Simulation exercise : the exponential distribution

Alnour Ribault

3 février 2018

Overview

This document is my submission for the first part of the final assignment of the Statistical Inference class in the Data Science specialization by the John Hopkins University on Coursera.
The aim is to illustrate the Central Limit Theorem applied to the exponential distribution.

Simulations

First, we declare the simulations parameters. `lambda` is the parameter we will use for the exponential distribution, `n` is the sample size and `B` is the number of samples.
We also set an arbitrary seed so that the content of this document, which includes random number generation, is reproducible.

```
lambda <- 0.2
n <- 40
B <- 1000
set.seed(238235493)
```

We then use the `rexp` function to generate `n*B` exponentials which we store in a matrix with `B` rows and `n` columns, so that each row is a sample of `n` exponentials: those are our simulations.

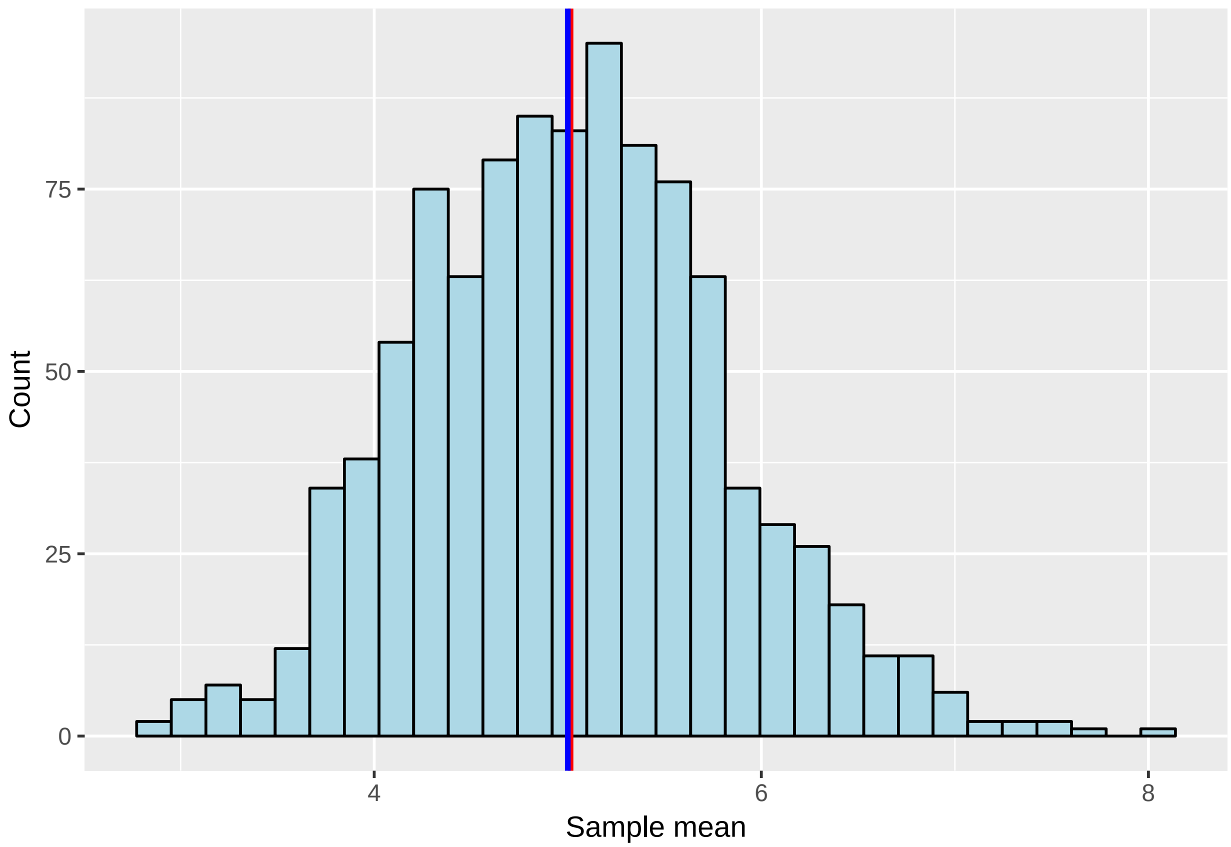
```
sims <- matrix(rexp(n*B, lambda), nrow = B, ncol = n)
```

Sample Mean versus Theoretical Mean

We compute the mean of each sample by using the function `apply` along the rows - that's why one argument is equal to `1`. We store the results in the variable `means`. We also compute both the mean of the sample mean, which is the mean of the `means` vector, and the theoretical means, which is equal to `1/lambda` - because the sample mean is an *unbiased estimator* of the population mean.

```
means <- apply(sims, 1, mean)
sampleMeanMean <- mean(means)
theoreticalMean <- 1/lambda
```

The mean of the sample mean is equal to 5.0142586 while the theoretical mean is 5 : they are indeed very close! We plot below a histogram of the sample mean. The red vertical bar represents the mean of the sample mean, and the blue vertical bar represents the theoretical mean.



Sample Variance versus Theoretical Variance

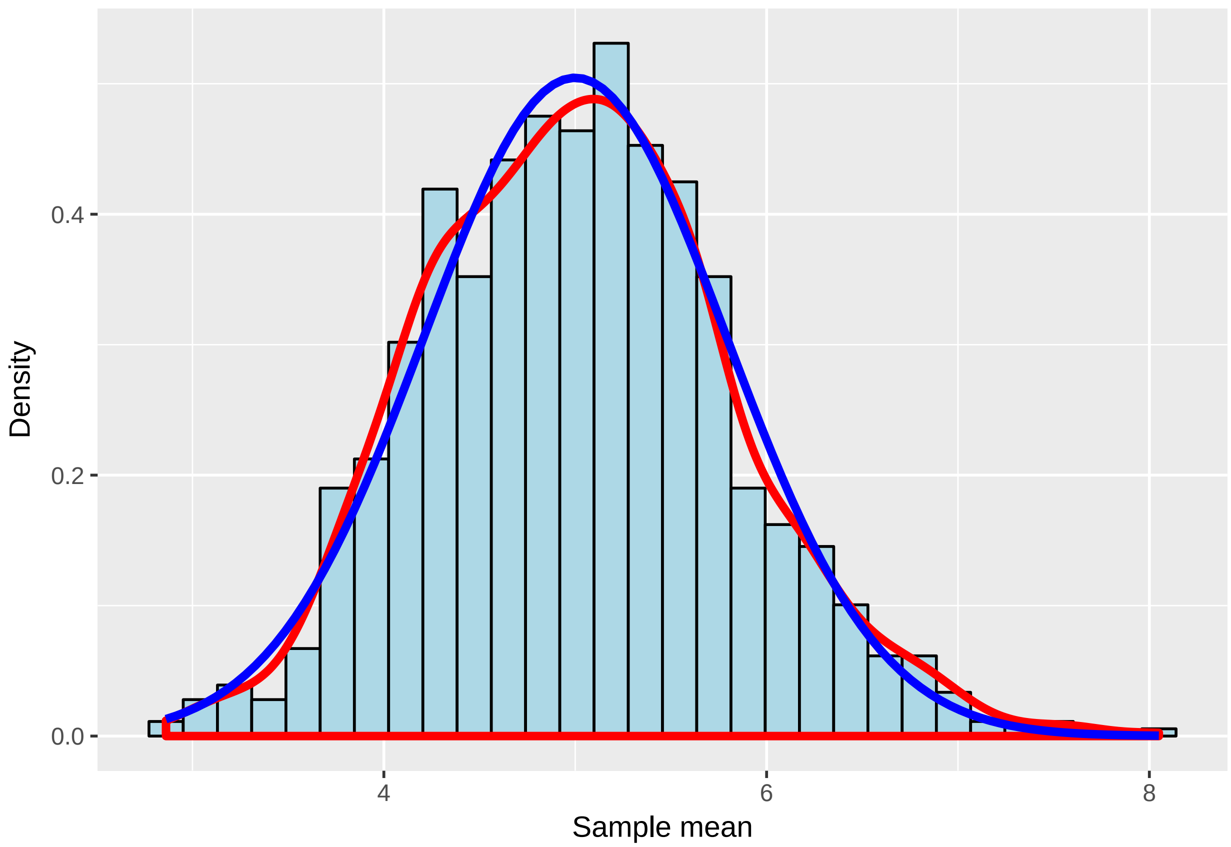
When we considerer samples of size `n` like we did, the theoretical variance of sample mean is equal to `sigma/n`. Let's compare this theoretical variance with the sample variance of the sample mean.

```
theoreticalVar <- 1/(lambda^2*n)
sampleVar <- var(means)
```

The theoretical variance is 0.625 and the sample variance 0.6346577. Those are pretty too ! If our sample size was bigger, the Central Limit Theorem tells us that the sample variance would converge towards the theoretical variance and therefore the difference would eventually get even smaller.

Distribution

The Central Limit Theorem tells us that the distribution of the sample means converges to a normal distribution with mean equal to the population mean, `1/lambda`, and standard deviation equal to the standard error, `1/(sqrt(n)*lambda)`.
In order to illustrate this result, we present another histogram of the sample means but we add two distributions to the plot. The red one is the estimated density of our sample means obtained with the `geom_density` function, and the blue one is the normal distribution with mean equal to the population mean and standard deviation equal to the standard error.
You can see with these plot that those distributions are indeed very close! And they would match even more when `n` gets bigger and bigger.


0.4
0.2
0.0
Density
4
6
8
Sample mean

Appendices

Code for "Sample mean versus theoretical mean" plot

```
g_mean <- ggplot(data.frame(means), aes(means)) +
    geom_histogram(color = "black", fill = "lightblue") +
    geom_vline(xintercept = sampleMeanMean, size = 1, color = "red") +
    geom_vline(xintercept = theoreticalMean, size = 1, color = "blue") +
    labs(main = "Sample mean versus theoretical mean", x = "Sample mean", y = "Count")
suppressMessages(print(g_mean))
```

Sample distribution vs theoretical distribution

```
g_distro <- ggplot(data.frame(means), aes(means)) +
    geom_histogram(aes(y =..density..), color = "black", fill = "lightblue") +
    geom_density(color = "red", size = 1.5) +
    stat_function(fun = dnorm,
                  args = c(mean = 1/lambda, sd = 1/(lambda*sqrt(n))),
                  colour = "blue", size = 1.5) +
    labs(main = "Sample distribution vs theoretical distribution",
         x = "Sample mean",
         y = "Density")

suppressMessages(print(g_distro))
```