Investigation of the CLT using 1000 simulated samples from random exponential distributions

Bert Lijnen

10 August 2017

Summary

In this report we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. Specifically, we will investigate the theoretical mean and the overall mean of our samples. We do the same thing for the variance and we will check whether our simulated data will be bell shaped as expected according to the CLT.

Simulation of an exponential distribution

In this section we will simulate an exponential distribution with rate parameter lamba equal to 0.2. We will run 1000 simulations and store the data in a matrix. Each of the 1000 simulations will contain 40 observations. To run the simulation in R we use the funcion rexp() with arguments n and lambda. To guarantee that the results are reproducible we set seed.

```
set.seed(1234)
lambda<-0.2
n<-40
sim<-1000
sim_matrix<-matrix(data=rexp(sim*n, lambda), nrow=sim, ncol=n)
```

We check the results.

```
dim(sim_matrix)
```

```
## [1] 1000   40
```

```
sim_matrix[1:10,1:2]
```

```
##              [,1]      [,2]
##  [1,] 12.50879302 3.7841509
##  [2,]  1.23379442 0.7843467
##  [3,]  0.03290978 7.7705637
##  [4,]  8.71373045 3.6451603
##  [5,]  1.93591292 0.0983085
##  [6,]  0.44974836 0.5168639
##  [7,]  4.12040757 4.6273571
##  [8,]  1.01308950 1.5763302
##  [9,]  4.19020160 3.6735403
## [10,]  3.80215150 6.3464934
```

The simulation has been run correctly since we have a matrix with 1000 rows (number of simulations) and 40 columns (= sample size of each simulation).

1. Sample mean and theoretical mean

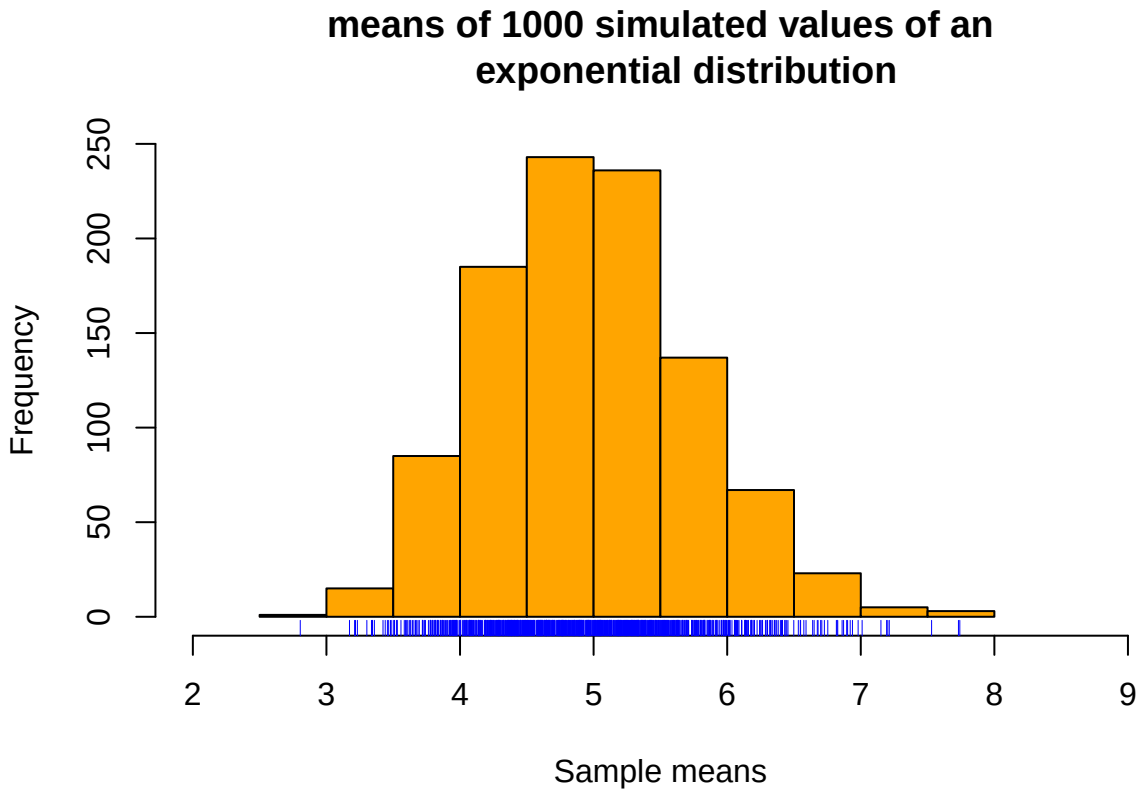
We now compute the mean for each simulation. Since each simulation represents a row in our matrix, we can easily find it by using the rowMeans() function. We store the results in an object sample_mean_sim

```
sample_mean_sim<-rowMeans(sim_matrix)
sample_mean<-mean(sample_mean_sim)
print(sample_mean)
```

```
## [1] 4.974239
```

Let's check if the mass is concentrated around 5 when plotting the 1000 means

```
hist(sample_mean_sim, col="orange", main="means of 1000 simulated values of an
     exponential distribution",xlab="Sample means",
     xlim=c(2,9), ylim=c(0,250))
rug(jitter(sample_mean_sim, amount=0.01), col="blue")
```



The rug and the bars clearly show that the mass is concentrated in the interval [4;6].

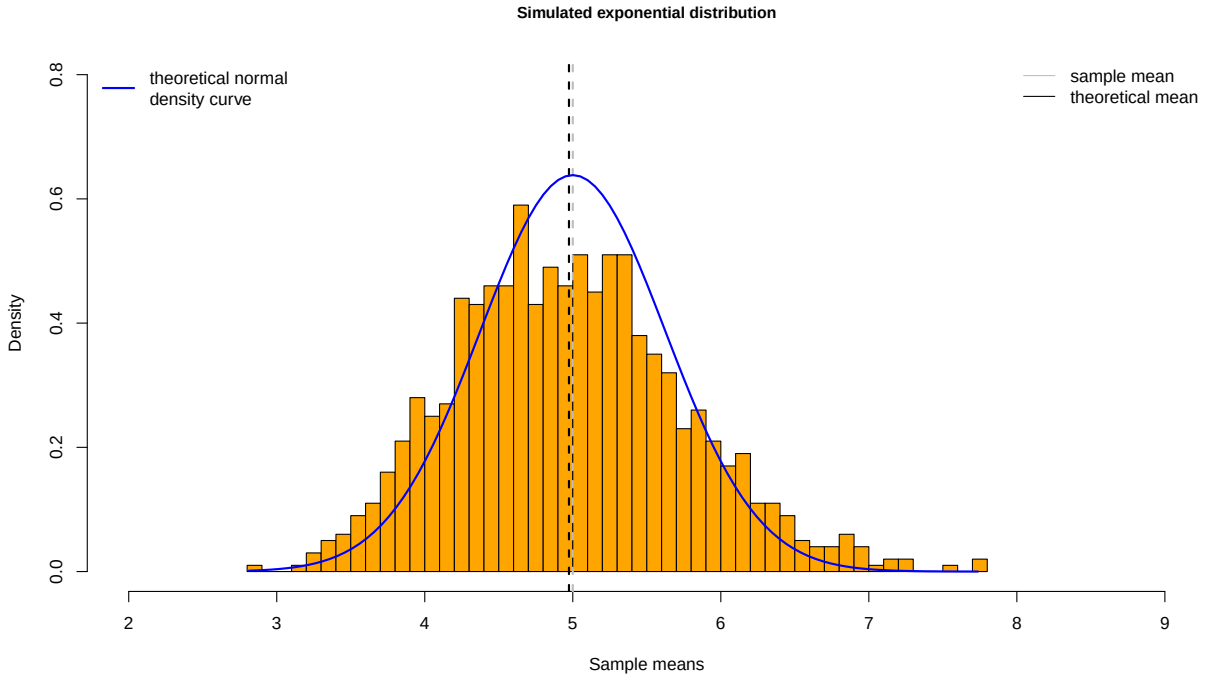
Now we want to compare the sample mean with the expected value or theoretical mean which for an exponential distribution equals 1/lambda.

```
exp_value<-1/lambda
print(exp_value)
```

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

We conclude that both means are nearly identical.The plot below shows that the means are indeed very close.

```r
hist(sample_mean_sim, col="orange",xlab="Sample means", main="", xlim=c(2,9),
     ylim=c(0,0.8), breaks=50, prob=TRUE)
title(main="Simulated exponential distribution", cex.main=0.9)
abline(v=sample_mean, col="black", lty=2, lwd=2)
abline(v=exp_value, col="grey", lty=2, lwd=2)
legend("topright", c("sample mean", "theoretical mean"), bty="n", lty=c(1,1),
       col=c("grey", "black"))
xfit<-seq(min(sample_mean_sim), max(sample_mean_sim), length=100)
yfit<-dnorm(xfit, mean=1/lambda,sd = ((1/lambda)^2)/n)
lines(xfit, yfit, pch=20, col="blue", lwd=2)
legend("topleft", c("theoretical normal\ndensity curve"), lty=1, lwd=2, bty="n",
       col="blue")
```



2. Sample variance of the distribution of means and theoretical variance of the distribution of means

First we compute the variance of the mean and the theoretical variance of the mean.

```r
sample_variance<-var(sample_mean_sim)
theo_variance<-((1/lambda)^2)/n
print(sample_variance)
```

```
## [1] 0.5949702
```

```r
print(theo_variance)
```

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

Both values are close, but the sample variance is higher than the theoretical variance.

3. Variance of the simulated exponentials

Now we have a look at the variances in the different samples. This means that we cannot use the aggregated data. Therefore, we write a function to calculate the variance of each line of the matrix with simulated exponentials allowing us to plot the histogram. In order to compare it with the theoretical variance we recalculate the latter without dividing it by n.

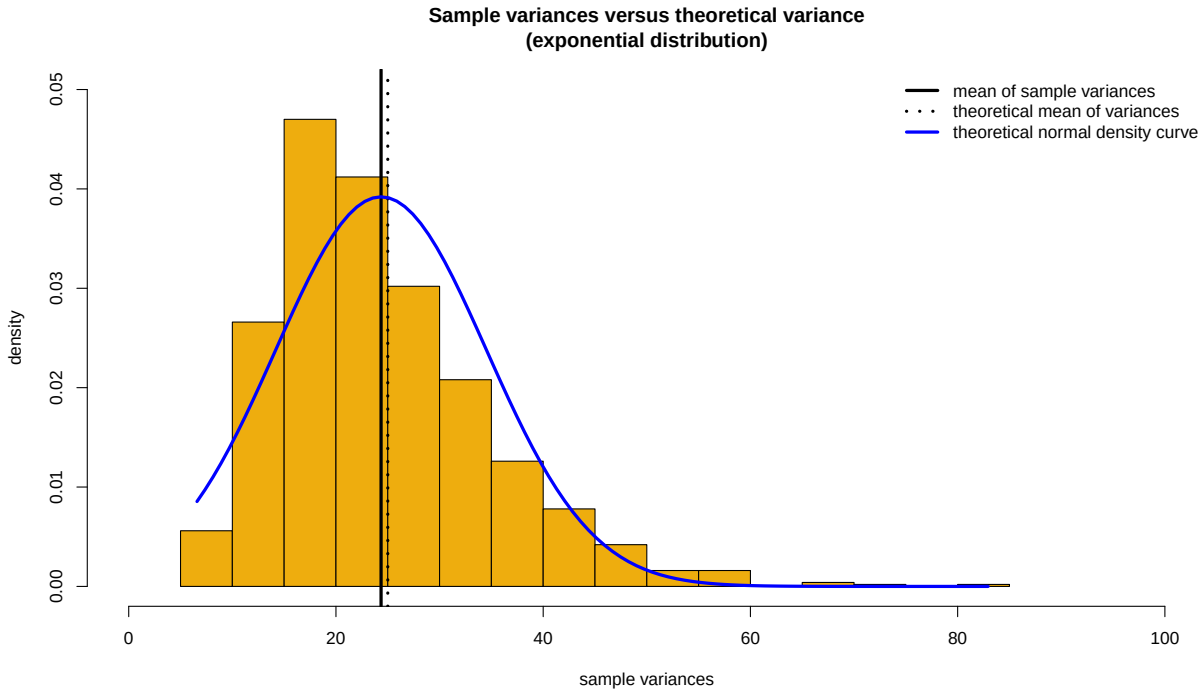
```
RowVar <- function(x) {
  rowSums((x - rowMeans(x))^2)/(dim(x)[2] - 1)
}
vars<-RowVar(sim_matrix)
head(vars)
```

```
## [1] 11.84129 36.35519 35.92965 16.79182 56.71738 13.67719
```

```
sample_var2<-mean(vars)
theo_var2<-((1/lambda)^2)
print(theo_var2)
```

```
## [1] 25
```

```
hist(vars, xlab="sample variances", ylab="density", col="darkgoldenrod2", prob=TRUE,
     main="Sample variances versus theoretical variance\n(exponential distribution)",
     xlim=c(0,100), ylim=c(0,0.05))
abline(v=sample_var2, col="black", lwd=3)
abline(v=theo_var2, col="black", lwd=3, lty=3)
xfit<-seq(min(vars), max(vars), length=100)
yfit<-dnorm(xfit, mean=sample_var2,sd=sd(vars))
lines(xfit, yfit, pch=20, col="blue", lwd=3)
legend("topright", c("mean of sample variances", "theoretical mean of variances",
                     "theoretical normal density curve"),
       bty="n", lty=c(1,3, 1), lwd=c(3,3, 3), col=c("black", "black", "blue"))
```



We can see that the distribution of the sample variances is centered at the same location as the theoretical variance.

4. Distribution

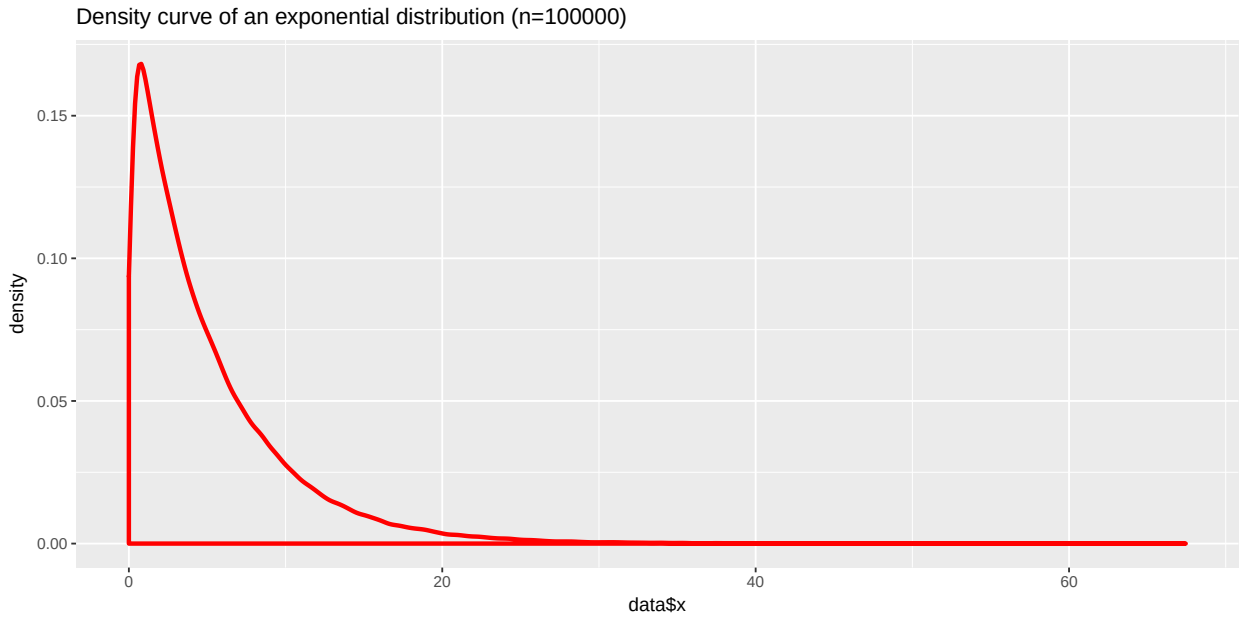
The Central Limit Theorem states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population. Furthermore, all of the samples will follow an approximate normal distribution pattern, with all variances being approximately equal to the variance of the population divided by each sample's size.

We have seen that the sample mean (5.002) is very close to the expected value or population mean(5). The histogram also suggests a normal distribution of the mean values. Furthermore, we have seen that the variance of the mean is close to the theoretical variance of means.

The distribution of the sample variances shows that the theoretical and the overall sample variance nearly coincide.

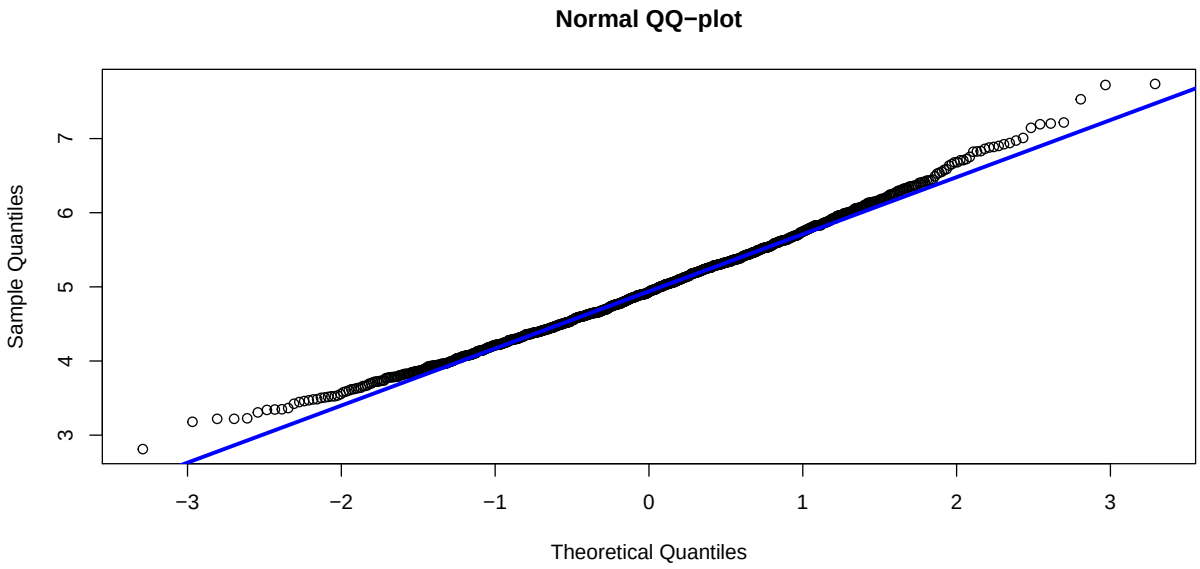
Our simulated exponential distribution looks approximately normal. This becomes obvious in the histograms overlaid with a theoretical normal density curve. The difference with a one-sample exponential distribution is clearly illustrated by the following plot:

```
library(ggplot2)
data<-data.frame(x=rexp(n=100000, rate=0.2))
ggplot(data, aes(x=data$x))+geom_density(col="red", lwd=1.2)+
        ggtitle("Density curve of an exponential distribution (n=100000)")
```



Let's also make a QQ-plot to investigate whether our simulated distribution approximates the gaussian distribution.

```
qqnorm(sample_mean_sim, main="Normal QQ-plot")
qqline(sample_mean_sim, col="blue", lwd=3)
```



The qqplot suggests that our simulated exponential distribution is indeed approximately normal since the theoretical quantiles match closely with the actual quantiles.