Comparaison of exponential distribution and Central Limit Theorem

Tomasz Jaskula

26 avril 2016

Overview

The goal of this project is to investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with `rexp(n, lambda)` where lambda is the rate parameter. The mean of exponential distribution is `1/lambda` and the standard deviation is also `1/lambda`. Set `lambda = 0.2` for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

The result of this project will be to illustrate the properties of the distribution of the mean of 40 exponentials and:

- Show the sample mean and compare it to the theoretical mean of the distribution
- Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution
- Show that the distribution is approximately normal

Analysis

This section explores several aspects of exponential distibution.

Load libraries

```
library(ggplot2)
```

Sample mean vs theoretical mean

The first step is to explore the sample mean and compare it to the theoretical mean. In order to have a significant data we would need to run 1000 simulations each containing 40 observations so we would be able to take means for each simulation. We can achieve it by runing the function `rexp(1000 * 40, 0.2)` and arranging it in the matrix of 100x40:

```
n <- 40 # number of observations
nbsim <- 1000 # number of simulations
lambda <- 0.2

set.seed(1111)

m <- matrix(rexp(n * nbsim, lambda), nbsim, n)
means <- apply(m, 1, mean)
```

Let's check what is the mean of 1000 averages of exponential distribution:

```r
mean(means)
```

```
## [1] 4.994883
```

Let's compare it to the theoretical mean of exponential distribution

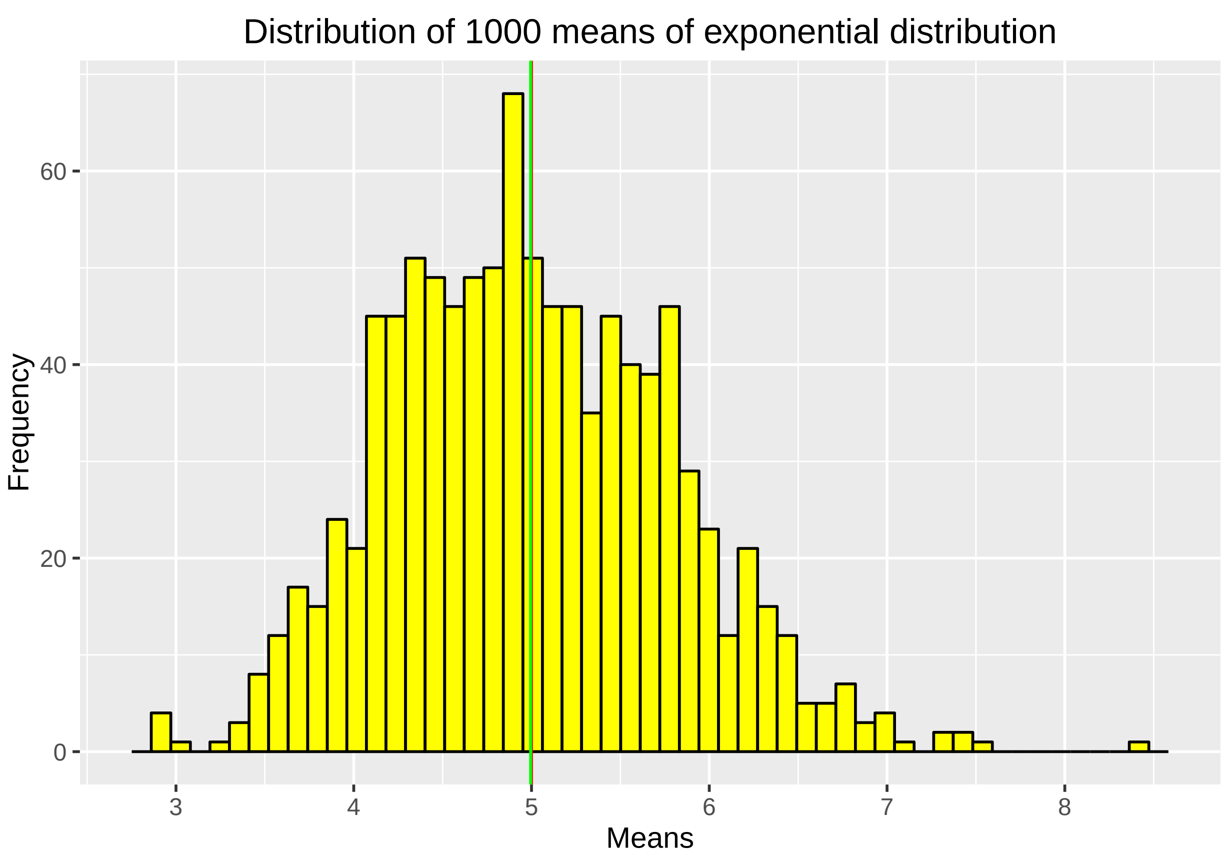
```r
theoretical_mean <- 1/lambda
theoretical_mean
```

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

The simulated mean of exponential distribution `4.994883` is very close to the therotetical mean `5`.

Next step is to look at the distribution of the simulated means.

```r
ggplot() +
  aes(means) +
  geom_histogram(bins=50, colour="black", fill="yellow") +
  geom_vline(xintercept = 5, colour = "red") +
  geom_vline(xintercept = 4.994883, colour = "green") +
  ggtitle("Distribution of 1000 means of exponential distribution") +
  xlab("Means") +
  ylab("Frequency")
```



Sample variance vs theoretical variance

The next step is to compare the variance present in the sample means of the 1000 simulations to the theoretical varience of the population.

The sample variance is

```
variance <- var(means)
variance
```

```
## [1] 0.6278933
```

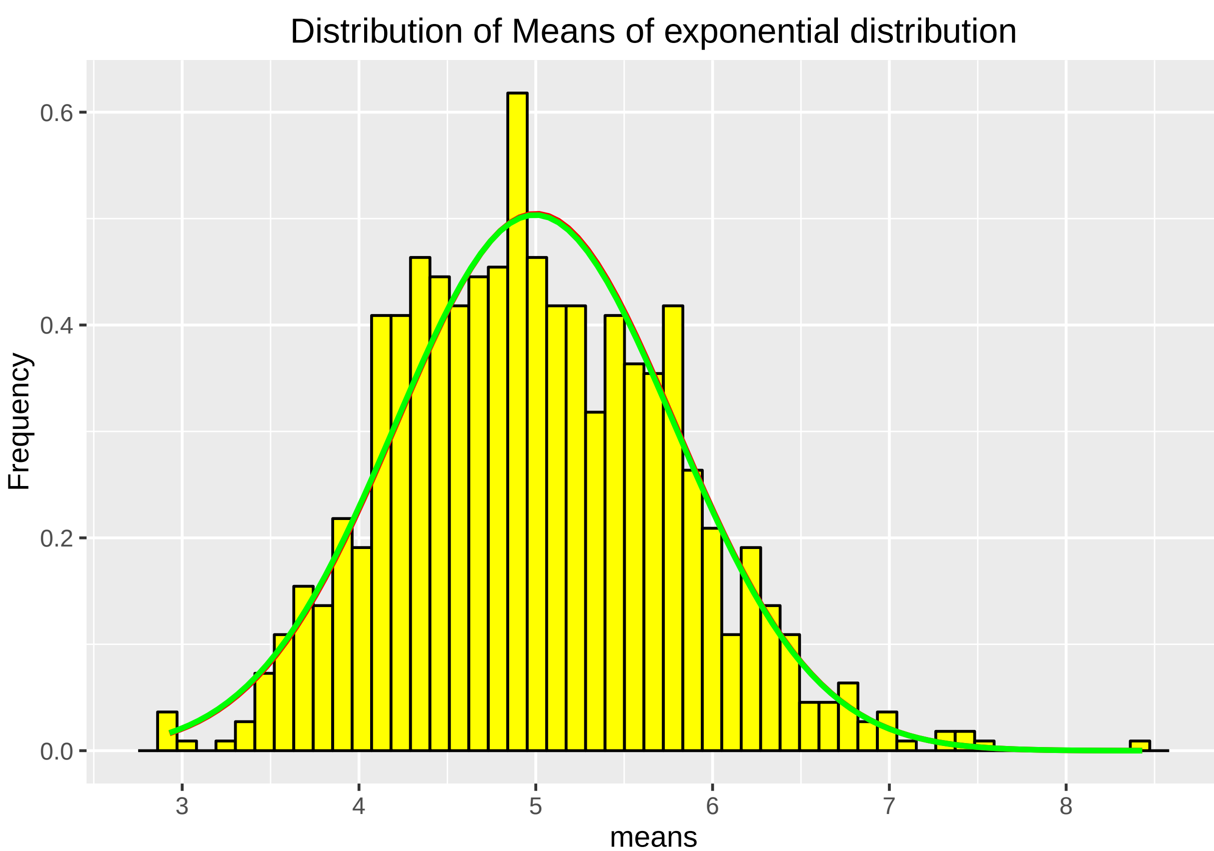
And the theoretical is

```
theoretical_variance <- ((1/lambda)^2)/n
theoretical_variance
```

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

The theoretical and sample variance are quite similar

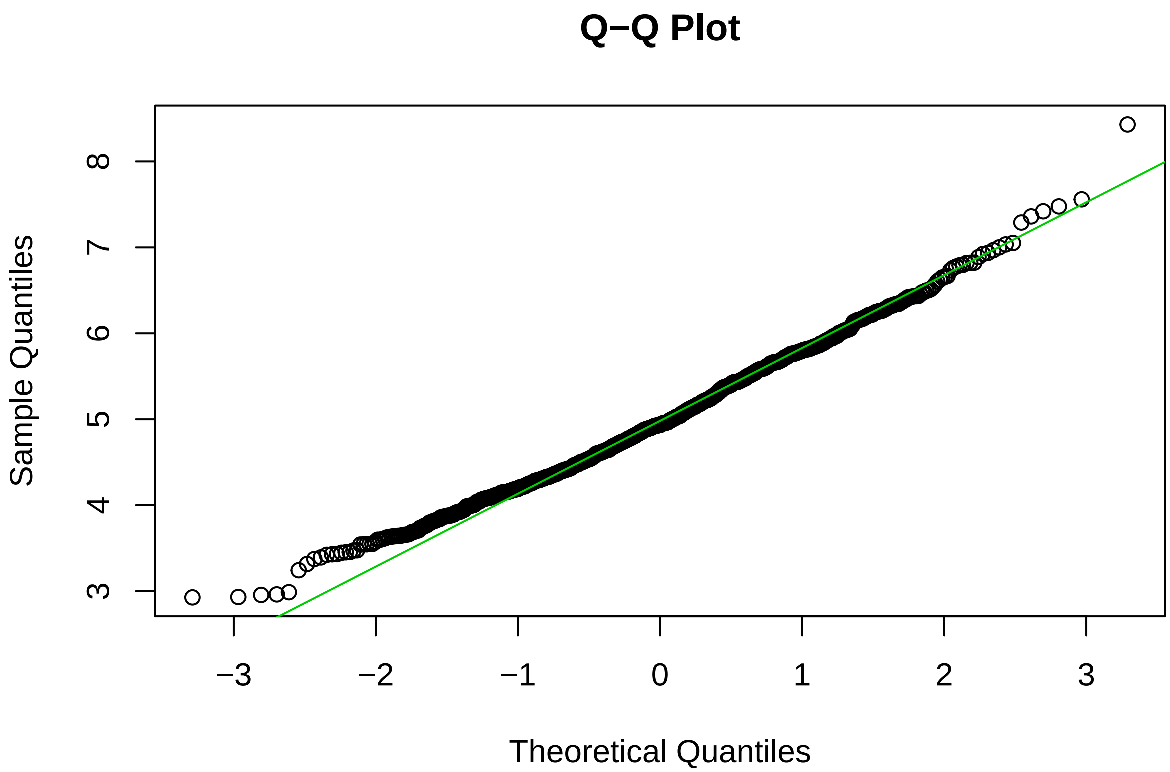
```
df <- data.frame(means)
ggplot(df,aes(x = means)) +
  geom_histogram(aes(y=..density..), bins=50, colour="black",fill="yellow") +
  labs(title="Distribution of Means of exponential distribution", y="Frequency") +
  stat_function(fun=dnorm,args=list( mean=1/lambda, sd=sqrt(theoretical_variance)), color = "red", size
  stat_function(fun=dnorm,args=list( mean=mean(means), sd=sqrt(variance)), color = "green", size = 1.0)
```



Show that the distribution is approximately normal

Because of the CLT, the means of exponential samples follow normal distribution. We can demonstrate in on the q-q plot which suggests that the means of samples of exponential distribution match the theoretical quantiles.

```
qqnorm(means, main ="Q-Q Plot")
qqline(means, col = "3")
```



Conclusion

We have showed that averaging over 1000 simulations of 40 observations for exponential distribution is very close to the theoretical distribution mean. We can say the same about the variance which is very close. The distribution is also noraml.