Statistical Inference - Project (Part I)

Benny96

August 12th 2016

Simulation Exercise: Exponential Distribution.

Author: Benny Galdós

Statement:

In this project you will investigate the exponential distribution in R and compare it 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. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

Overview:

As the statement of the project indicates, the aim of this short report is to obtain further knowledge about the exponential distribution and comparing it with the CLT, based on some simulations. To do so, I will be using some of the packages R offers. The outline, considering the statement given by the instructors, will be the following:

- Simulations.
- Sample Mean vs Theoretical Mean.
- Sample Variance vs Theoretical Variance.
- Distribution

Simulations:

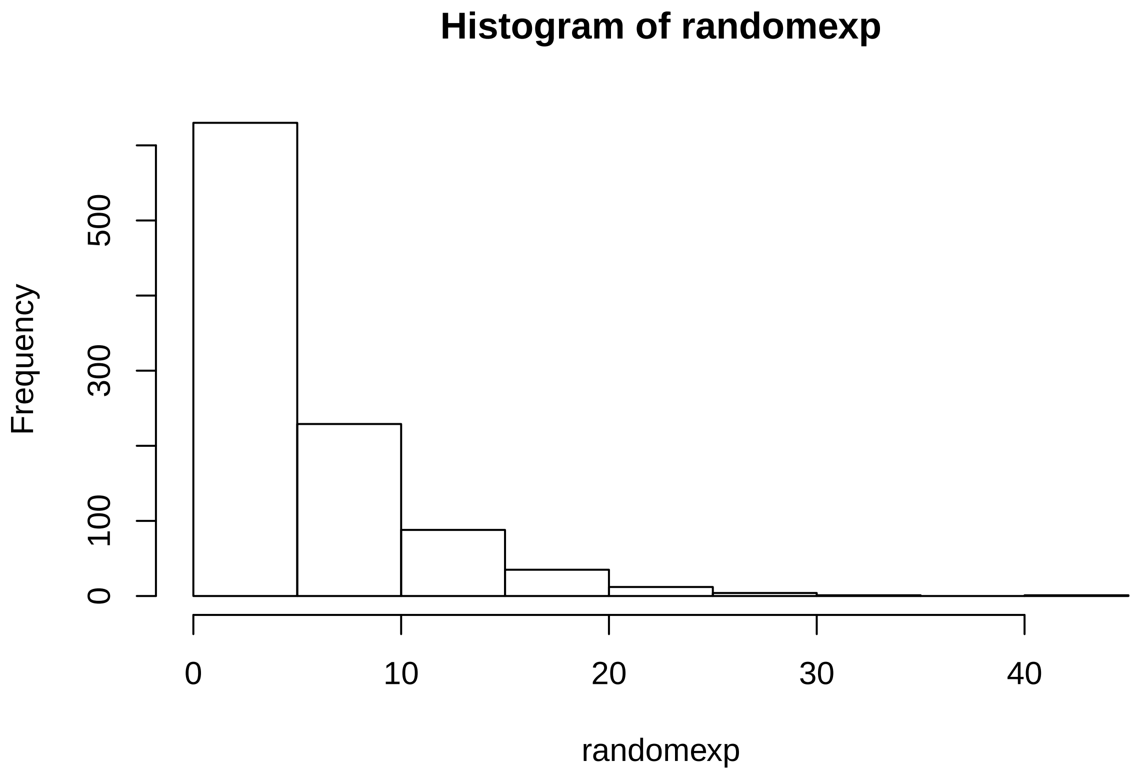
Let's start loading the required libraries, variables and seed:

```r
library("ggplot2")
n <- 40;lambda <- 0.2; sim <- 1000
avgdata = NULL
set.seed(12354)
```

Plotting the data:

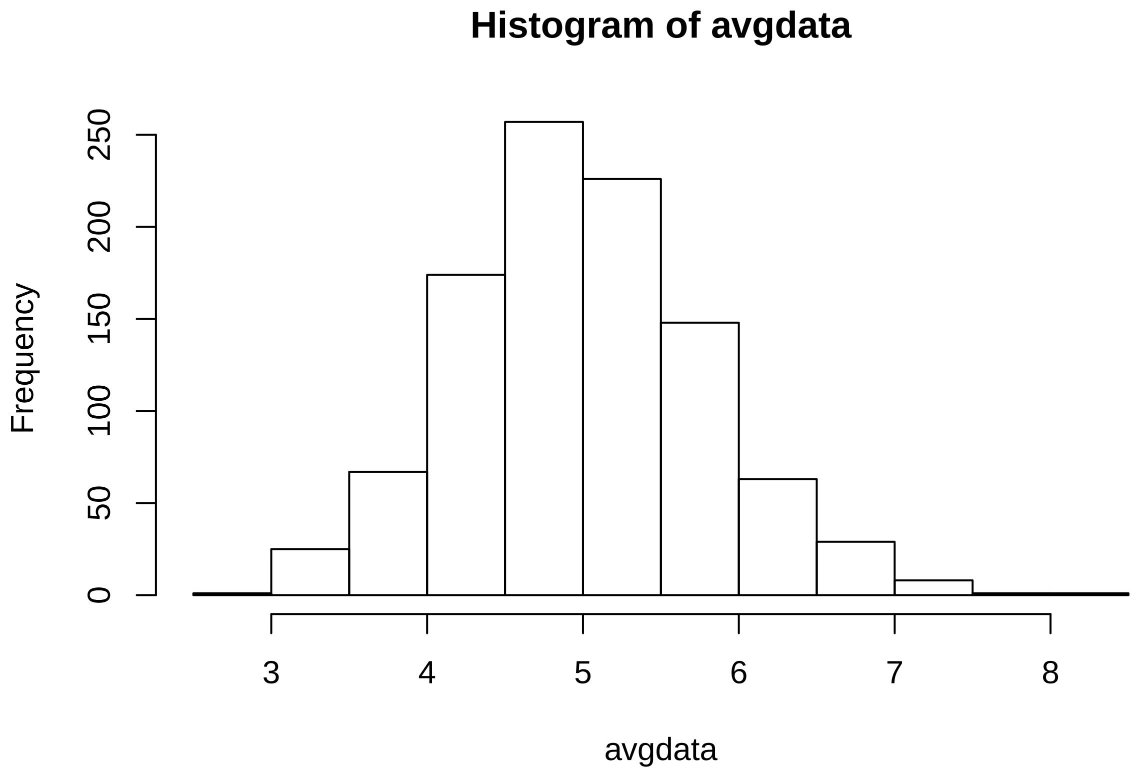
First, we'll see what the exponential distribution looks like, plotting 1000 random exponentials.

```
randomexp <- rexp(sim,lambda)
```



Now, we'll plot "avgdata", which holds the distribution of 1000 averages of 40 random exponentials.

```
for (i in 1 : sim)
{avgdata = c(avgdata, mean(rexp(n,lambda)))}
```



Reminder:

According to the statement, we already know theoretically that:

- Mean = 1/lambda
- Standard deviation = 1/lambda
- Variance = (Standard deviation)^2 = (1/lambda)^2

So, we will store these values for future comparisons:

```
Tmean <- 1/lambda
Tsd <- 1/lambda
Tvar <- (1/lambda^2)
```

Aswell as the values from the data we simulated (stored in the variable "avgdata"):

```
Smean <- mean(avgdata)
Ssd <- sd(avgdata)
Svar <- var(avgdata)
```

Sample Mean vs Theoretical Mean:

Displaying the obtained data:

```
Smean ## Mean of the chosen sample.
```

```
## [1] 4.996111
```

```
Tmean ## Theoretical mean.
```

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

We notice that the difference is quite low. This means that in large numbers, both distributions tend to have mean 5, in this case.

Sample Variance vs Theoretical Variance:

```
Svar ## Variance of the chosen sample.
```

```
## [1] 0.6118028
```

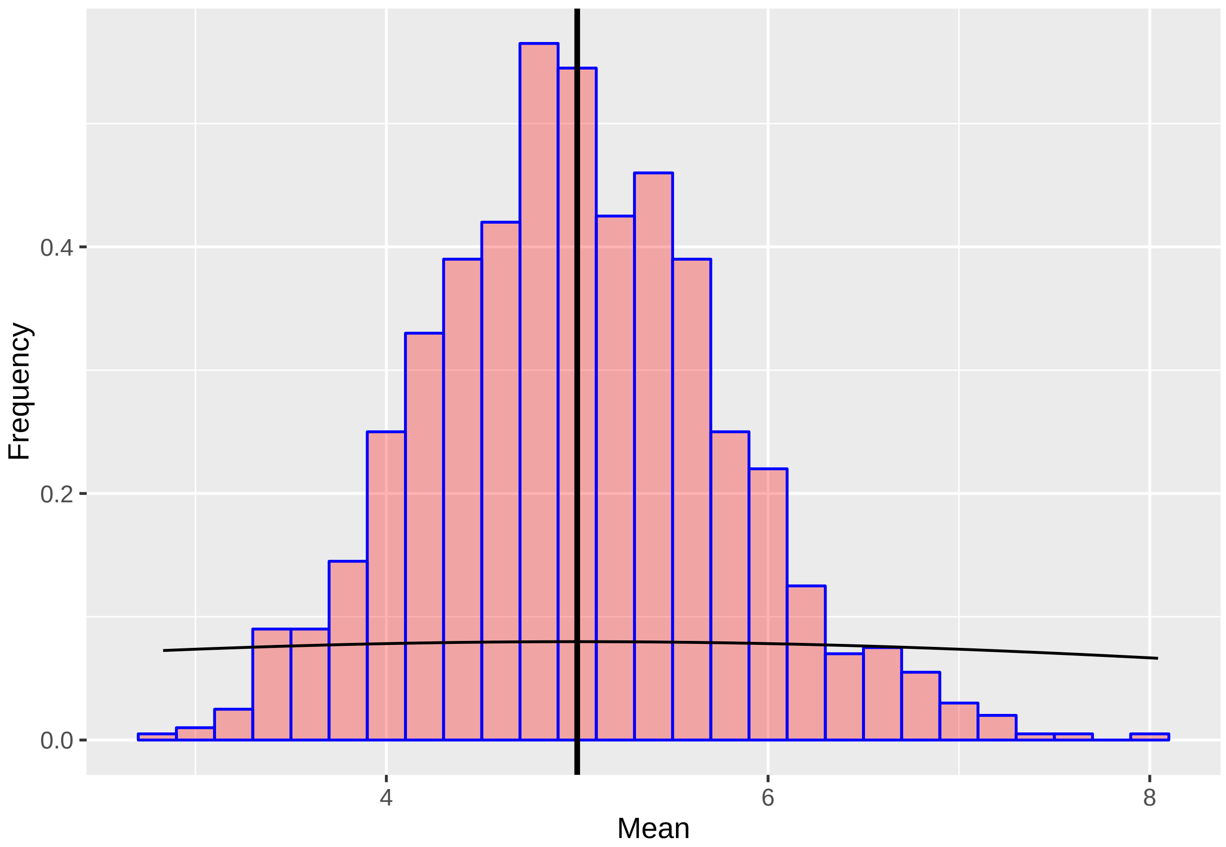
```
Tvar ## Theoretical variance.
```

```
## [1] 25
```

However, regarding the variance, we have a big difference between them. The sample's variance is somewhere near 0.65, while the theoretical variance is 25. This is somehow explained because of the difference of distribution they show, and it just tells us that the values of the exponential distribution will be much more spread than the values obtained from the averages.

Distribution:

In this case, we have to mention that even if both calculations have been originally made from randomly generated exponential numbers, the distribution of 1000 averages of 40 random exponentials is quite different. According to the CLT (Central Limit Theorem), it tends to form a Gaussian distribution. And, in fact, we can assert that it is true looking at the following histogram:


0.4
0.2
0.0
Frequency
4
6
8
Mean