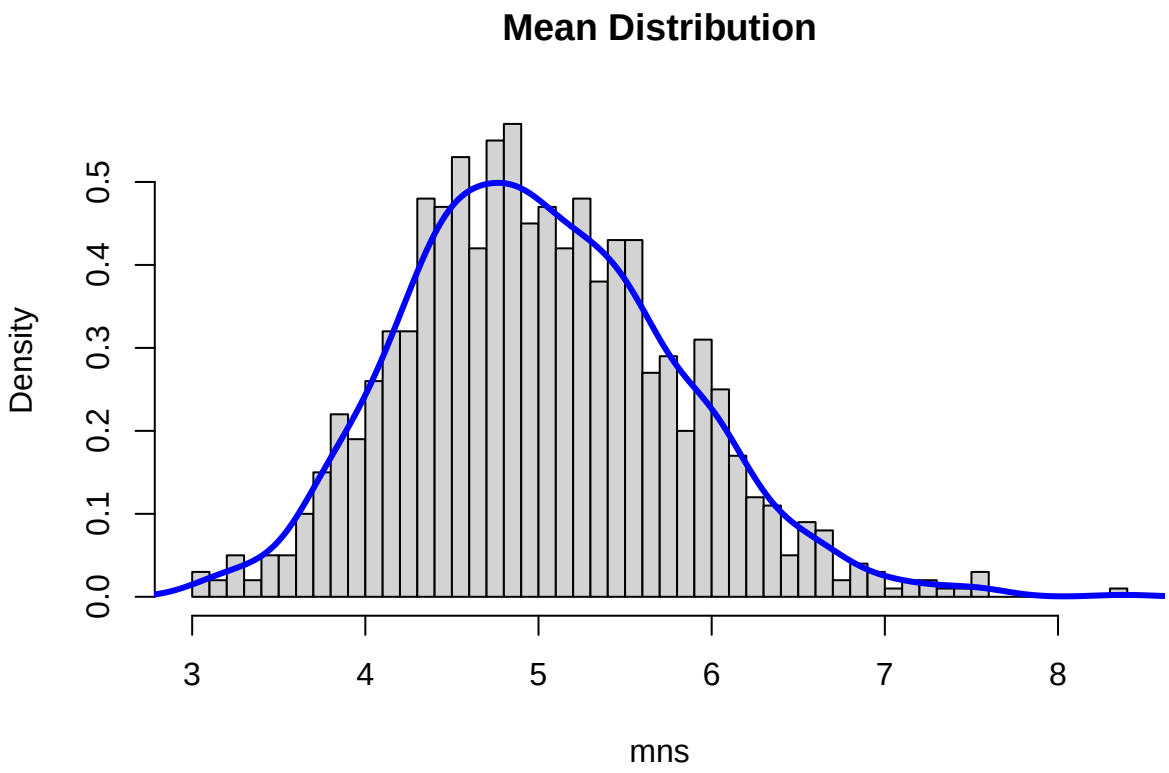
```
sd(mns)^2
```

```
## [1] 0.6135066
```

3. Distribution.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
#Show that the distribution is approximately normal.
#need that prob = TRUE to turn into density
hist(mns, prob = TRUE, main = "Mean Distribution", breaks = 50)
lines(density(mns), lwd = 3, col = "blue")
```



This histogram has been adapted to show how this simulation tracks normality. This related to the CLT and how more samples will make this data appear more normal.