Statistical Inference

Part 1: Central Limit Theorem review

MATIAS, V. A.

05/04/2021

Overview

The first part of the project makes one thousand simulations of the exponential function to analyzes the Central Limit Theorem application under the means distribution.

Central Limit Theorem

The Central Limit Theorem (CLT) says that the averages of any distribution tend to the normal as the amount of data grows. The first topic examines this fact.

Simulation

We'll create a thousand exponential distributions and extract the means to build a new distribution that should be next to the normal. The probability density function (pdf) of the exponential distribution is calculated as:

$f(x) = \lambda e^{-\lambda x}$

Where:

- λ = average rate parameter;
- x = Random variable (numeric).

In this report will be created 1000 of these distributions. Each of them has 40 random variables, and lambda equals 0.2, so the formula can be adapted to $f(x) = 0.2e^{-0.2*x}$. The plot below is generated by extracting the averages for each of the exponential distributions.

Given the random data, all simulations will result in curves with little differences. However, every simulation will return an approximately normal distribution.

Sample statistics X Theoretical statistics

The plot above shows the sample mean and the theoretical mean. You can see that it differs, but both are very close. Sample variance and theoretical variance follow the same idea, look at the simulation results for means and variances.

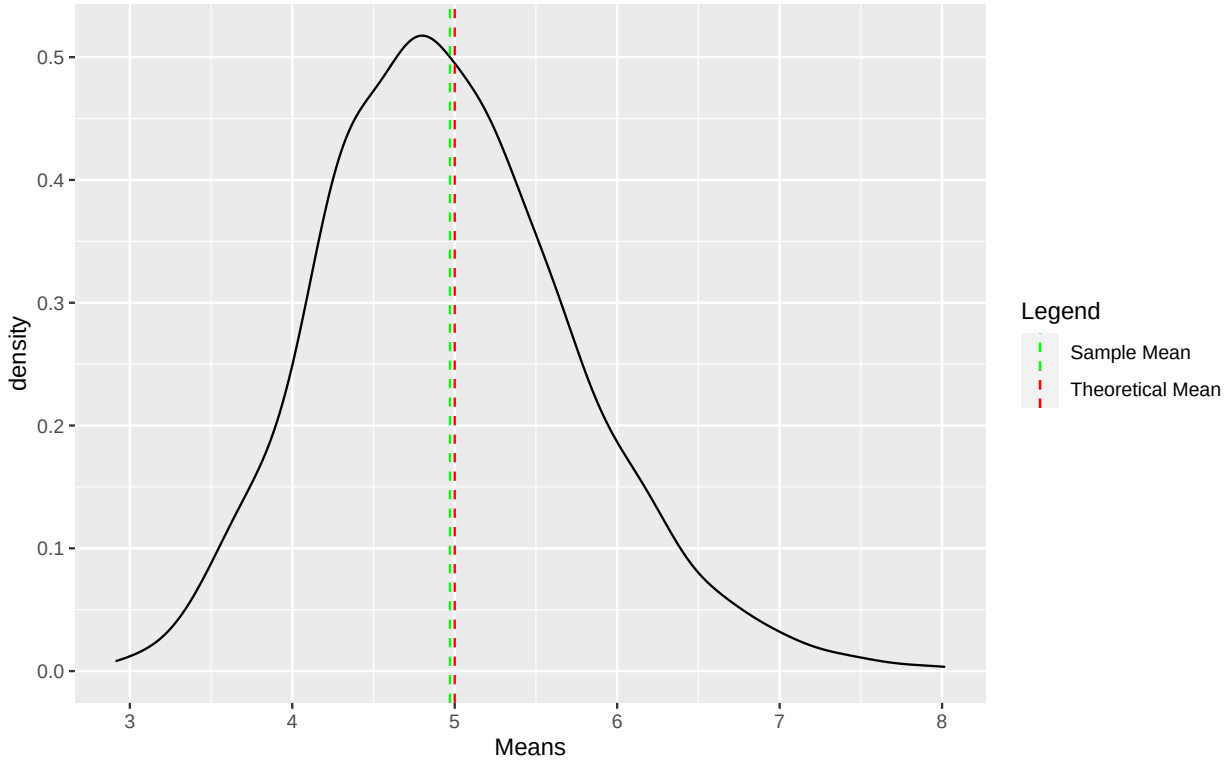

Figure 1: Averages PDF

```
## Theoretical Mean: 5
## Sample Mean: 4.97066687956162
## Theoretical Variance: 0.625
## Sample Variance: 0.618734330549701
```

The theoretical mean (in other words, the expected value) for exponential distribution is calculated as $E(X) = 1/\lambda$ and the variance as $Var(X) = (1/\lambda^2)/n$, so, this results will be constant for the same values of λ and n.

For one thousand simulations, the sample mean will be very close to the theoretical mean of 5 (that is, 1/0.2), with a difference close to 2 decimal places for most of the simulations made. The sample variance should be next to the theoretical too. This makes sense because we run a lot of simulations, and this can reduce the relevance of many outliers.

Distribution

To understand how the distribution of the averages is close to the normal, look at the plot below.

The line that follows the histogram is the normal for the theoretical mean and standard deviation. Even with variations of the original curve, the data follow a normal distribution because they have a high concentration of averages around the average of the distribution, while divergent values become less likely as standard deviations from the mean increase.

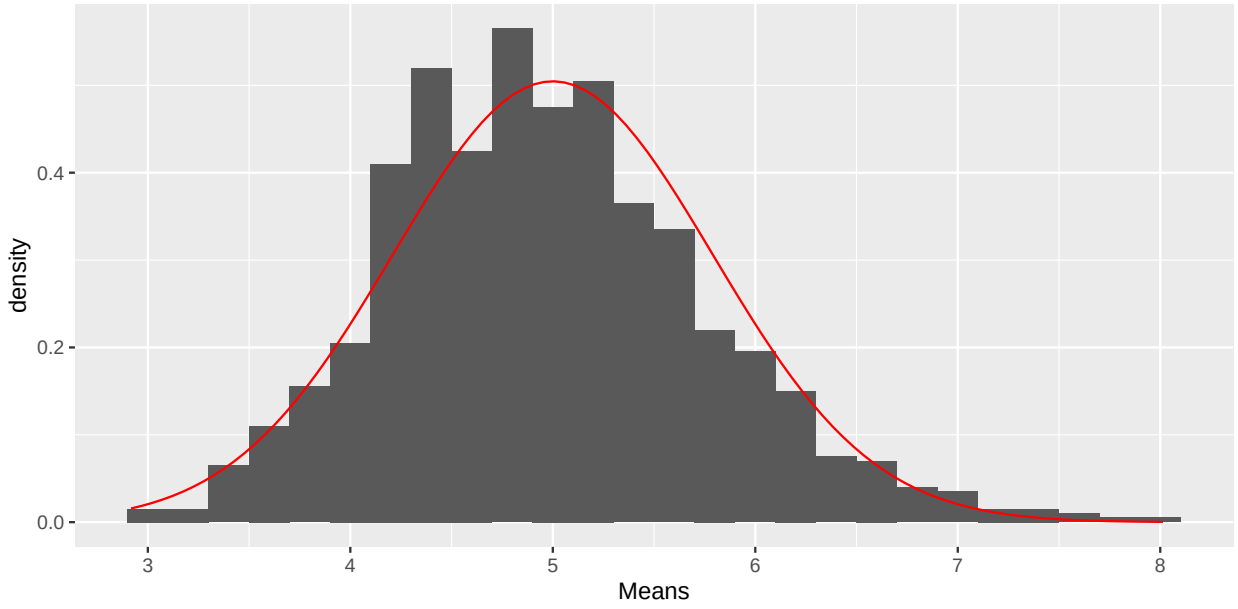

Figure 2: Histogram of the data and the normal curve of the theoretical distribution

To clarify the change, the third plot shows the comparison between the distribution of the means and one of the exponential function results (n = 40, lambda = 0.2)

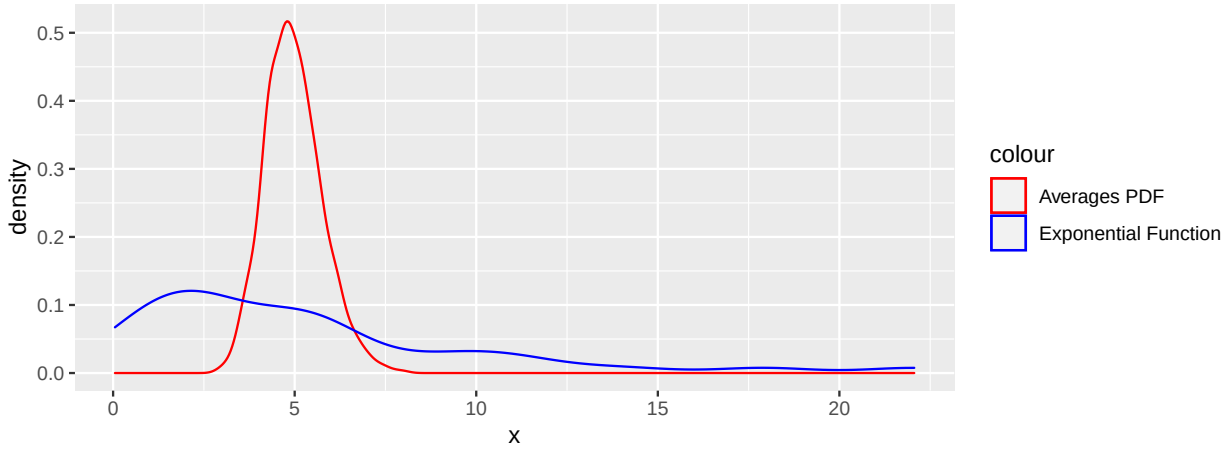

Figure 3: Comparison between the distribution of means and one distribution of the exponential function

As the exponential distribution has lambda 0.2, the average of this distribution must be close to 5, which is notable for the graph. However, the values vary for different values of x, causing a decrease/increase as new simulations are carried out. Making many distributions and collecting their averages, therefore, it is visible that the lower / higher values will be less frequent, causing a concentration of averages, but liable to more extreme values with a lower probability of occurrence.

Apendix

```
#Figure 1 code
library(ggplot2)

n <- 40
lambda <- 0.2

mns <- c()
for (i in seq(1000)) {
  mns <- c(mns, mean(rexp(n, lambda)))
}

pdf <- ggplot() +
  geom_density(aes(x = mns)) +
  xlab("Means") +
  geom_vline(aes(xintercept = mean(mns), color = "Sample Mean"),
             linetype = "dashed") +
  geom_vline(aes(xintercept = 1/lambda, color = "Theoretical Mean"),
             linetype = "dashed") +
  scale_color_manual(name = "Legend",
                     values = c("Sample Mean" = "green", "Theoretical Mean" = "red"))


pdf
```

```
#means and variances code
theor_mean <- 1/lambda
smp_mean <- mean(mns)
theor_var <- (1/lambda^2)/n
smp_var <- var(mns)

message("Theoretical Mean: ", theor_mean,
        "\nSample Mean: ", smp_mean,
        "\nTheoretical Variance: ", theor_var,
        "\nSample Variance: ", smp_var)
```

```
#Figure 2 code
ggplot() +
  geom_histogram(aes(mns, ..density..),
                 binwidth = lambda) +
  xlab("Means") +
  stat_function(fun=dnorm,
                color="red",
                args = list(mean = theor_mean,
                            sd = sqrt(theor_var)))
```

```
#Figure 3 code
exp_results <- rexp(n, lambda)

ggplot() + geom_density(aes(mns, color = 'Averages PDF')) +
  geom_density(aes(exp_results,color = 'Exponential Function')) +
```

```
  scale_color_manual(values = c('Averages PDF' = 'red',
                                'Exponential Function' = 'blue')) +
xlab("x")
```