Investigating Exponential distribution with simulations

Harish Kumar Rongala

October 21, 2016

Overview

This is an attempt to investigate **Exponential distribution** in R and compare it with **Central Limit Theorem**. We will try to achieve that in following steps

1. Generate 1000 simulations of Exponential distribution in R
2. Calculate its theoretical **mean** and **variance**, and compare it with generated sample distribution
3. Compare the averaged distribution with original distribution
4. Form new distribution according to **Central Limit Theorem** and find if it is normally distributed

1. Simulations

In our simulations, we use **lambda** equal to **0.2** for Exponential distrbution. Initially We create a variable "sample", to store our average sample distribution. We generate 40 random exponentials, find their average and append to our "sample" variable, we do this 1000 times. For exponential distributions, theoretical mean and variance can be calculated using following formulae

- Mean: $1/\lambda$
- Variance: $1/\lambda^2$

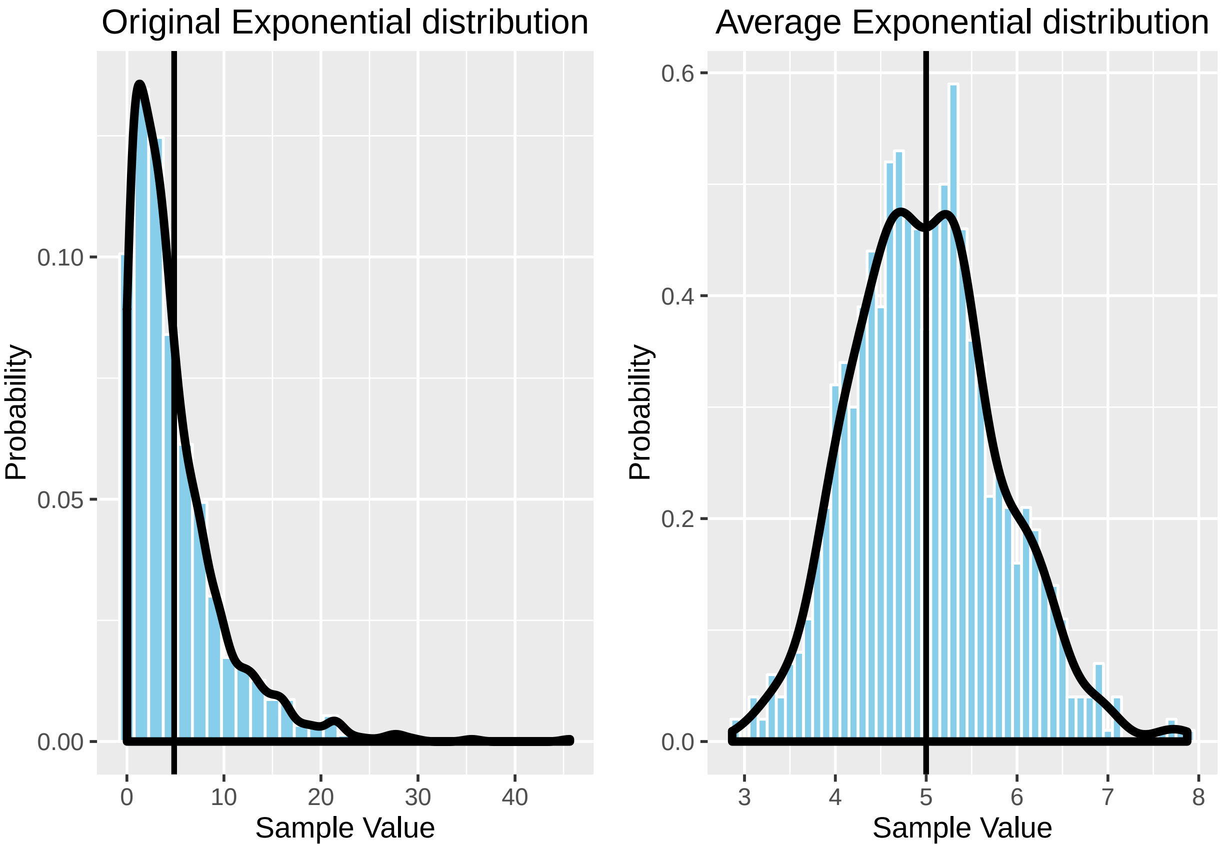
The following table compares the theoretical random variables with calculated random variables of our averaged sample distribution

```
##          Theoretical   Sample
## Mean        5.000000 4.999107
## Variance    0.625000 0.646437
```

Result: From the above table, sample mean and variance are consistent with theoretical mean and variance respectively.

2. Let's Visualize

Let's see how our sample distribution looks like. We could do this by comparing it with the original (unaltered) exponential distribution. The following plot depicts the comparision. It also shows where our distributions are centered, i.e Mean of the distribution. In our plot, black vertical lines represents the **mean** of that distribution. In our case original distribution mean is **4.871228** and averaged distribution mean is **4.999107**.



Result: From the above plot, we can notice that the averaged sample distribution looks more **gaussian** than the original distribution and they are almost centered at the same point.

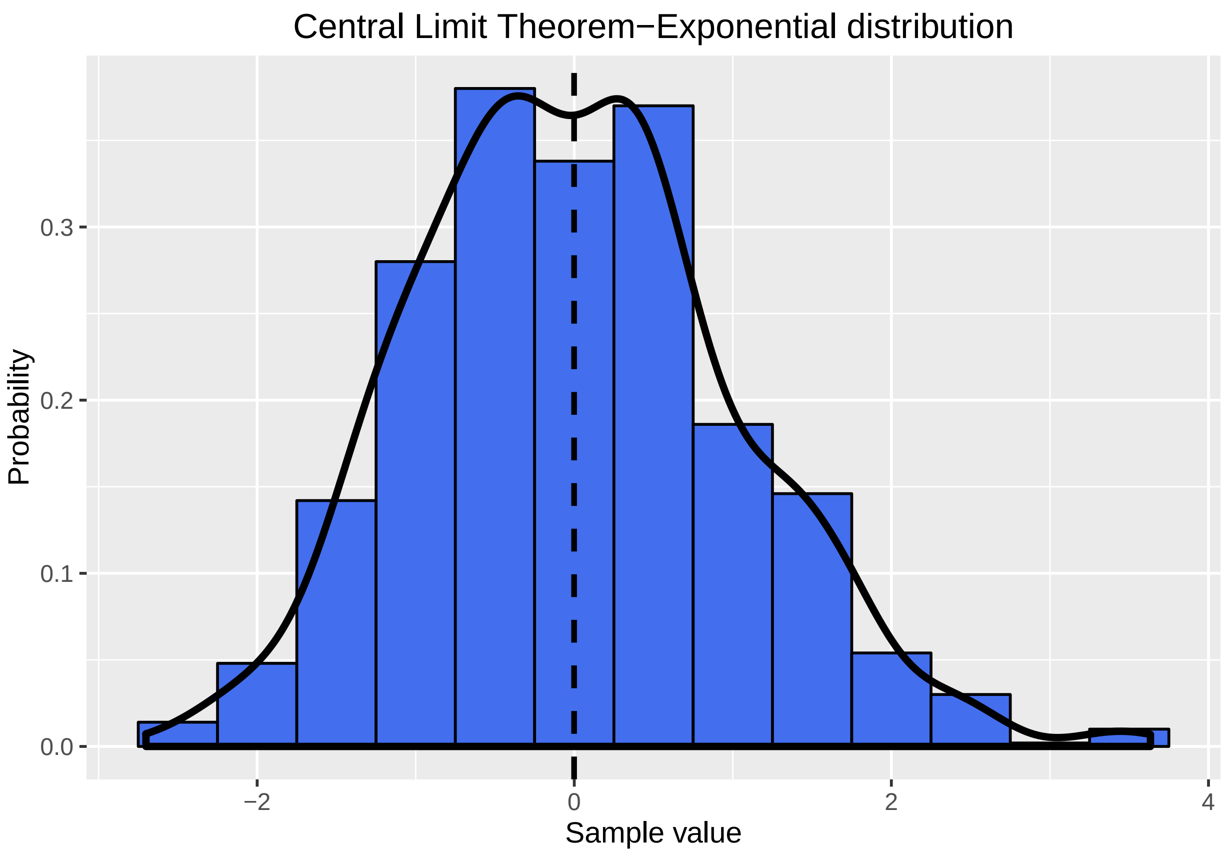
3. Comparing it with CLT

Central Limit Theorem states that distribution of averages of iid variables becomes that of a standard normal as the sample size increases. The following formula's result will have a distribution like that of a standard normal for larger sample size(n)

$$\sqrt{n}(X_n - \mu)/\sigma$$

Where,

- n = Sample size
- X_n = Sample mean
- μ = Population mean (theoretical mean)
- σ = Population standard deviation (theoretical sd)



Mean of thus formed distribution is

```
## [1] 0
```

Variance of thus formed distribution is

```
## [1] 1.03
```

Result: Mean and variance of standard normal is **0** and **1** respectively. Therefore, Central Limit Theorem is verified with Exponential distribution.

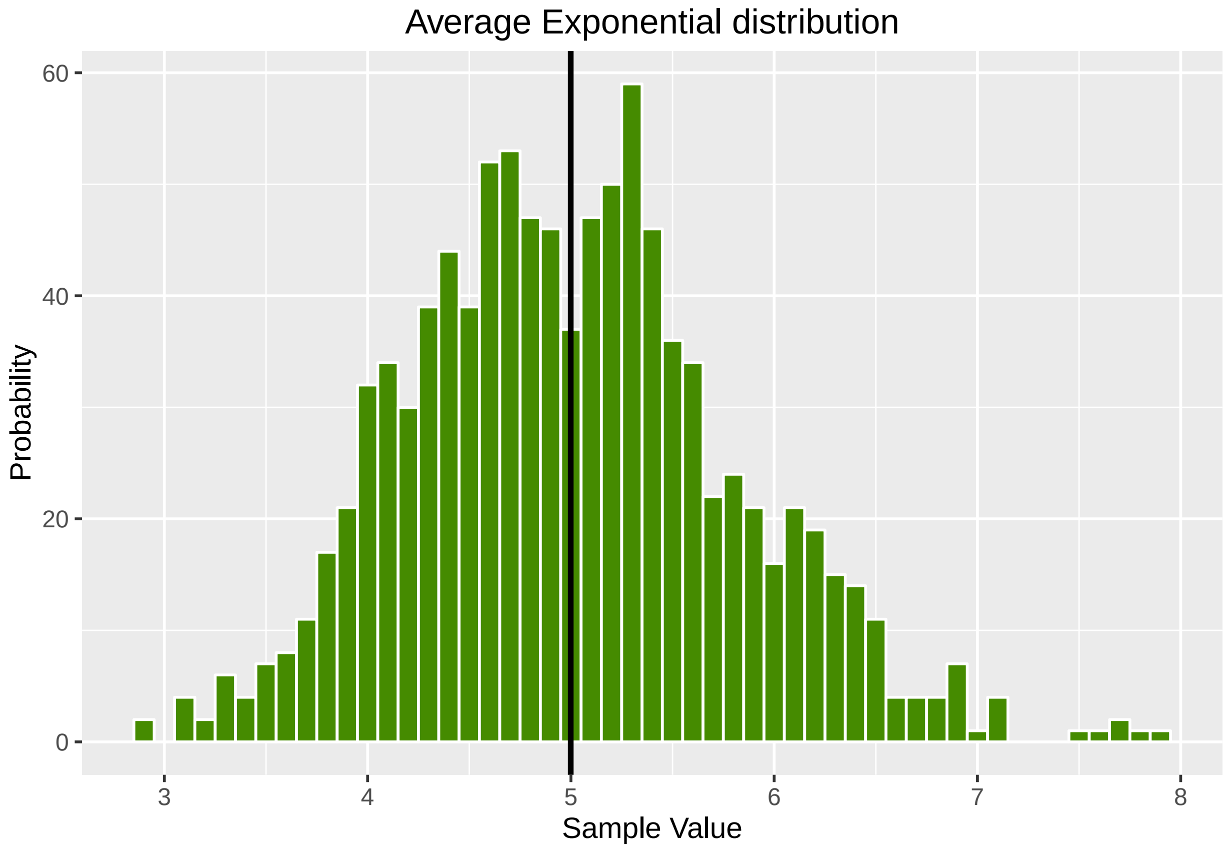
4. Appendix

The following code is used to generate simulations and tabulate results in section 1

```
# For reproducibility we set random generator seed as following
set.seed(1993)
# Variabe to hold Exponential distribution
sample=NULL
# We take lambda as 0.2
lambda=0.2
for(i in 1:1000){
sample=c(sample,mean(rexp(40,lambda)))
}
```

```
# Theoretical mean and variance
theoretical_mean=1/lambda
theoretical_var=1/lambda^2/n
# Mean and Variance of our simulations
sample_mean=mean(sample)
sample_var=var(sample)
# Tabulating the results
results<-matrix(c(theoretical_mean,sample_mean,theoretical_var,sample_var),byrow = T,ncol = 2)
rownames(results)<-c("Mean","Variance")
colnames(results)<-c("Theoretical","Sample")
results<-as.table(results)
print(results)
```

The following plot can clearly show where the averaged distribution is centered at. In this plot, black vertical line represents the **sample mean** which in our case is **4.999107**.



Code used in section 2 can be found here

```
library(ggplot2)
library(gridExtra)
# For reproducibility we set random generator seed as following
set.seed(1993)
# Variabe to hold Exponential distribution
sample=NULL
# We take lambda as 0.2
lambda=0.2
```

```
for(i in 1:1000){
        sample=c(sample,mean(rexp(40,lambda)))
}
# Calculate mean of the sample
sam_mean<-mean(sample)
# For the purpose of plotting, convert it into a dataframe
sample<-as.data.frame(sample)
# Plot the averaged distribution - sample
plot1<-ggplot(sample,aes(sample))+geom_histogram(aes(y=..density..),col="white",fill="skyblue",binwidth
set.seed(1993)
# Generate original data
orig_data<-rexp(1000,lambda)
org_mean<-mean(orig_data)
orig_data<-as.data.frame(orig_data)
# Plot original data
plot2<-ggplot(orig_data,aes(orig_data))+geom_histogram(aes(y=..density..),col="white",fill="skyblue",bin
grid.arrange(plot2,plot1,ncol=2)
```

Code used in section 3

```
set.seed(1993)
library(ggplot2)
pop_mean<-1/0.2
pop_sd<-1/0.2
se<-pop_sd/sqrt(40)
test=NULL
for(i in 1:1000){
        sam<-rexp(40,0.2)
        sam_mean<-mean(sam)
        test<-c(test,(sam_mean-pop_mean)/se)
}
test<-as.data.frame(test)
sam_mean<-mean(test$test)
sam_sd<-sd(test$test)
plot1<-ggplot(test,aes(test))+geom_histogram(aes(y=..density..),binwidth = 0.5,col="black",fill="royalbl
print(plot1)
```