Exponential Distribution and Central Limit Theorem

Christophe Chevalier

May 2015

Overview

Using R we investigate the exponential distribution by doing some simulations. We first generate a 1000 sets of 40 exponentials each and then analyze the resulting distribution of the mean of each set (sample mean and sample variance). The distribution of the mean of 40 exponentials for a 1000 simulations is shown to be normal in accordance with the Central Limit Theorem.

1. A 1000 simulations of 40 exponentials

```
library(ggplot2)
set.seed(8765432)
# Exponential distribution and simulations parameters
lambda  <- 0.2
n <- 40
nosim <- 1000
```

We perform the request simulations using the R function *rexp(n, lambda)* and store the results in a matrix (*simus*), each row containing a sample of 40 exponentials. We then compute the mean value of each row in order to get the required distribution of the mean of 40 exponentials, $\bar{X}$, for a 1000 simulations (vector *sim_mean*)

```
simus <- matrix(rexp(nosim * n, lambda), ncol = n, byrow = TRUE)
sim_mean <- apply(simus, 1, mean)
```

A demo of a single sample distribution of 40 exponentials is given in Appendix A.

2. Central Limit Theorem

According to the Central Limit Theorem (CLT), the distribution of sample means, $\bar{X}$, is approximately normal with mean = μ and variance = σ^2/n where μ and σ^2 are respectively the mean and the variance of the original distribution (n being the number of samples).

2.1 Sample Mean versus Theoritical Mean

In this study case, the mean of the original exponential distribution is: $\mu = 1/\lambda$

```
# Sample Mean: CLT Theory vs Simulated
c(1/lambda, mean(sim_mean))
```

```
## [1] 5.00000 5.02662
```

2.2 Sample Variance versus Theoritical Variance

In this study case, the variance of the original exponential distribution is: $\sigma^2 = (1/\lambda)^2$

```
# Variance: CLT Theory vs Simulated
c((1/lambda)^2 /n, var(sim_mean))
```

```
## [1] 0.6250000 0.6206097
```

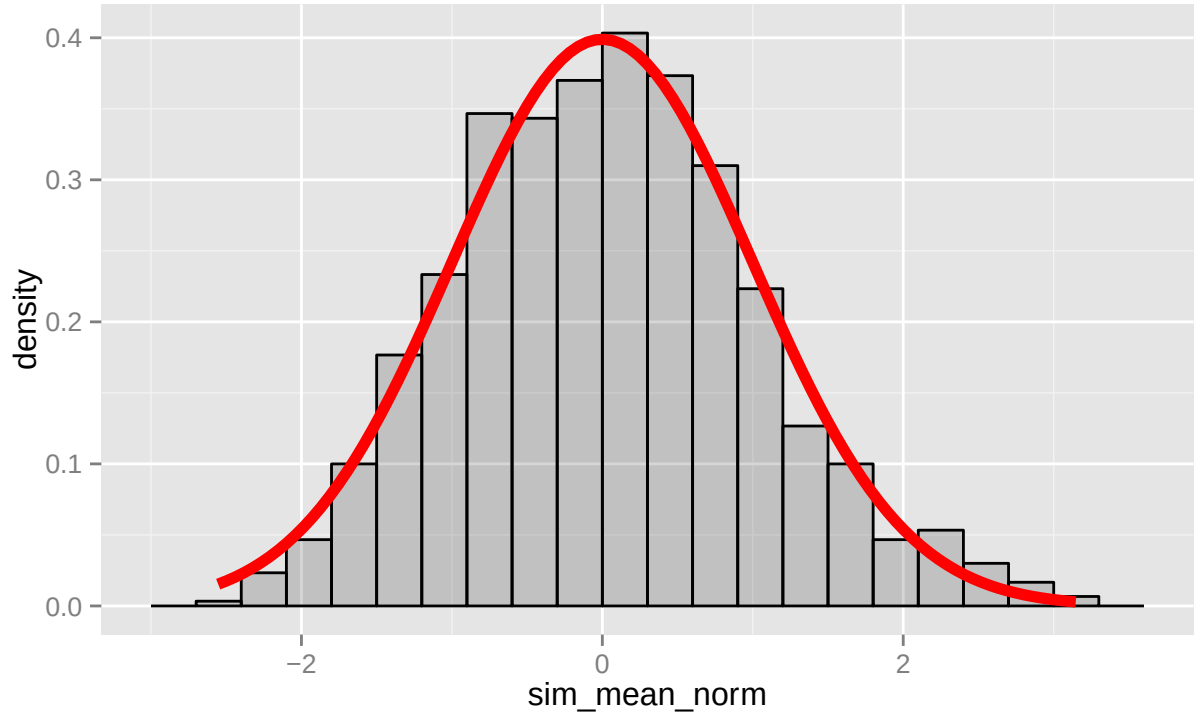
Both the sample mean and the sample variance are in agreement with their respective theoritical values.

3. Distribution

3.1 Histogram of the normalized distribution of the means of 40 exponentials

The histogram of the 1000 simulated means of 40 exponentials is given in Appendix B in its non-normalized form (original). Here we normalized the resulting distribution and compare its histogram to the standard normal distribution.

```
# CLT normalize function applied to sim_mean
clt_func <- function(x, n) (mean(x) - 5) / (5 / sqrt(n))
sim_mean_norm <- apply(simus, 1, clt_func, n)

df_sim_mean_norm <- data.frame(sim_mean_norm)
h2 <- ggplot(df_sim_mean_norm, aes(x = sim_mean_norm))
h2 <- h2 + geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..density..))
h2 <- h2 + stat_function(fun = dnorm, size = 2, colour = "red")
h2
```

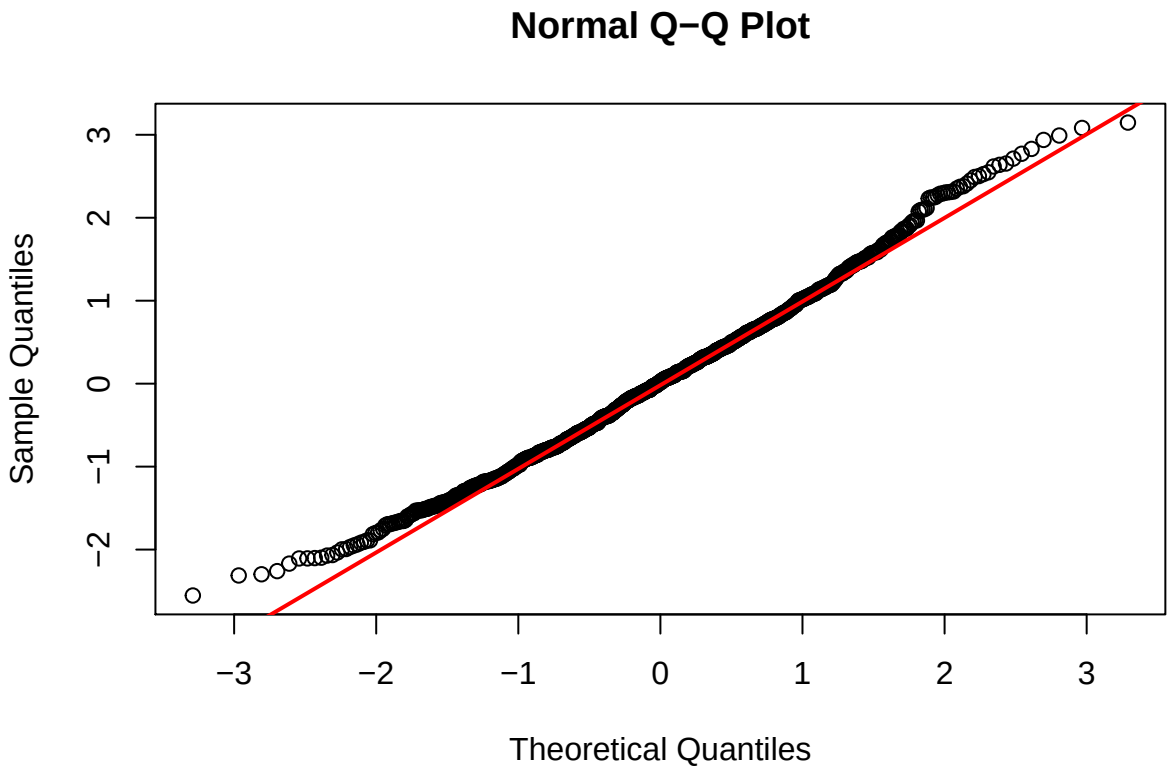


The resulting normalized distribution (histogram plot) and the standard normal distribution (red curve) are in close agreement.

3.2 Q-Q plot

As suggested in the *hist* online help it is recommended to compare data with a model distribution using the *qqplot* command. This is illustrated in the graph below (using the normalized distribution of the means of 40 exponentials)

```
qqnorm(sim_mean_norm)
qqline(sim_mean_norm, col = "red", lwd = 2)
```



The agreement between the quantiles of normalized data (Sample Quantiles) and the ones of the standard normal distribution (Theoritical Quantiles) is rather good.

Appendices

A. Demo of a sample distribution of 40 exponentials

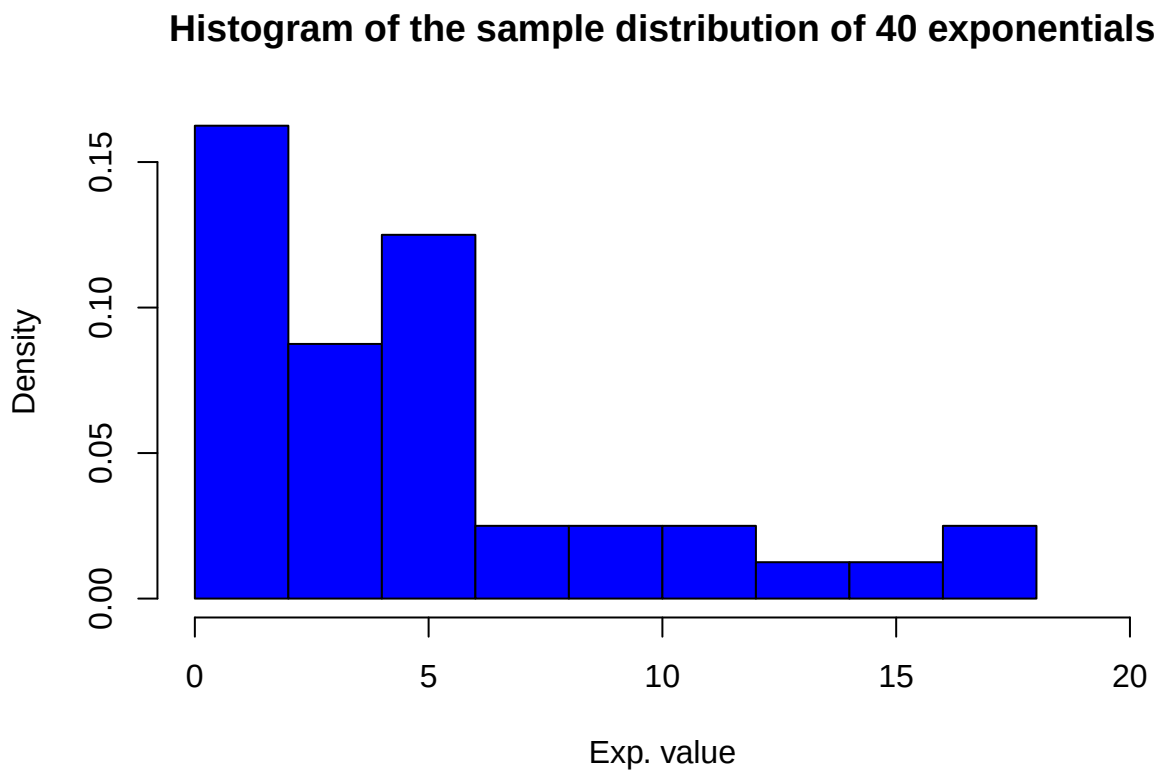
```
demo_exp  <- rexp(n, lambda)

# Mean of the demo distribution: Theory vs Computed demo
c(1/lambda, mean(demo_exp))
```

```
## [1] 5.000000 5.138034
```

```
# Standard deviation of the demo distribution: Theory vs Computed demo
c(1/lambda, sd(demo_exp))
```

```
## [1] 5.000000 4.618291
```

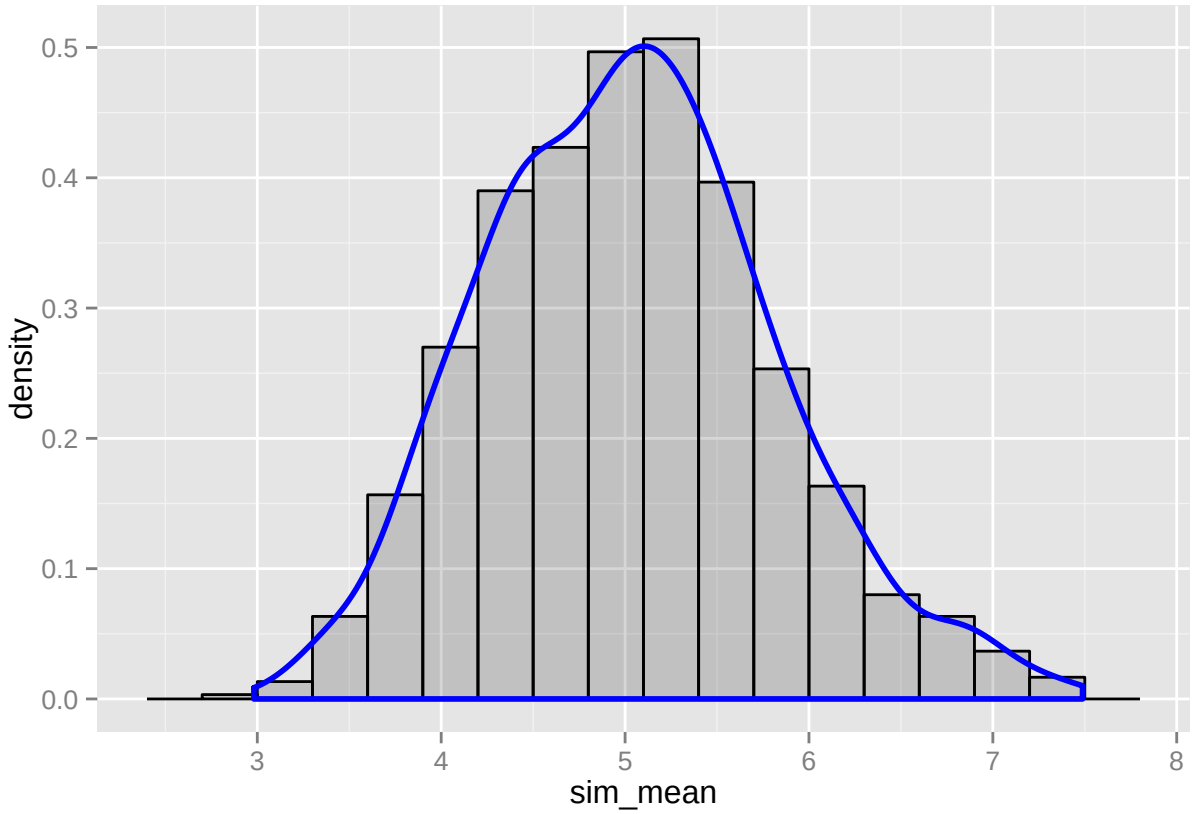
```
# Histogram
hist(demo_exp, freq = FALSE, col = "blue",
     xlim = c(0, 20),
     xlab = "Exp. value",
     main =  "Histogram of the sample distribution of 40 exponentials")
```



B. Distribution of the means of 40 exponentials

The graph below represents the original distribution of the means of 40 exponentials with its associated density (blue line)

```
df_sim_mean <- data.frame(sim_mean)
h1 <- ggplot(df_sim_mean, aes(x = sim_mean))
h1 <- h1 + geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..density..))
h1 <- h1 + geom_density(size = 1, colour = "blue")
h1
```

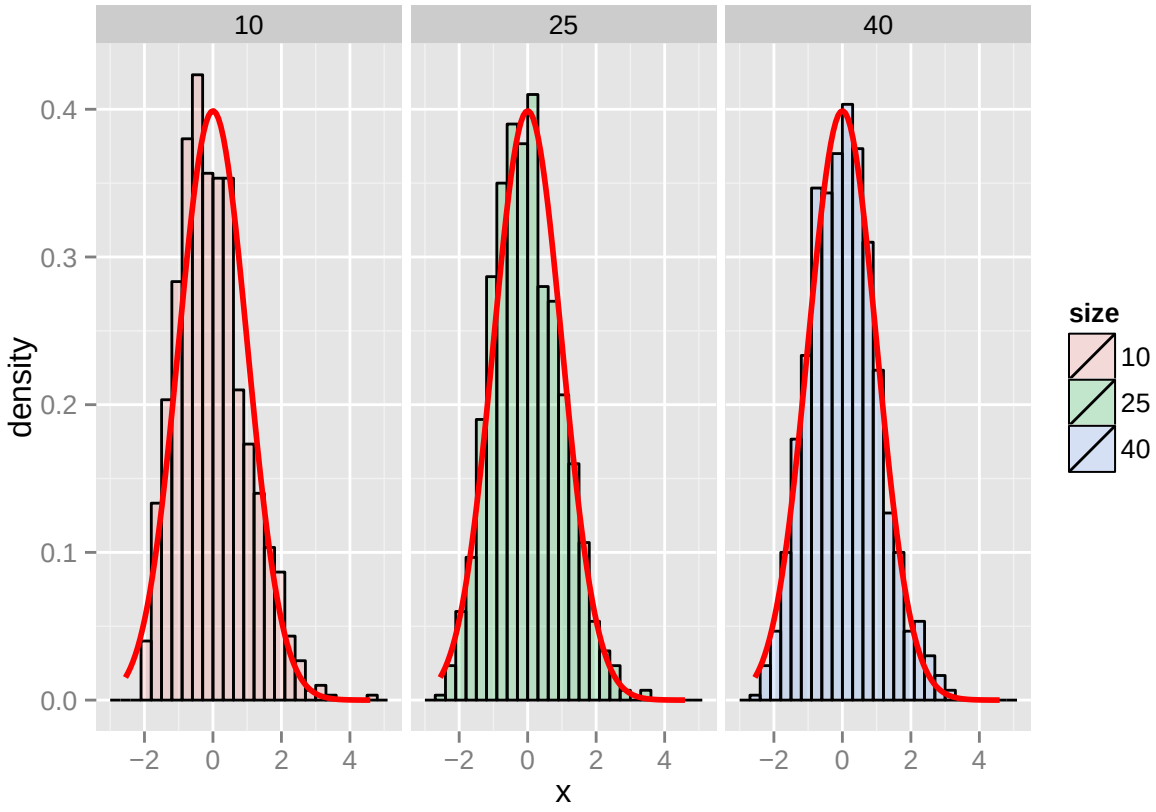


The resulting distribution looks like a normal distribution with mean around 5.

C. Distribution of the means as a function of n

Below is an exploration of different resulting distributions of the means of n exponentials with $n = 10$, 25 or 40 as also performed in the course, lesson 7 (see slide 9/31 for instance).

```
# We reuse the existent simulations with n = 10, 25 or 40. Resulting means are
# normalized using the clt_func defined in 3.1
dat <- data.frame(
  x = c(apply(simus[ , 1:10], 1, clt_func, 10),
        apply(simus[ , 1:25], 1, clt_func, 25),
        apply(simus[ , 1:40], 1, clt_func, 40)),
      size = factor(rep(c(10, 25, 40), rep(nosim, 3)))
  )

g <- ggplot(dat, aes(x = x, fill = size))
g <- g + geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..density..))
g <- g + stat_function(fun = dnorm, size = 1, colour = "red")
g + facet_grid(. ~ size)
```



From these plots we can see that the resulting normalized distribution is getting closer to the standard normal distribution (red curve) with increasing values of n.