Statistical Inference Project 1

Isaac G Veras

05/10/2023

Introduction:

In this project, the exponential distribution in R and the comparison with the Central Limit Theorem will be investigated. The exponential distribution will be simulated in R with rexp(n, lambda), where lambda is the rate parameter. The mean of the exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Setting lambda = 0.2 for all simulations. In this way, the distribution of the means of exponential 40 will be investigated. It will be necessary to do a thousand simulations.

Part 1: Simulation Exercise Instructions

Questions:

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.

Package installation:

```
if (!require("pacman")) install.packages("pacman")
```

```
## Carregando pacotes exigidos: pacman
```

```
pacman::p_load(pacman,     # Package Manager
               knitr,      # Transform R Markdown documents into various output formats
               plyr,       # Data manipulation
               data.table, # Manipulate, process and analyze large data sets
               tidyverse   # Data organization
)
```

Pre-processing:

Setting a seed for reproducibility:

```
set.seed(111)
```

Setting a rate parameter lambda:

```
lambda = 0.2
```

Setting a size of a sample:

```
n = 40
```

Setting a number of simulations:

```
number_simulations = 1000
```

Generating 1000 samples of 40 exponentials and calculating their mean values:

```
samples <- replicate(number_simulations, rexp(n, lambda))
dim(samples)
```

```
## [1]   40 1000
```

```
class(samples)
```

```
## [1] "matrix" "array"
```

We can see that “samples” is a matrix of 40 rows and 1000 columns. Since each column contains a sample of 40 random exponentials we’ll apply mean() to columns to get 1000 sample means.

```
exponencial_means <- apply(samples, 2, mean)
```

1. Show the sample mean and compare it to the theoretical mean of the distribution.

Sample(empirical) mean:

```
empirical_mean <- mean(exponencial_means)
empirical_mean
```

```
## [1] 5.02562
```

Theoretical mean:

```
theoretical_mean<- 1/lambda
theoretical_mean
```

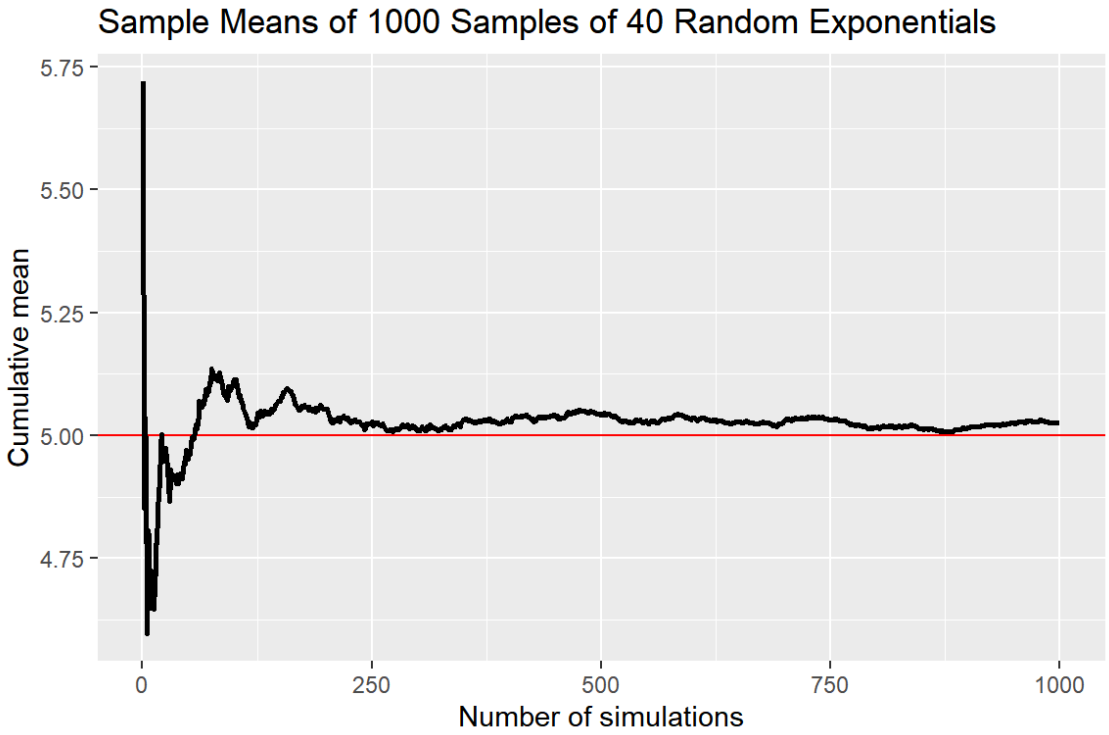
```
## [1] 5
```

We can see that the sample mean 5.02562 is good approximation of the theoretical mean t_mean = 5.

```
means <- cumsum(exponencial_means)/(1:number_simulations)
library(ggplot2)
plot_var <- ggplot(data.frame(x = 1:number_simulations, y = means), aes(x = x, y = y))
plot_var <- plot_var + geom_hline(yintercept = theoretical_mean, colour = 'red') + geom_line(si
ze = 1)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
plot_var <- plot_var + labs(x = "Number of simulations", y = "Cumulative mean")
plot_var <- plot_var + ggtitle('Sample Means of 1000 Samples of 40 Random Exponentials ')
plot_var
```



As we can see from the graph, empirical mean is a consistent estimator of theoretical mean, because it converges to the value of theoretical mean.

2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

Sample variance:

```
sample_var <- var(exponencial_means)
sample_var
```

```
## [1] 0.6069798
```

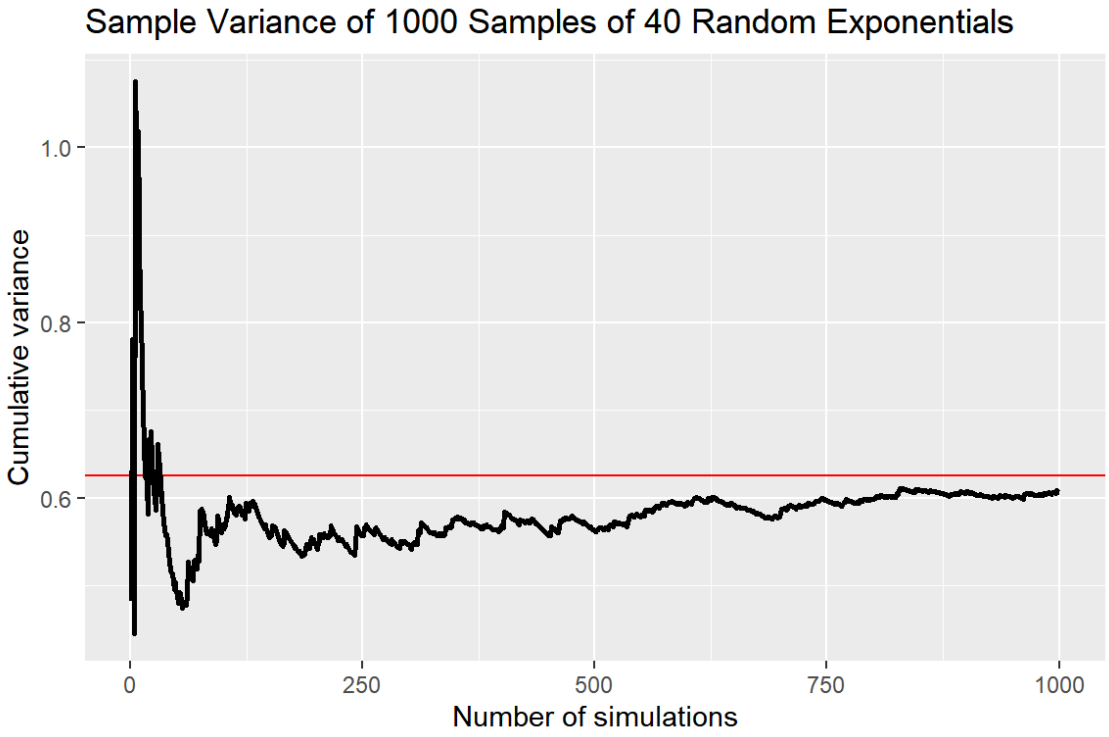
Theoretical variance of the distribution of sample means:

```
theorical_var <- (1/(lambda*sqrt(n)))^2
theorical_var
```

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

As we can see, both empirical and theoretical variances of the distribution of sample means have value close to 0.6.

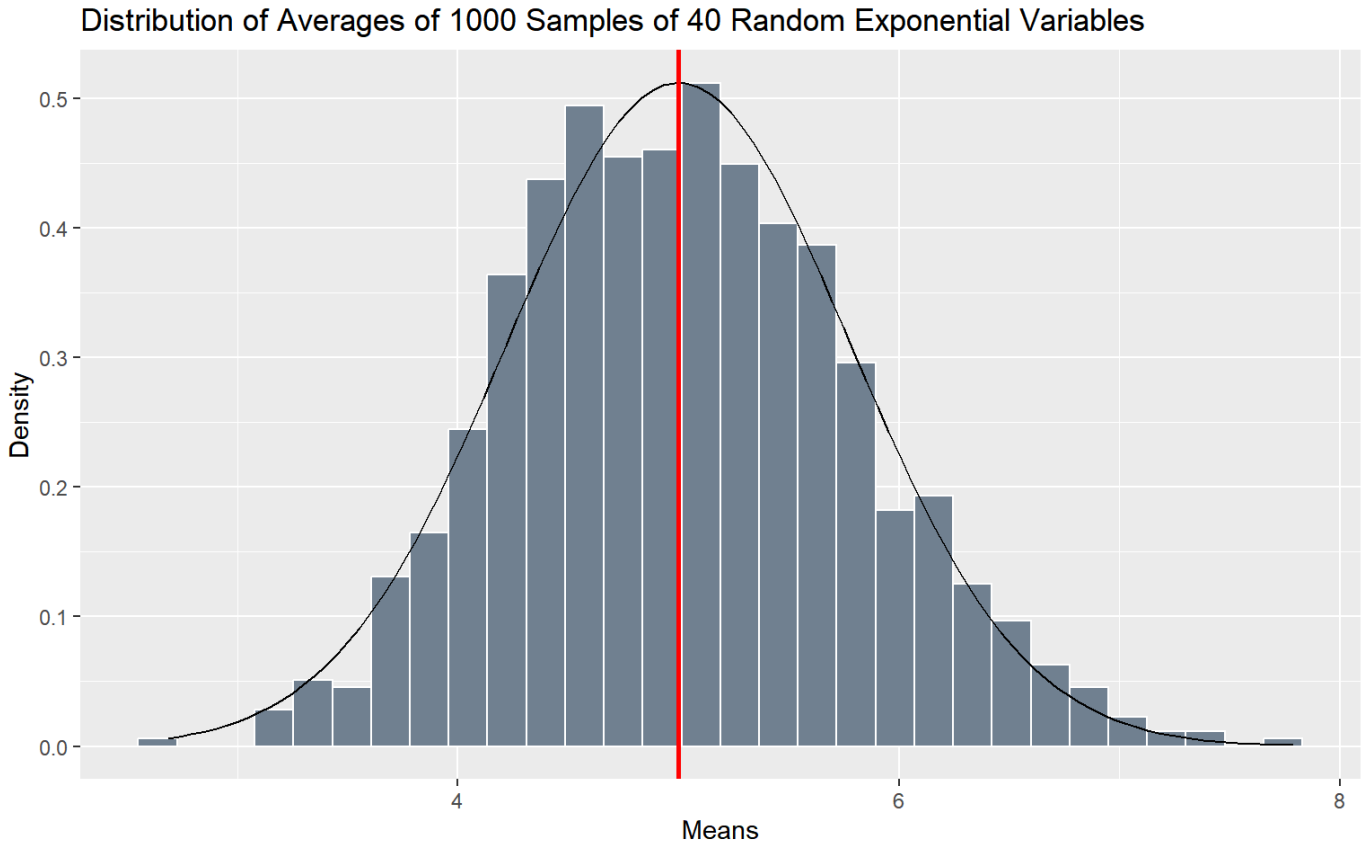
```
cumvar   <- cumsum((exponencial_means - empirical_mean)^2)/(seq_along(exponencial_means - 1))
plot_var <- ggplot(data.frame(x = 1:number_simulations, y = cumvar), aes(x = x, y = y))
plot_var <- plot_var + geom_hline(yintercept = theorical_var, colour = 'red') + geom_line(size
= 1)
plot_var <- plot_var + labs(x = "Number of simulations", y = "Cumulative variance")
plot_var <- plot_var + ggtitle('Sample Variance of 1000 Samples of 40 Random Exponentials ')
plot_var
```



As we can see from the graph, sample variance is a consistent estimator of the theoretical variance, because it converges to the value of the theoretical variance.

3. Show that the distribution is approximately normal.

```
se       <- sd(exponencial_means)
plot_var <- ggplot(data.frame(x = exponencial_means), aes(x = x))
plot_var <- plot_var + geom_histogram(aes(y = ..density..), colour = 'white', fill = 'slategra
y')
plot_var <- plot_var + stat_function(fun = dnorm, colour = 'black', args = list(mean = theoreti
cal_mean, sd = se))
plot_var <- plot_var + geom_vline(xintercept = theoretical_mean, colour = 'red', size = 1)
plot_var <- plot_var + ggtitle('Distribution of Averages of 1000 Samples of 40 Random Exponenti
al Variables')
plot_var <- plot_var + xlab('Means')
plot_var <- plot_var + ylab('Density')
plot_var
```



As we can see from the graph, the distribution of averages of 1000 samples of 40 iid exponentials is approximately normal.