Exponential distribution comparison with CLT

Marçal Boix

Nov 21, 2015

In this project we will 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.
We will investigate the distribution of averages of 40 exponentials.
We will set lambda = 0.2 for all of the simulations.

In this document we will do the following tasks:

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

Initialization

In order to complete the project we need to initialize R with a few statements.

```
options(warn=-1)
set.seed(11525)
library(ggplot2)
```

Simulation

We will create a matrix of 1000 rows and 40 columns. This matrix will be used to save the 1000 simulations of 40 samples each one. After that, we will calculate the mean of each simulation and save in a data frame.

```
lambda = 0.2
numberOfSimulations = 1000
sampleSize = 40
simulations = matrix(rexp(numberOfSimulations * sampleSize, rate=lambda),
                     numberOfSimulations, sampleSize)
simulationMeans = data.frame(rowMeans(simulations))
names(simulationMeans) = c("mean")
```

Mean comparison. Sample Mean versus Theoretical Mean.

Expected mean 'mu' of a exponential distribution with rate 'lambda' is: mu = 1/lambda

```
mu = 1/lambda
mu
```

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

Average of 1000 simulation means:

```
meanOfSimulationMeans = mean(simulationMeans[,1])
meanOfSimulationMeans
```

```
## [1] 4.962325
```

As expected, two values are very close: 5.00 and 4.96.

Variance comparison. Sample Variance versus Theoretical Variance.

First of all we will calculate standard deviation.
Expected standard deviation 'sd' of a exponential distribution with rate 'lambda' is: sd = (1/lambda)/sqrt(n)

```
sd= (1/lambda)/sqrt(sampleSize)
sd
```

```
## [1] 0.7905694
```

And the variance is 'sd^2':

```
var = sd^2
var
```

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

We will calculate the variance for the 1000 simulations

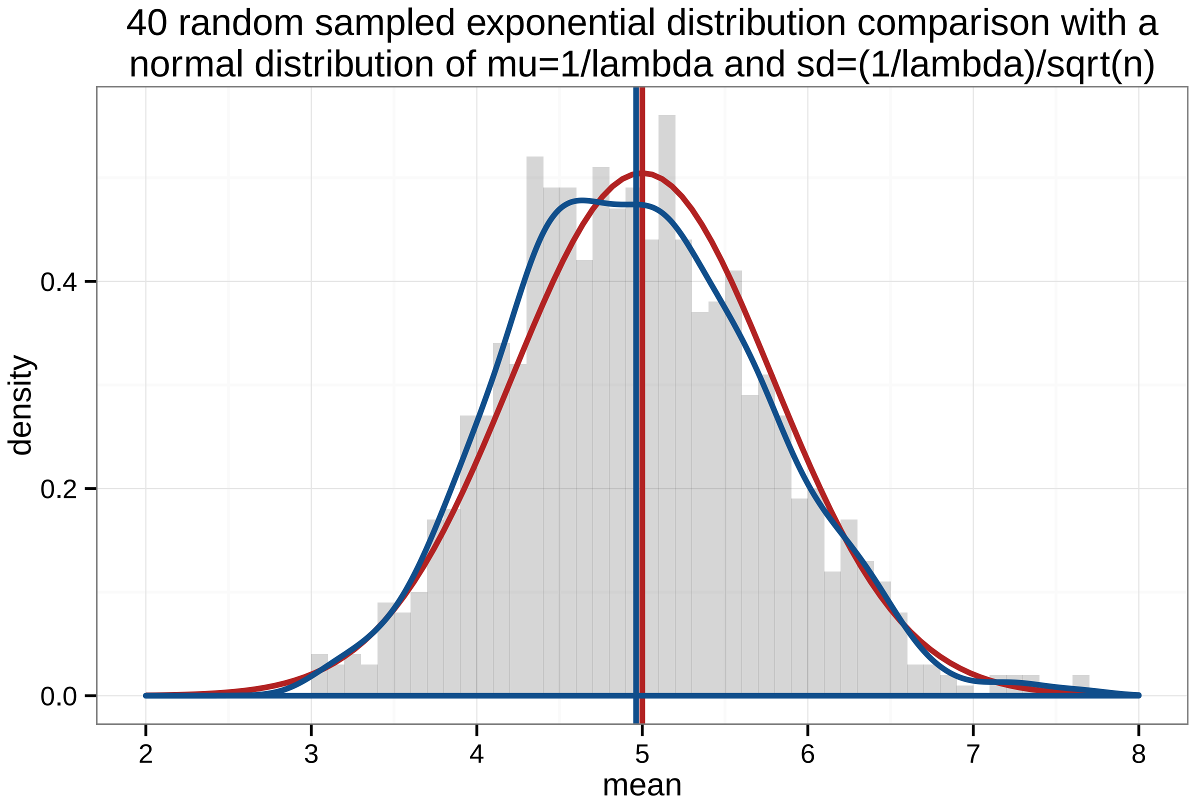
```
varSimulation = var(simulationMeans)
varSimulation
```

```
##           mean
## mean 0.6106535
```

Distribution

The following graph shows the calculated distribution of means of 40 random sampled exponential distributions and the normal distribution with mean=mu and sd=(1/lambda)/sqrt(n).

```
strTitle = "40 random sampled exponential distribution comparison with a"
strTitle = paste(strTitle, "normal distribution of ", sep="\n")
strTitle = paste(strTitle, "mu=1/lambda and sd=(1/lambda)/sqrt(n)", sep="")
ggplot(data = simulationMeans, aes(x = mean)) +
  geom_histogram(binwidth=0.1, aes(y=..density..), alpha=0.2) +
  stat_function(fun = dnorm, arg = list(mean = mu , sd = sd), colour = "firebrick", size=1) +
  geom_vline(xintercept = mu, size=1, colour="firebrick") +
  geom_density(colour="dodgerblue4", size=1) +
  geom_vline(xintercept = meanOfSimulationMeans, size=1, colour="dodgerblue4") +
  scale_x_continuous(breaks=seq(mu-3,mu+3,1), limits=c(mu-3,mu+3)) +
  ggtitle(strTitle) +
  theme_bw()
```



Blue line is the distribution of the simulation means.
Red line is the normal distribution.
The sample distribution is approximately normal as we see in the graph.
This is due to the **central limit theorem**: averages of samples follow normal distribution.