Central Limit Theorem: A Pragmatic Approach via Simulations Using an Exponential Distribution

Francisco Ganhão

10/19/2016

Overview

The current document will study the Central Limit Theorem (CLT), using an exponential distribution through the use of repeated simulations.

It will be concluded through the Law of Large Numbers (LLN), that the distribution of the sample means, $\bar{X}$, from a population with mean μ and standard deviation σ, adjusted as $\frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ has a normal distribution with mean 0 and variance 1, i.e. $N(\mu = 0, \sigma = 1)$.

Based on the aforementioned we can conclude that using sample statistics from a given population are a good estimator of that same population.

Simulations

The exponential distribution of this assignment will be simulated using the $rexp(n, \lambda)$ function from R, where the mean is $\mu = \frac{1}{\lambda}$ and the standard deviation is also $\sigma = \frac{1}{\lambda}$.

For all simulations, λ will assume the value 0.2, where we will have a sample of $n = 40$ exponentials for a total of $m = 1000$ simulations.

The R code shown below, retrieves $m \times n$ samples of the exponential distribution, i.e. $40,000$ samples, and adjusts those samples to a matrix with 1000 rows (the number of simulations) by 40 columns (the number of samples per simulation) into the variable *simulations*.

```
m <- 1000
n <- 40
lambda <- 0.2
simulations <- matrix(rexp(m*n,lambda),m,n)
```

Sample Mean versus Theoretical Mean

Below the mean sample values, saved in *means*, from each simulation are computed as follows.

```
means <- apply(simulations,1,mean)
```

To see the effect of the CLT in action, we can compute the Expected mean value for a given number of simulations as shown in the following code.

```
mean_i <- cumsum(means) / (1  : m)
```

As for the mean value from the *theoretical* mean, in variable *tmean*, is equal to $\mu = \frac{1}{\lambda}$ as follows.

```
tmean <- 1/lambda
```

Below we can see the Law of Large numbers in action and observe that as the number of simulations increases, m, the sample mean tends to converge almost closely to the theoretical mean (in the red color).

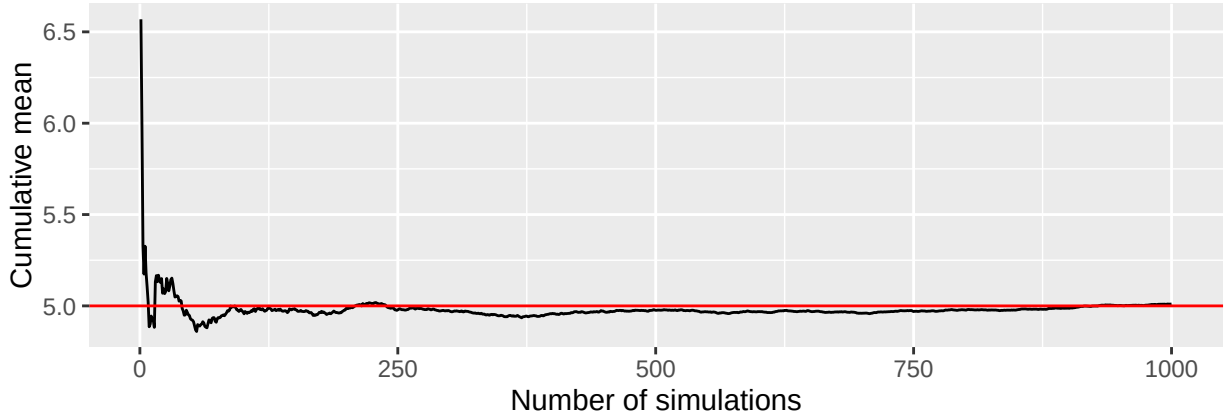


Figure 1: Cumulative mean over a given number of simulations.

Sample Variance versus Theoretical Variance

Below the sample variance values, saved in *vars*, from each simulation are computed as follows.

```
vars <- apply(simulations,1,var)
```

As for the variance value from the *theoretical* population variance, saved in variable *tvar*, is equal to $\sigma^2 = \frac{1}{\lambda^2} = 25$.

```
tvar <- (1/lambda)^2
```

Like the mean, below the cumulative average variance is computed.

```
var_i <- cumsum(vars) / (1   : m)
```

And as seen in the figure, the cumulative average variance approximately converges to the theoretical variance, in the red color, as the number of simulations increases.

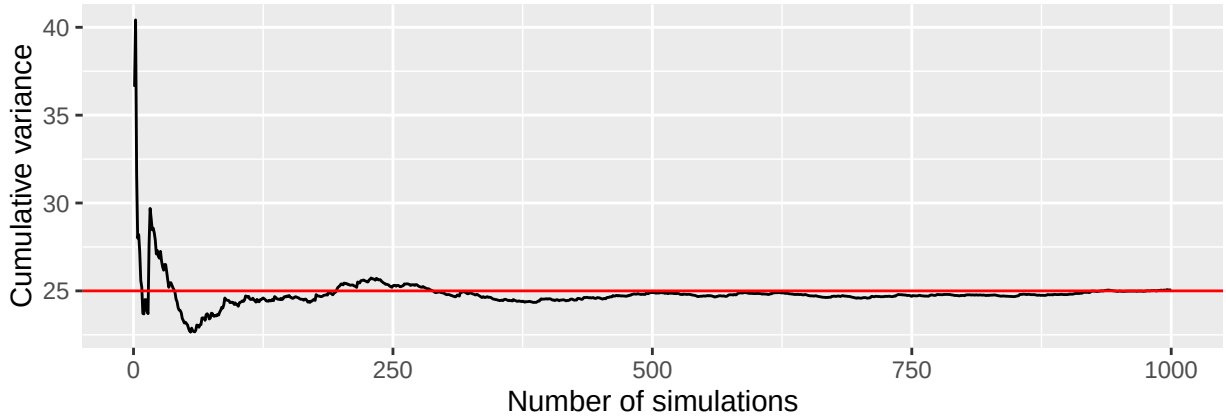


Figure 2: Cumulative variance over a given number of simulations.

Distribution

As stated by the CLT, the mean sample values in the variable *means* should form a normal distribution centered in *tmean* with standard deviation *tsd* as computed below.

```
tsd <- sqrt(tvar)/sqrt(n)
```

To see the average mean adjusted as a normal distribution as $N(0, 1)$, we could subtract the sample means by the theoretical average and divided by the expected standard deviation from the sample means according to the CLT.

```
snormal <- (means - tmean)/tsd
```

The sample mean distribution, adjusted by the CLT, is plotted against a normal distribution in a red line with mean 0, and standard deviation 1. Based on the previous plot we can conclude that the simulated values approximately follow the CLT Normal distribution with mean 0 and standard deviation 1.

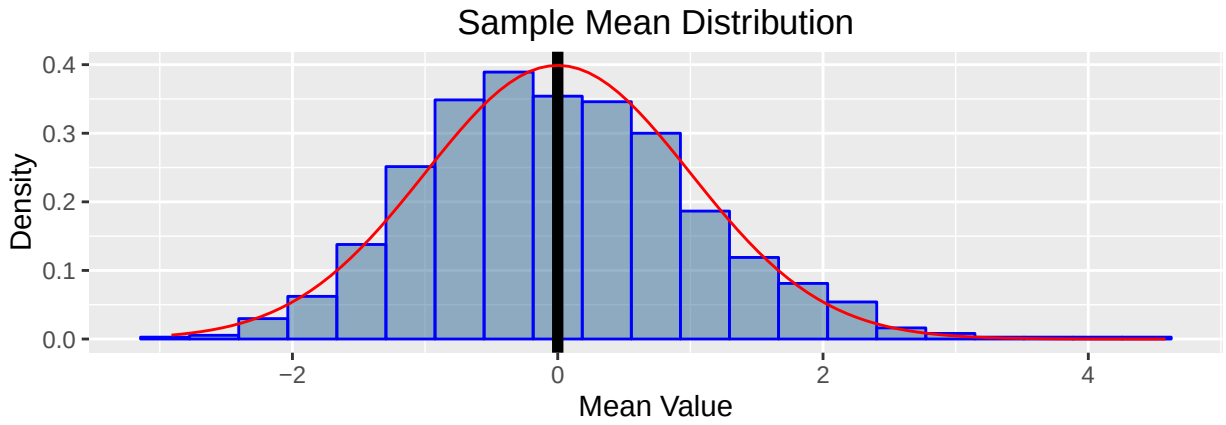


Figure 3: Sample mean distribution, adjusted by the CLT, versus a normal distribution with mean 0 and standard deviation 1.

Conclusions

From the results presented in the document we could conclude that the LLN applies when collecting several simulation samples of a given population's distribution.

After a large number of simulations, the average sample mean and the average sample variance converged respectively to the theoretical mean and respective theoretical variance.

Finally, through the CLT, we could observe that the expected sample means has a normal distribution centered on the population's average value, μ, and with standard deviation, $\frac{\sigma}{\sqrt{n}}$. The former distribution when adjusted as $\frac{\bar{X}-\mu}{\sigma}\sqrt{n}$, it follows as normal distribution $N(0,1)$.

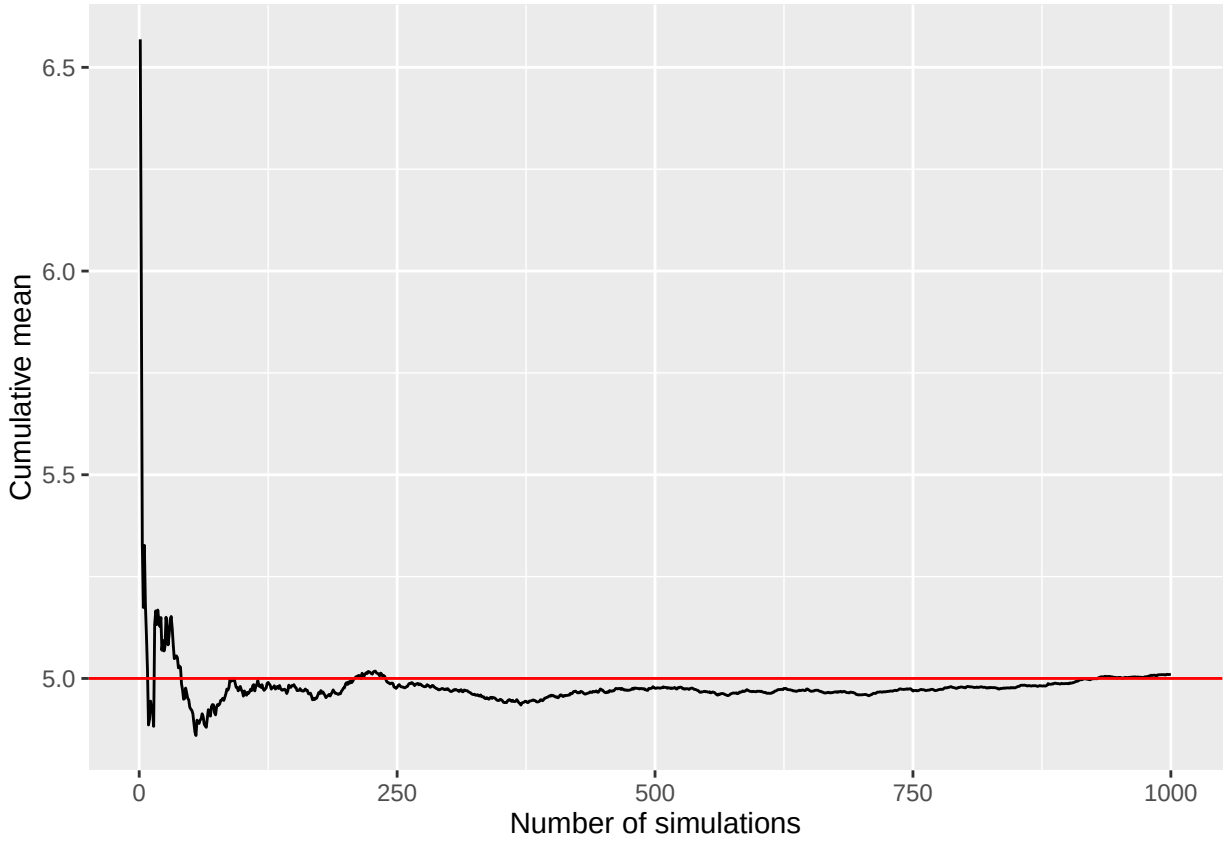
Appendix

In the appendix section, the code used to produce each plot is shown in the following sub-sections.

Sample Mean versus Theoretical Mean

Below you can find the code that produces the cumulative sample mean plot, as observed in **figure 1**.

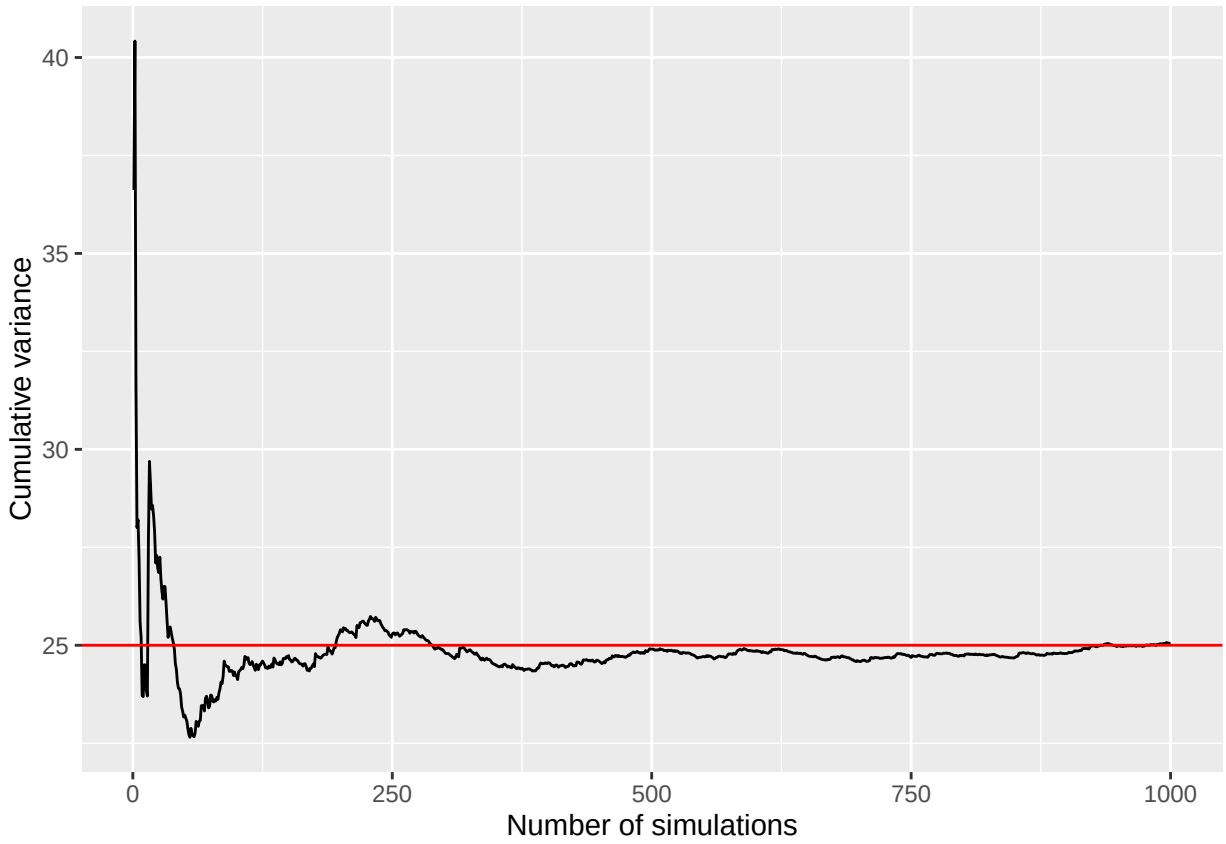
```
library(ggplot2)
g <- ggplot(data.frame(x=1:length(mean_i), y = mean_i),aes(x=x,y=y)) +
    geom_line() +
    geom_hline(yintercept = tmean, color = "red") +
    labs(x = "Number of simulations", y = "Cumulative mean")
g
```



Sample Variance versus Theoretical Variance

Below you can find the code that produces the cumulative sample variance plot, as observed in **figure 2**.

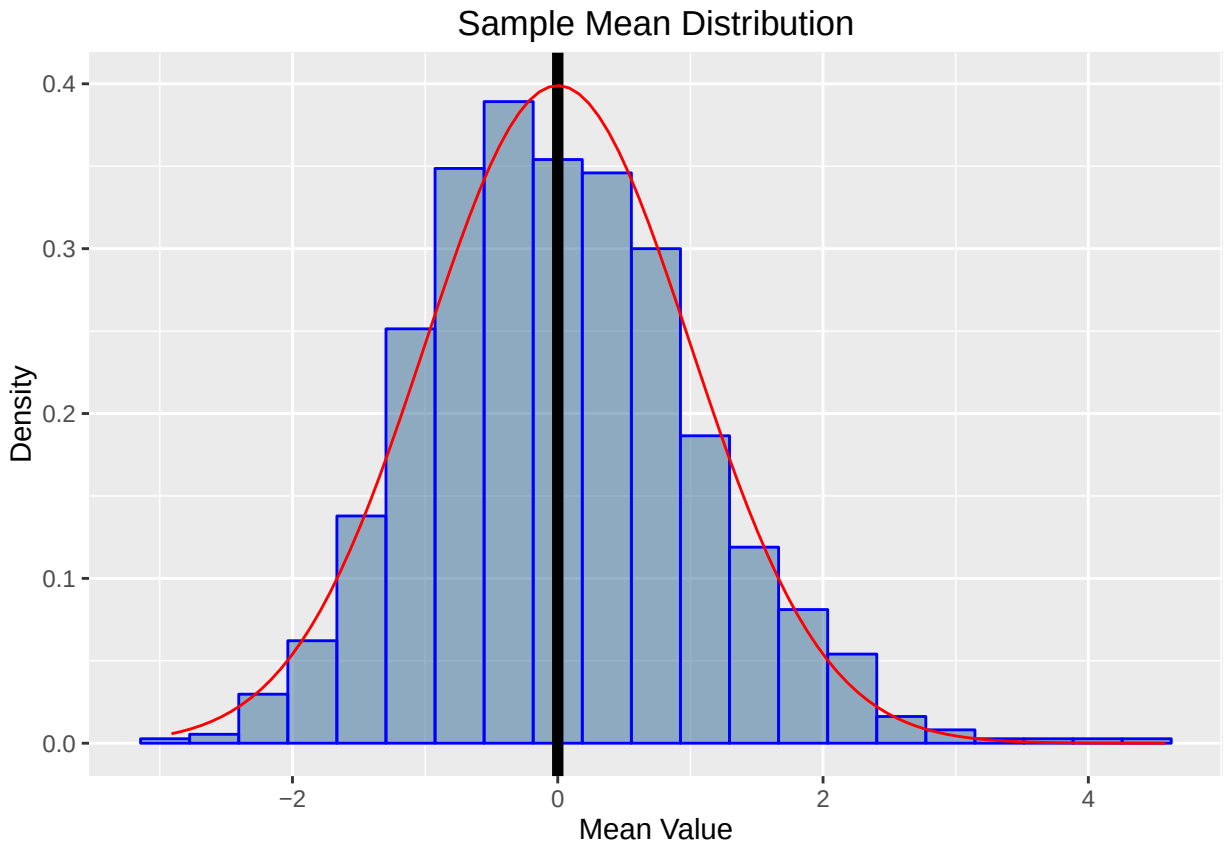
```
g <- ggplot(data.frame(x=1:length(var_i), y = var_i),aes(x=x,y=y)) +
    geom_line() +
    geom_hline(yintercept = tvar, color = "red") +
    labs(x = "Number of simulations", y = "Cumulative variance")
g
```



Distribution

Below you can find the code that produces the CLT distribution plot, as observed in **figure 3**.

```
df <- data.frame(distribution = snormal)
g<-ggplot(df, aes(x = distribution,fill=3)) +
    geom_histogram(alpha = 0.5, binwidth = 0.37, colour = "blue", aes(y = ..density..)) +
    geom_vline(xintercept = 0, size=2) +
    stat_function(fun = dnorm, colour="red", args = list(mean=0,sd=1)) +
    guides(fill=FALSE) +
    labs(title = "Sample Mean Distribution",
         x = "Mean Value",
         y = "Density")
g
```


Sample Mean Distribution
Density
0.4
0.3
0.2
0.1
0.0
−2
0
2
4
Mean Value