Course Project - Part1: Exponential Distribution compared with the Central Limit Theorem in R

Author: Imène Drir | Source: github

Synopsis

This project looks at simulations of averages of an **Exponential Distribution** and approximate the outcome to a normal distribution as per the principles of the **Central Limit Theorem** in **R**. Link: (CLT) Theorem.

In this analysis, we will perform 1000 simulations and explore the distribution of averages of 40 exponentials to:

- Look at means of averages the 40 exponentials vs. the theroetical mean.
- Look at the Variance of our resulting distribution.
- Determine if we can approximate the distribution of averages of 40 exponentials to a normal distribution.

Set-up & Process

We start by loading packages we will use throughout this study like stats and ggplot2.

We will also set a seed in order to make this study reproducible in other users' R environements.

```
# Set random seed
set.seed(2016)
```

To perform 1000 simulations, which will be comprised of the averages of 40 exponentials. We will use the R function: rexp(n, lambda) with *lambda* (λ) as the rate parameter. We will set $\lambda = 0.2$. See R code:

```
# Set our variables for the exponential distribution
lambda = 0.2 # lambda value for this simulation exercise (rate)
nExp = 40; simNo = 1000 # number of exponentials n & number of simulations needed simNo
simExp <- rexp(nExp*simNo, rate = lambda)
simMatrix <- matrix(simExp, simNo, nExp)
```

1. Calculation of the sample mean of the exponential distribution vs.the theoretical mean of the distribution.

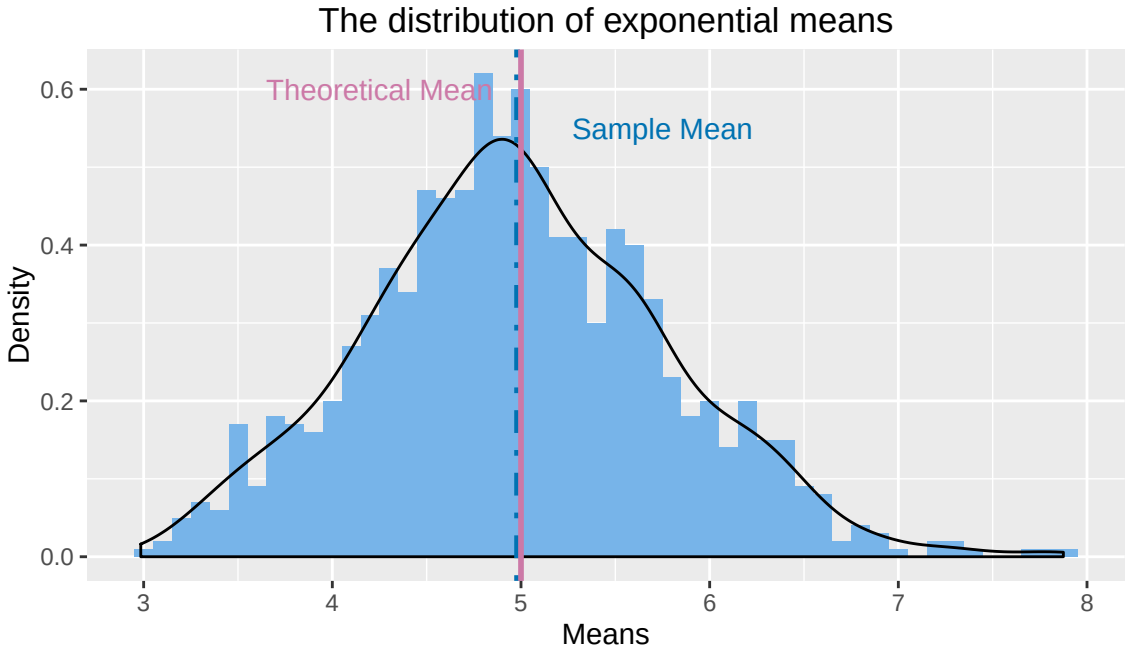
The mean of exponential distribution is $1/\lambda$, it will be referred to as the theoretical mean moving forward. We will start by calculating the sample mean and compare it to the theoretical mean. Then, we will plot the two statistics in a histogram that shows the density of the distribution of means resulting from the simulations.

Plot for the distribution of the averages

In the plot, we can see the distribution of our calculated means for the sample dataset.

```
sample_mean <- mean(simMean)
theoretical_mean <- 1 / lambda
```

Figure 1 (See appendix for R code)



Observation: in order to highlight where the data centers, the sample mean from all the caluculated averages and the actual theoretical mean $(1/\lambda)$ are plotted as vertical lines.The plot seems to indicate that averages of 40 exponentials are centered around the value of $x = 5$ which is the mean of our theoretical distribution.

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

To see if the resulting dataset had some variability, we'll proceed to calculate the sample standard deviation and variance based on our sample mean. In addition, we will include the theoretical standard deviation and variance. Per the CLT, the theoreritcal variance equals to the variance of the underlying distribution $(1/\lambda)$ divided by the sample size: $(1/\lambda)$ / (sample size). See R code:

```
# Sample Values for standard deviation and Variance
sample_sd <- sd(simMean)
sample_var <- sample_sd^2
# Per CLT we know the theoretical value for standard deviation is: sd = (1 / lambda) /(sample size)
theoretical_sd <- (1/lambda)/sqrt(nExp)
theoretical_var <- theoretical_sd^2
#Summarising in a dataframe to compare sample & theoretical values
dtresults <- data.frame(c(sample_mean, theoretical_mean),
                        c(sample_var, theoretical_var), c(sample_sd, theoretical_sd),
                        row.names = c("Sample", "Theoretical"))
colnames(dtresults) <- c("Mean", "Variance", "Standard Deviation")
dtresults
```

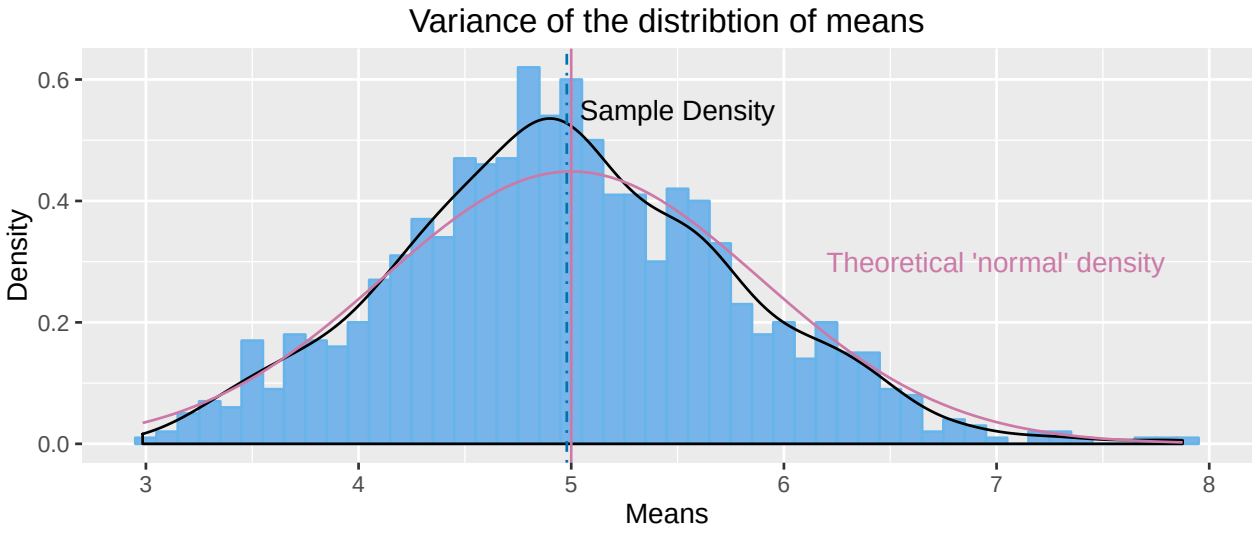
```
##                Mean  Variance Standard Deviation
## Sample      4.979186 0.6379013          0.7986872
## Theoretical 5.000000 0.6250000          0.7905694
```

Conclusion: From the summary table, we can see that the Sample variance 0.6379013 is pretty close to the theoretical variance of 0.625.

3. Show that the distribution is approximately normal.

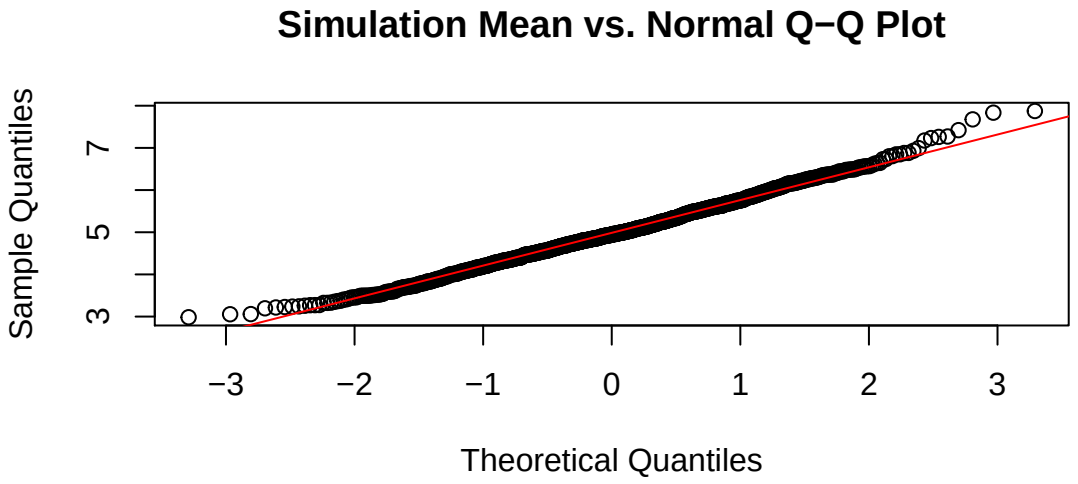
In order to assess the 'normality' of our exponential distribution, we can first plot our observation's density with the one of a normal in our previous graph to see how our data is cented. As a second step We will use the qqnorm and qqline function in R, which will result in a plot comparing our dataset of 1000 samples of 40 exponentials (on the y axis) with a standard normal distribution (on the x axis). If our data set is normally distributed, then the data points should fall along the line produced by the qqline function.

Figure 2 (See appendix for R code)



QQ Norm Plot

```
qqnorm(simMean, main ="Simulation Mean vs. Normal Q-Q Plot")
qqline(simMean, col = "2")
```



Conclusion: From the QQNorm plot, we can see that our average calculations from the exponential distribution are centered around the Normal QQ line as per the Central Limit theorem. We can assume that by increasing the number of simulation from 1000 to 10000 the result would probably look much more aligned.

Based on these calulculation of mean, standard deviation and variance, we can infer that the averages of the *exponential distribution* follow the princicple of the *Centrel Limit Theorem* in regards of following a *normal distribution.*

Appendix

Figure 1 R code

```
# Plot the mean of 40 exponentials
dtMean <- data.frame(simMean)
g <- ggplot(data = dtMean, aes(x = simMean)) +
  geom_histogram(binwidth=0.1, fill = "#77B4E9", aes(y=..density..))+
  labs(x="Means") + labs(y="Density")+
  ggtitle("The distribution of exponential means")+ geom_density() +
  geom_vline(xintercept = sample_mean, size = 1, color ="#0072B2", linetype = 4)+
  annotate("text", x = 5.75, y = 0.55, label = "Sample Mean", color = "#0072B2")+
  geom_vline(xintercept = theoretical_mean, size = 1, color = "#CC79A7")+
  annotate("text", x = 4.25, y = 0.6, label = "Theoretical Mean", color = "#CC79A7")
g
```

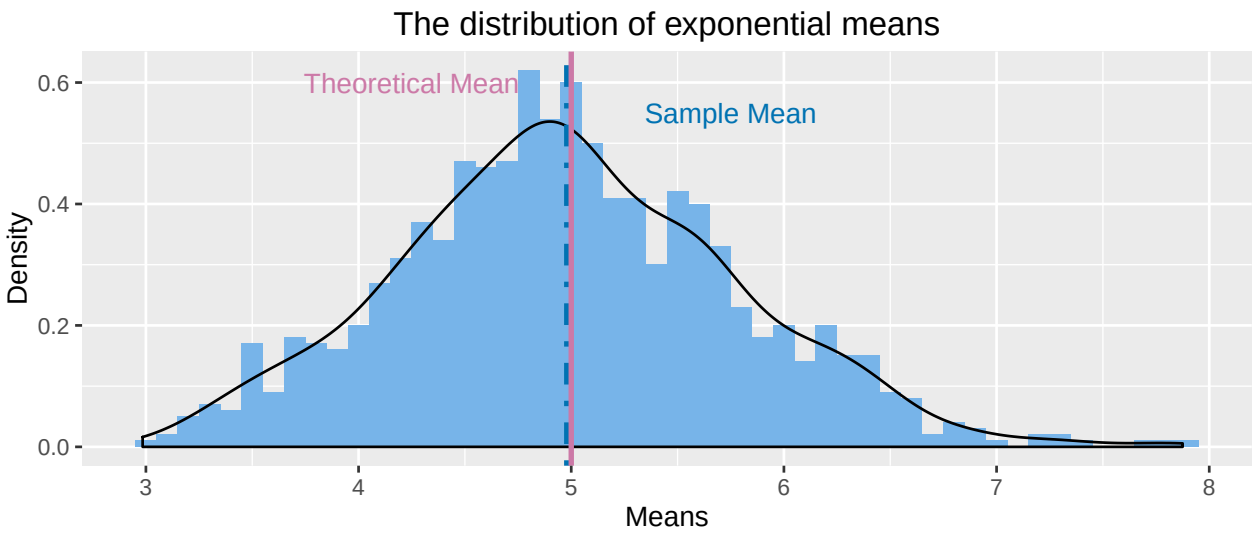
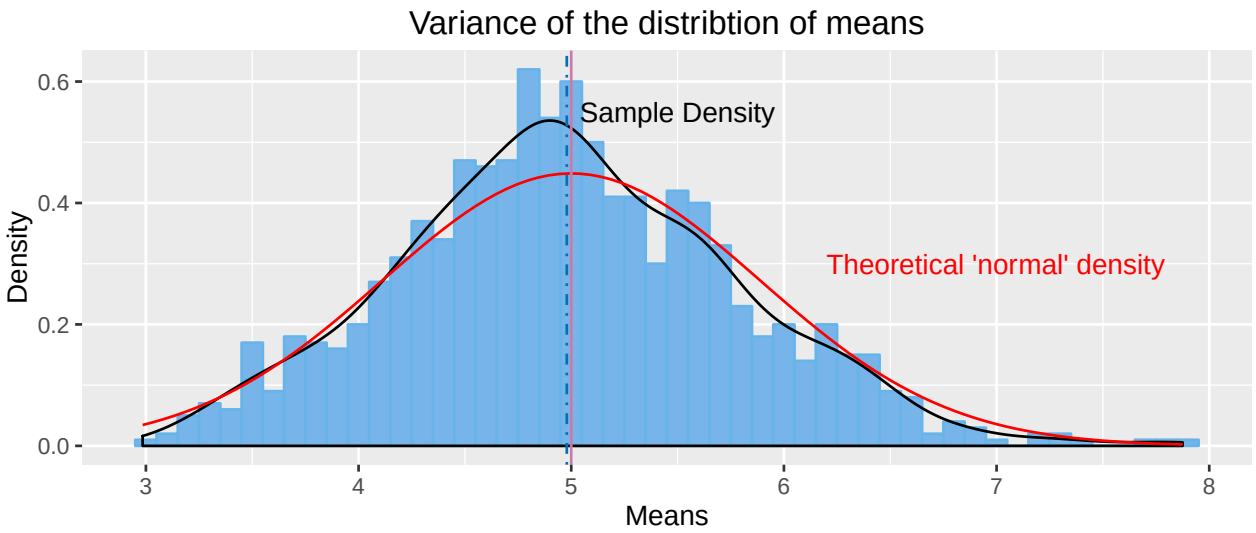


Figure 2 R code

```
g <- ggplot(data = dtMean, aes(x = simMean)) +
geom_histogram(binwidth=0.1, fill = "#77B4E9", color = "#66B4E9", aes(y=..density..))+ geom_density() +
stat_function(fun = dnorm, args=list(mean=1/lambda, sd=sqrt(theoretical_sd)), colour = "red")+
labs(x="Means")+
labs(y="Density")+
ggtitle("Variance of the distribtion of means")+
geom_vline(xintercept = sample_mean, size = 0.5, color ="#0072B2", linetype = 4)+
annotate("text", x = 5.5, y = 0.55, label = "Sample Density")+
geom_vline(xintercept = theoretical_mean, size = 0.5, color = "#CC79A7")+
annotate("text", x = 7, y = 0.3, label = "Theoretical 'normal' density", color = "red")
g
```



Disclaimer

This project is part of the Statistical Inference Course developped by the Johns Hopkins School of Public Health and presently available on Coursera: link.