Exponential Distribution Investigation with the CLT comparison

Teo Lo Piparo

10/11/2020

Overview

The project consists of two parts:

- A simulation exercise.
- Basic inferential data analysis.

Part 1

The 1st part will investigate the exponential distribution in R and compare it with the Central Limit Theorem taking into consideration the distribution of averages of 40 exponentials, throughout thousand simulations.

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

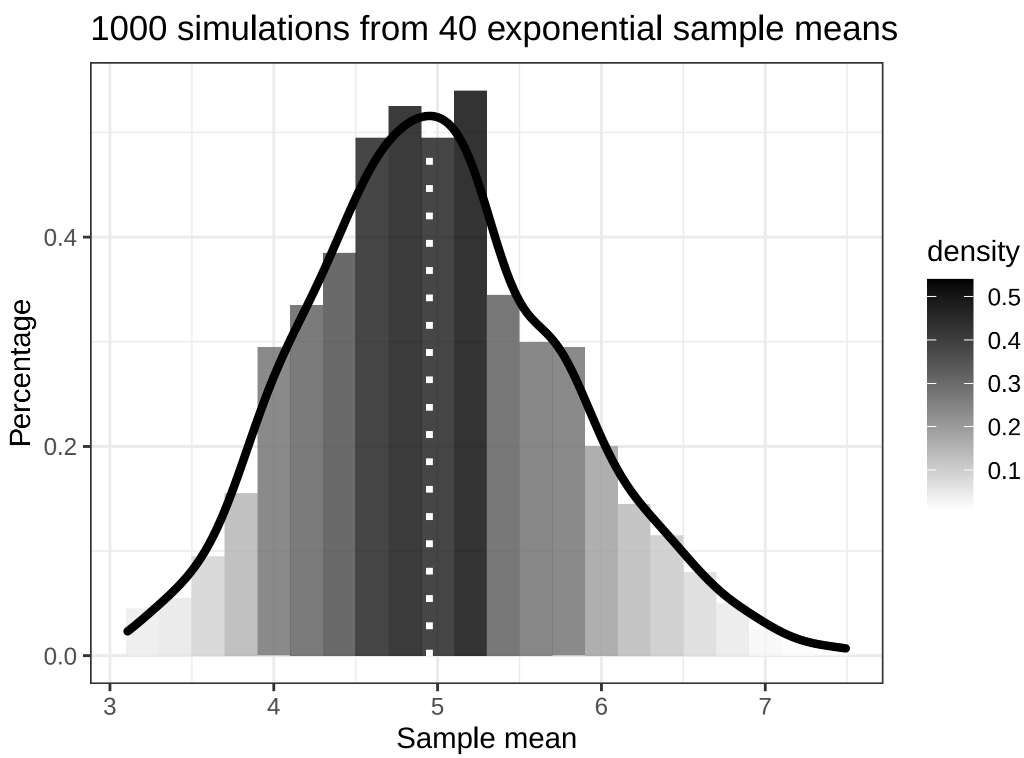
Set the seed for reproducibility Simulate $n = 40$ standard normal random exponential variables with and σ of $1/\lambda = 5$ Let X_i be the variable i Then note that $= E[X_i] = 1/\lambda = 5$ $Var(X_i) = (1/\lambda)^2 = 25$ Take average of each simulation of $n = 40$ standard normal random exponential variables Repeat this over and over for $t = 1000$ times

```
set.seed(1)

n = 40 ## sample size
t = 1000 ## no. of simulations
lambda = 0.2 ## lambda as the rate paramenter
theor.mu = 1/lambda ## theoretical mean
theor.sd = 1/lambda ## theoretical standard deviation
df = data.frame("mu" = c(1:t)) ## dummy data frame for the 1000 simulations to store the means
```

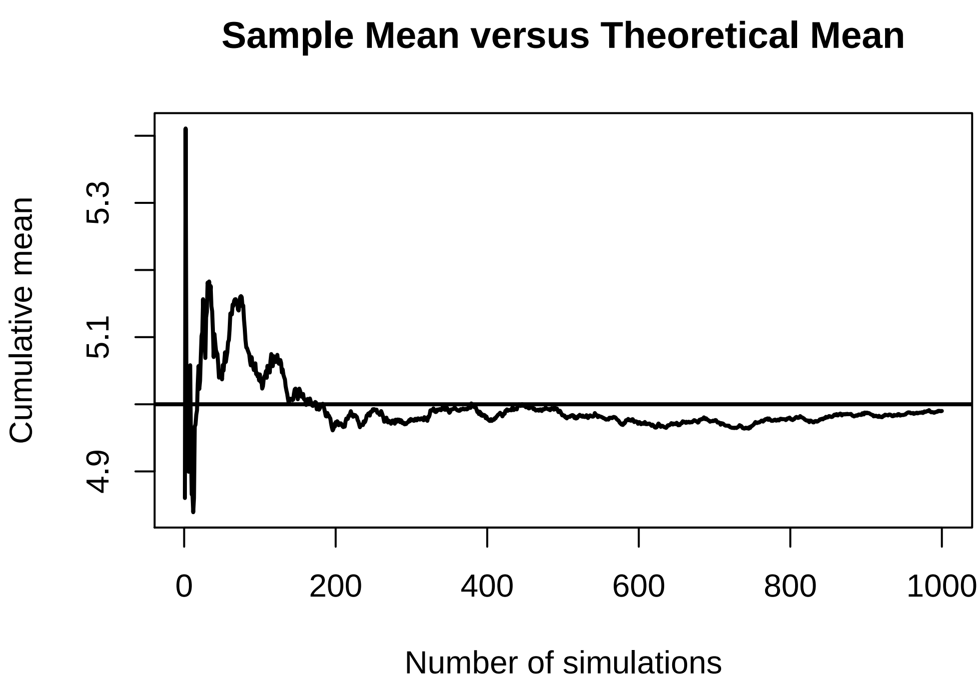
The below function generate, for 1000 times, 40 random exponential with λ set at 0.2. Then, for each simulation, the is stored in the dummy data frame.

The following plot shows the distribution of the 1000 simulations at $\bar{X}$ of sample size of 40, and the theoretical of $1/\lambda$ is close to the simulated one.



Histogram

```
means <- cumsum(df$mu)/(1:t)
```



Cumulative line plot

The simulation and the theoretical approach show a ~ outcome and is presented as follow.

```r
data.frame("Theory mean" = theor.mu, "Simul. mean" = mean(df$mu))
```

```
##   Theory.mean Simul..mean
## 1           5    4.990025
```

As t increases, the $\bar{X}$ converges to the theoretical μ:

$\bar{X}$ of 1000 simulations of average of 40 exponentials is 4.990025.

Theoretical   is $1/\lambda = 5$

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

Calculate the theoretical variance using the know variables λ and n. Knowing that the Var_{X_i} is $(1/\lambda)^2/n$, we can compare it with the calculated variance of the simulation.

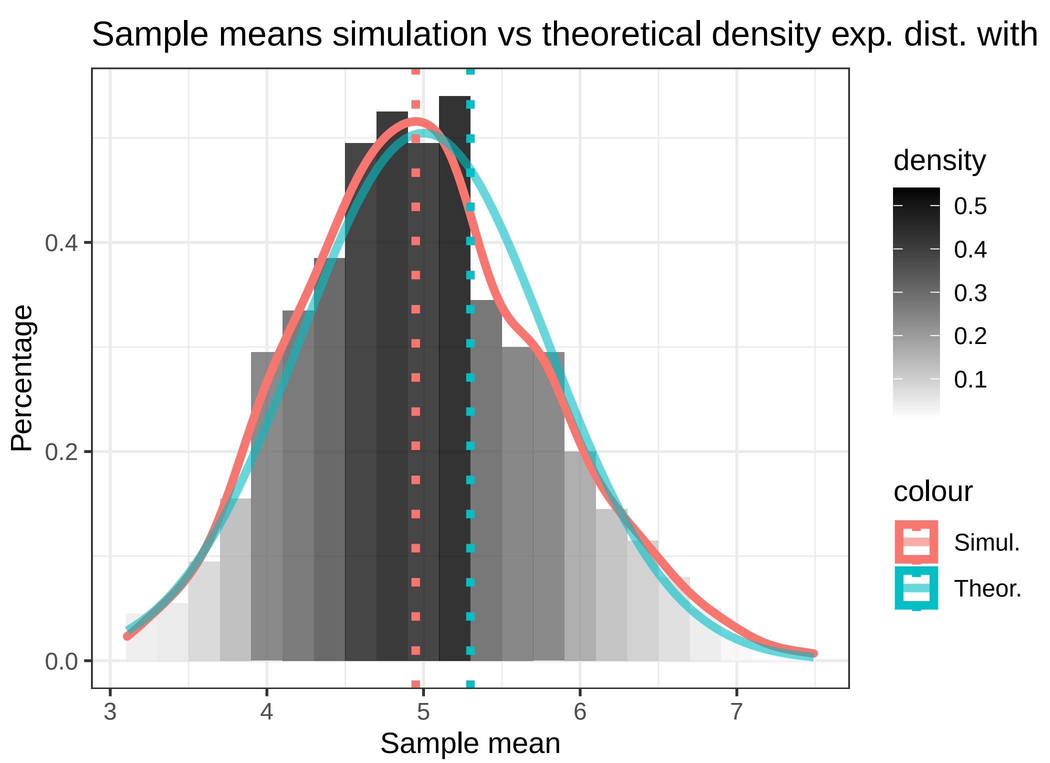
```r
data.frame("Theory variance" = theor.var, "Simul. variance" = simul.var)
```

```
##   Theory.variance Simul..variance
## 1           0.625       0.6111165
```

3) **Show that the distribution is approximately normal**

Create an overlay for the known $\bar{X}$ distribution plot that shows the theoretical distribution, in order to identify similarity in their densities. Identify the X ais by the sequence of 100 binwidth from the min value of the $\bar{X}$ and the max value of thr $\bar{X}$, in order to get feasible quantile distribution of the X axis array. Then, identify the Y axis by using the density function of the normal distribution to determine the $\bar{X}$ by using the theoretical   which we previously confirmed to be equal to $1/\lambda = 1/0.2 = 5$, and the theoretical σ which equal to $\frac{1/\lambda}{\sqrt{n}}$

```r
df$theor.x.density = seq(min(df$mu), max(df$mu), length=100) ## Divide by 100 times to get quantiles
df$theor.y.density = dnorm(df$theor.x.density, mean=theor.mu, sd=(1/lambda/sqrt(n))) ## Density normal
```

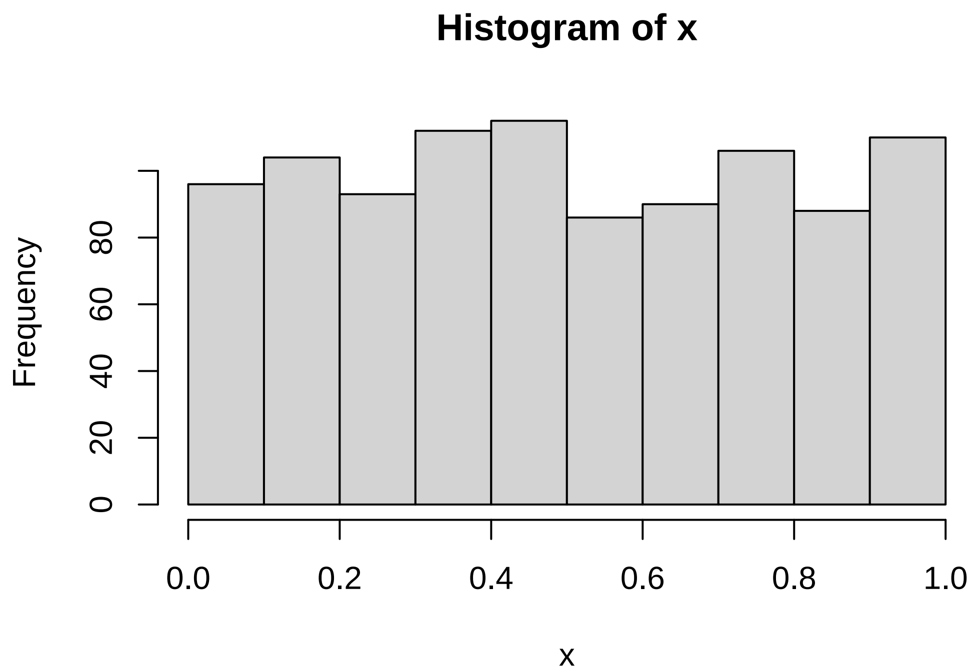


Appendix A

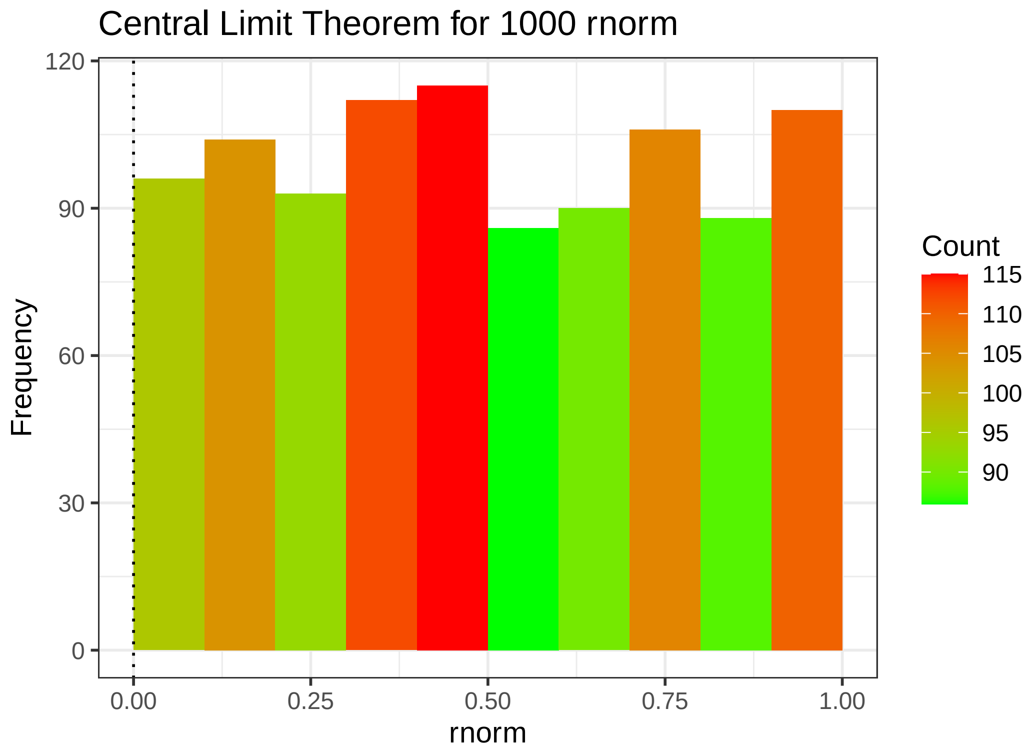
CLT explaination for asynthopic values, with infinite possibilities. This Appendix sere a ground theory for the assignment.

Having a population of random uniforms of size 1000. The cumulative mean for each observation is then calculated by the cumulative sum of the population divided by each observation.

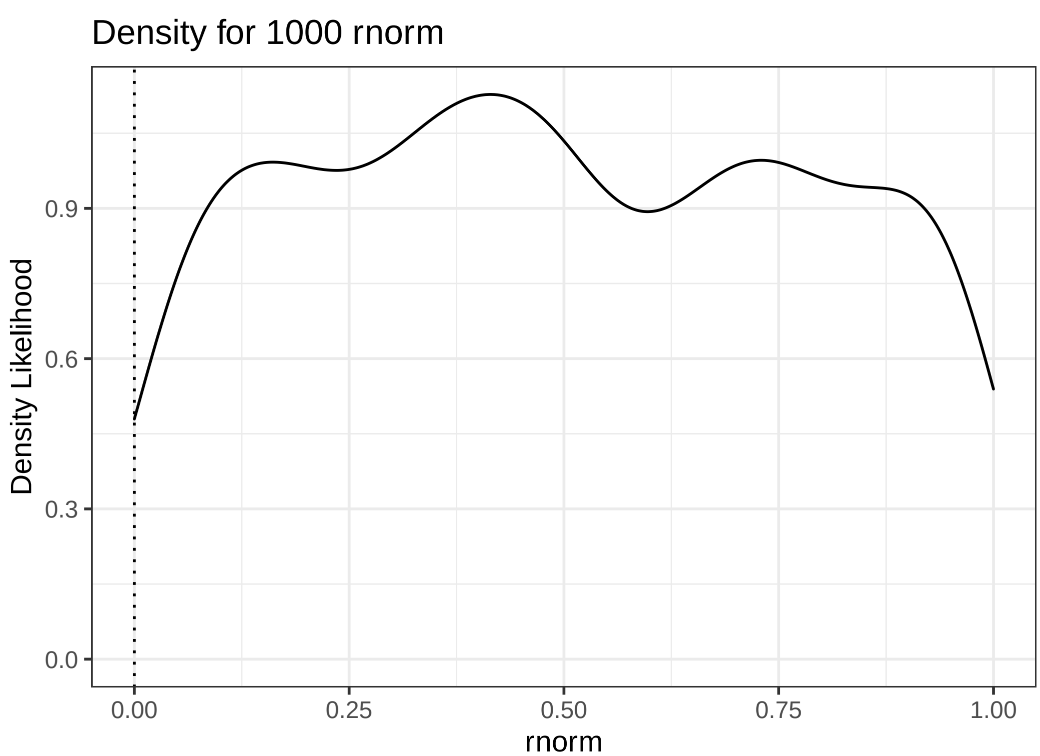
The frequency of the values for the population looks as follow.



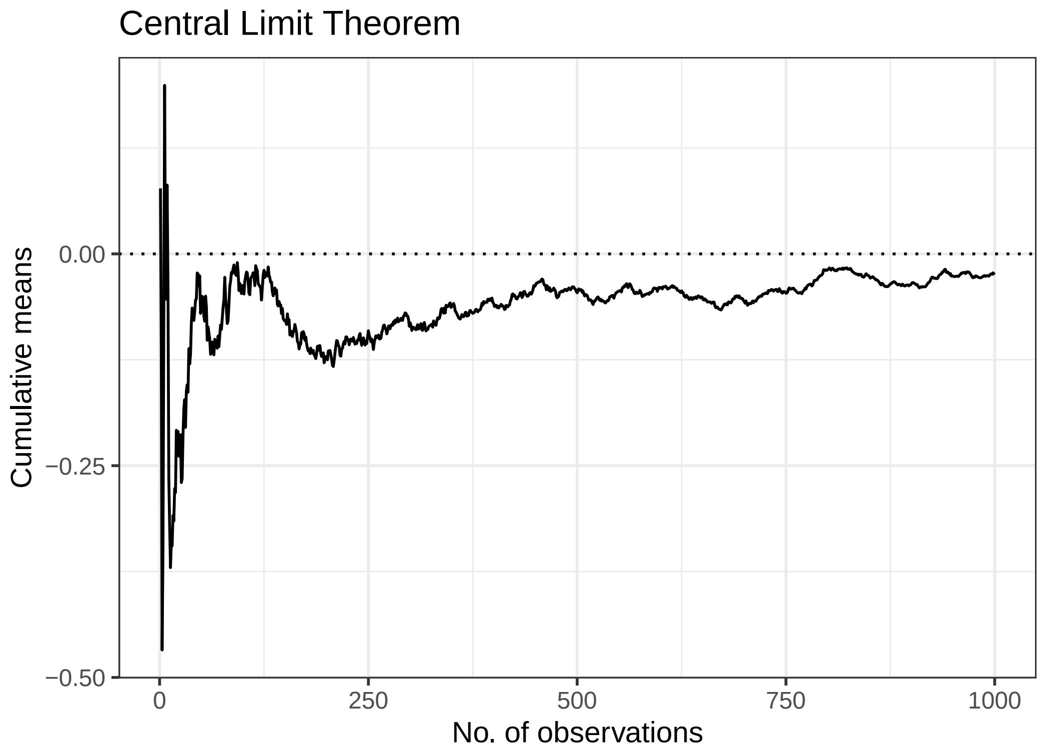
Furthermore, to better understand the Central Limit Theorem, some color code for the frequency is introduced.



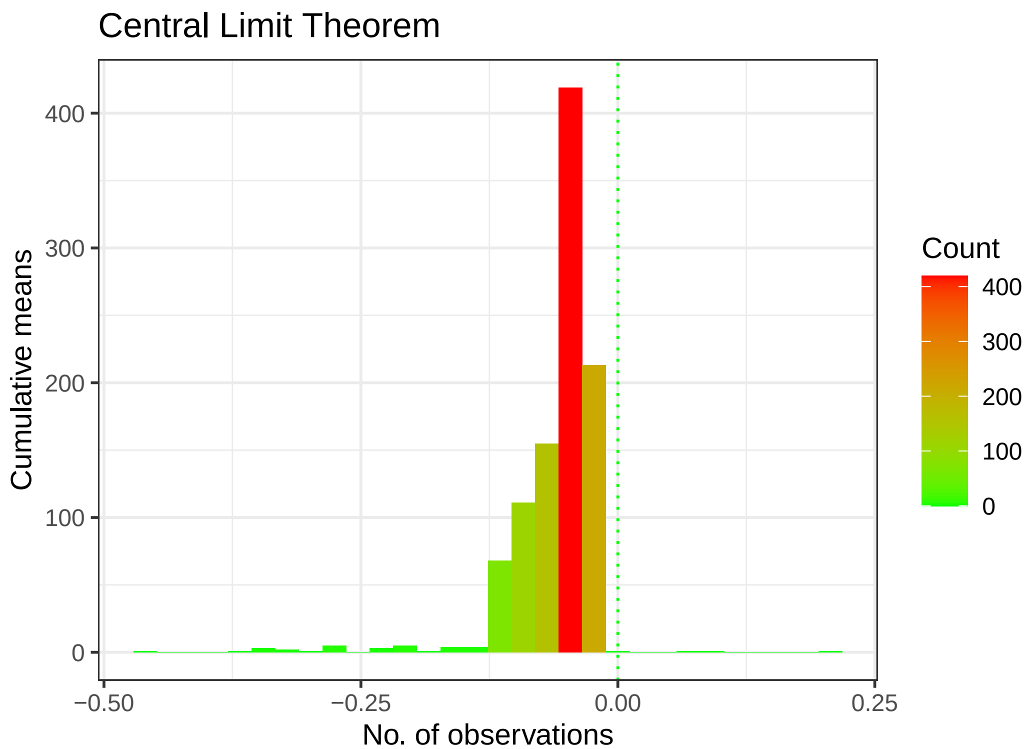
To show the the distribution of the value likelihood is equal for the population, a density visualization is provided.



Now, doing the same thing with the cumulative means of each observation the Central Limit Theorem is nicely explained, showing an alignment with the true population value, which is 0. This works only on large population sizes.



Equally, to better visualize the message, a similar color coded histogram can better clarify the CLT by showing the distribution of the cumulative means.



This prove, in practice, what in theory is explained by the CLT and presented by the next example. In fact, given an asymptotic population with infinite values the population true value will align on 0. However, the theorem is only explainable with large population sizes, as the cumulative mean will get closer to 0 as the number of observations increase as well.