CLT - Lectures

Luiz Bergo

03/11/2019

Central Limit Theorem - CLT

1. Introduction

This is a brief review about the important *Central Limit Theory* commom called *CLT*. It is itent to show a practical approach, based on numeric and graphic presentation, so for further studies are highly recommended.

The main idea is to apply the theory on a exponencial distribution and check whether is possible to conclude about the population distribution based on observed data.

The probability densidity function to an exponencial distribution takes form:

$f(x;\lambda) = \lambda e^{-x\lambda}$, $x \geq 0$

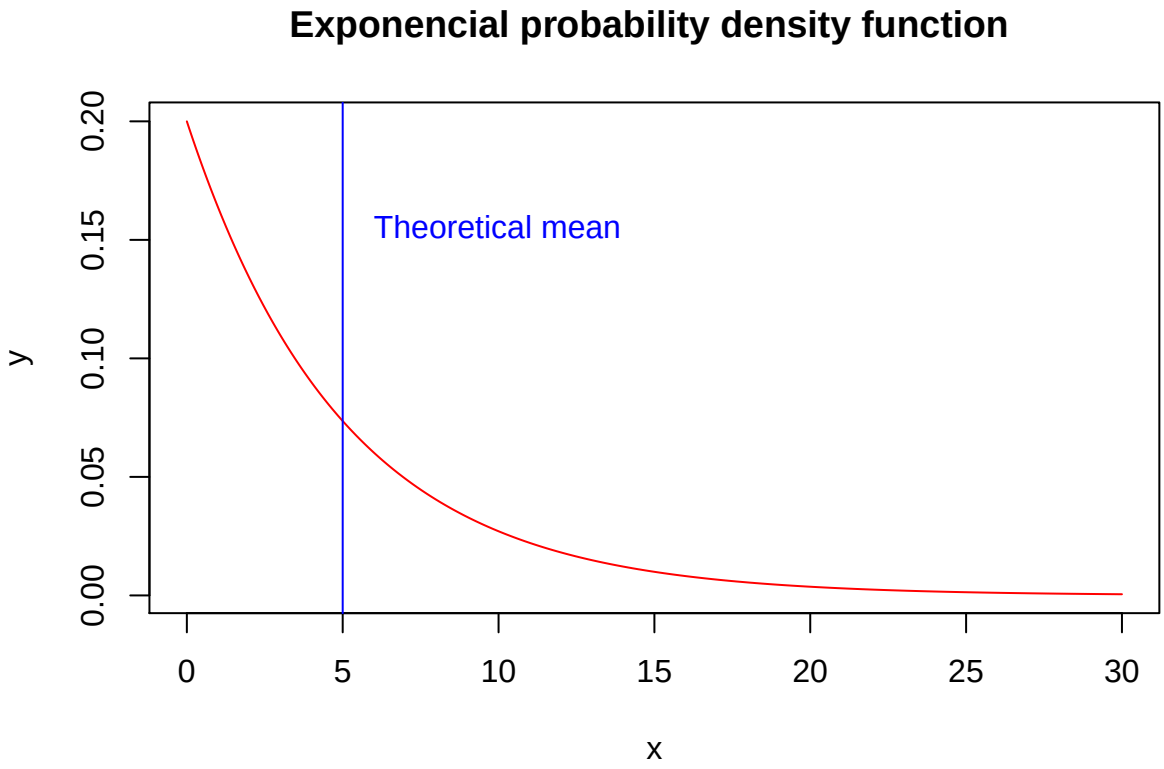
For such distribution we have the population parameter as follow:

$\mu = \sigma = \frac{1}{\lambda}$

In this work simulation assume $\lambda = 0.2$

```
x <- seq(0,30, 0.01)
lambda = 0.2
dist_exp <- function(x,lambda) (lambda)*exp(-x*lambda)
y <- dist_exp(x, lambda)

plot(x,y, type= "l", col="red", main = "Exponencial probability density function ")
abline(v= 1/lambda, col= "blue" )
text(6, 0.15, "Theoretical mean", col = "blue", adj = c(0, -.1))
```



2. Simulations

This section presents the graphtics that lead us to understand some of the main ideas concerning the *Central Limit Theory*.

2.1 Shape of simulated distributions

Three experiments composed by 1000 simulations each were conducted using 10, 20 and 40 samples of exponencial distributions.

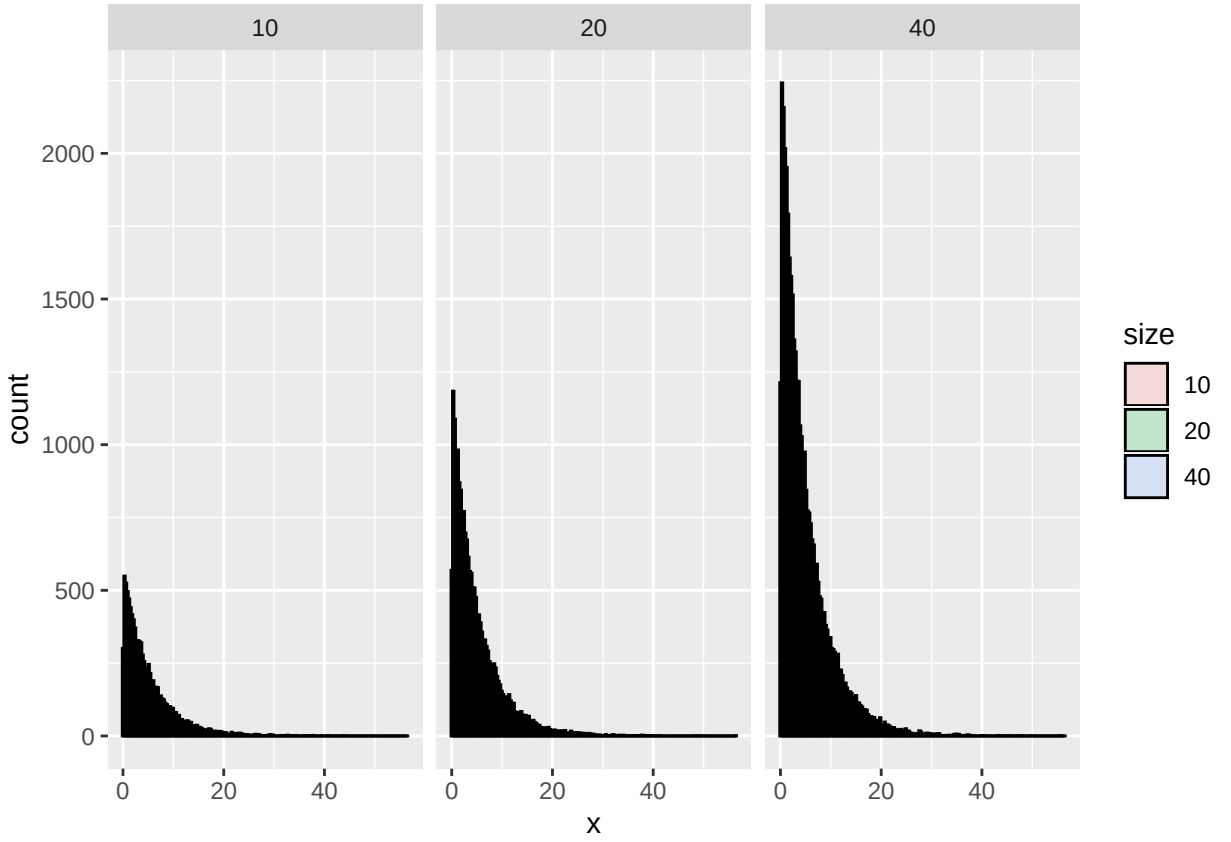
```
lambda = 0.2
nosim = 1000
n1 = 10
n2 = 20
n3 = 40
set.seed(63)

data1 <- rexp(nosim*n1,lambda)
data2 <- rexp(nosim*n2,lambda)
data3 <- rexp(nosim*n3,lambda)


data <- data.frame(
  x = c(data1, data2, data3),
  size = factor(rep(c(n1,n2,n3),c(length(data1),length(data2),length(data3))))
```

```
)

g <- ggplot(data, aes(x = x, fill = size)) + geom_histogram(alpha = .20, binwidth=.3, colour = "black",
g + facet_grid(. ~ size)
```

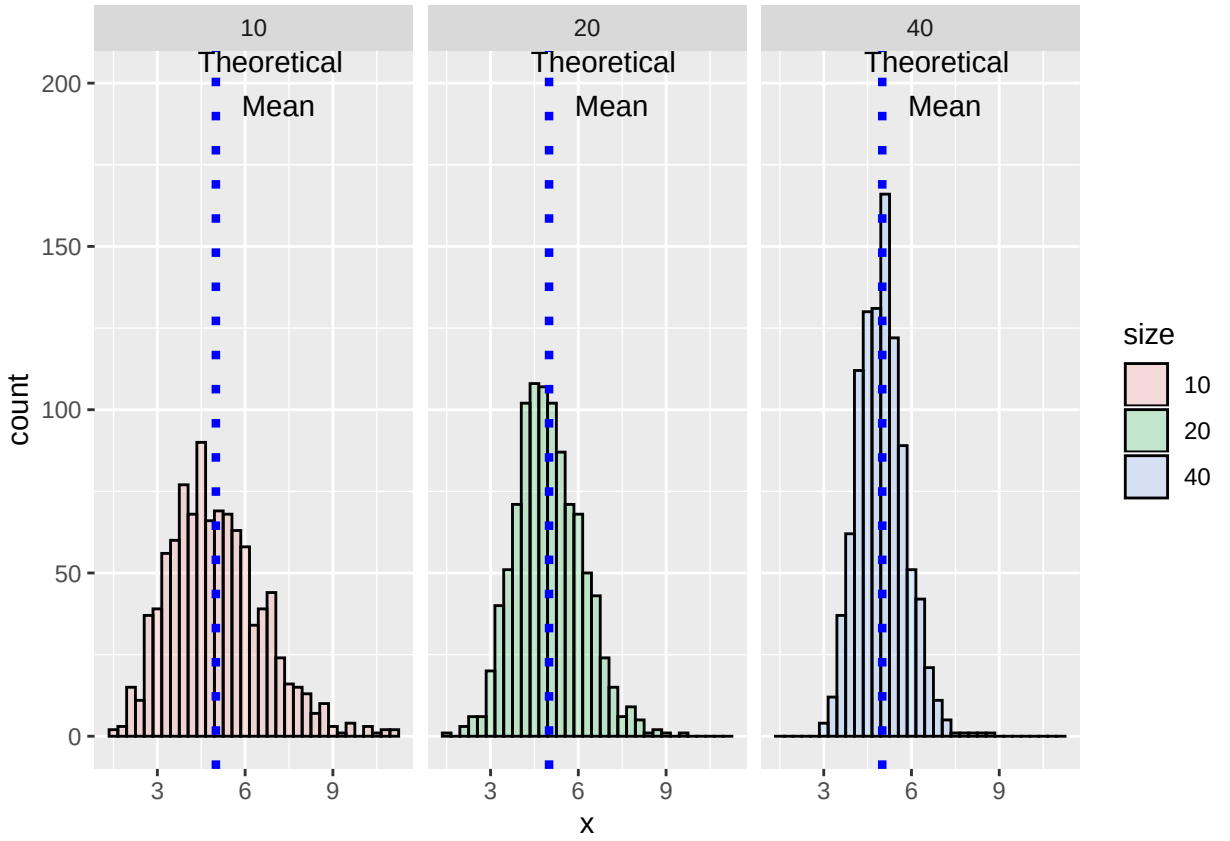


2.2 Distribution of means

The nest step is calculate the mean of samples for every simulation and the result follow below. It´s worthy to note the distribution approximation of the theoretical mean. As far as n is increased the distribution get narrowed indicating less variance as expect.

```
medias <- data.frame(
  x = c(apply(matrix(data1, nosim), 1, mean),
        apply(matrix(data2, nosim), 1, mean),
        apply(matrix(data3, nosim), 1, mean)
  ),
  size = factor(rep(c(n1, n2, n3), rep(nosim, 3))))

g <- (ggplot(medias, aes(x = x, fill = size))
      + geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..count..))
      + geom_vline(xintercept = 1/lambda, linetype = "dotted", size = 1.5, color = "blue")
      + annotate("text", label = "Theoretical \n Mean", size = 4, x = 1.4*(1/lambda), y = 200)
      )
g + facet_grid(. ~ size)
```



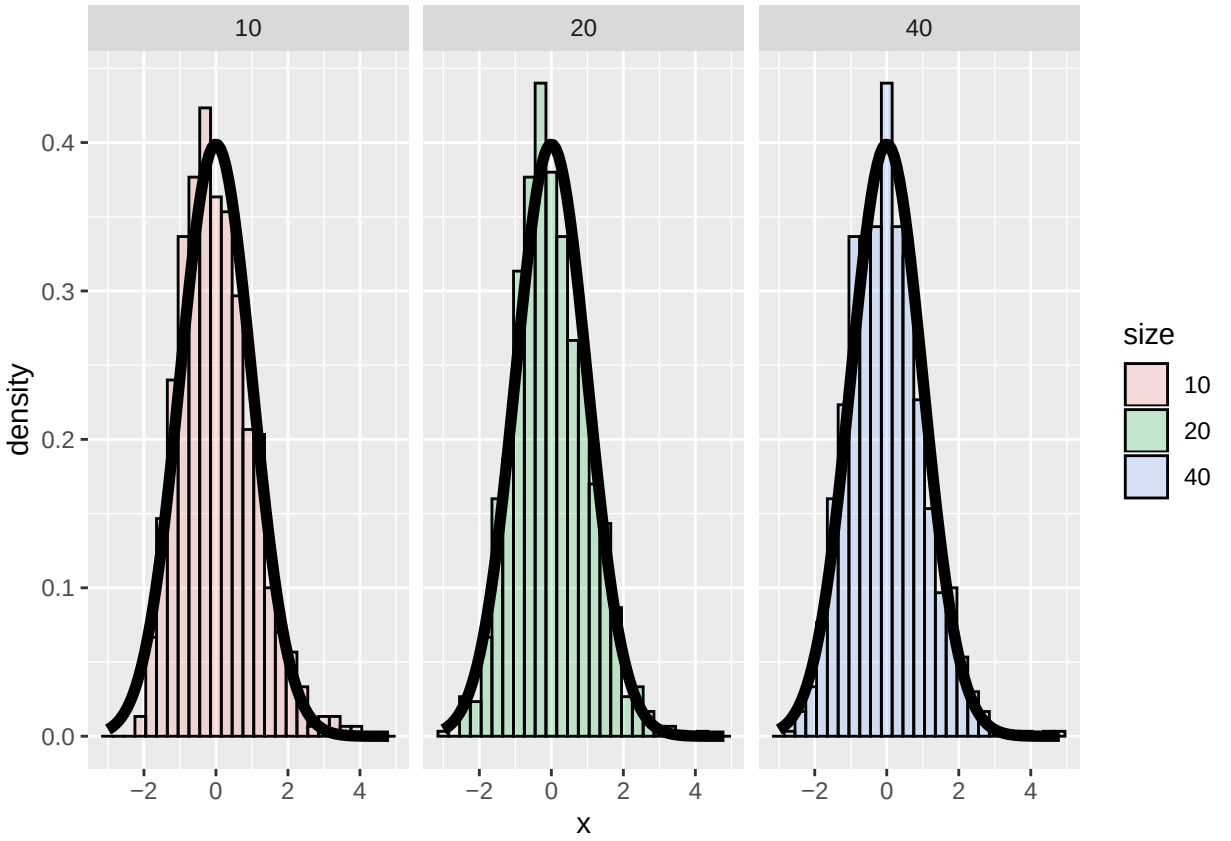
2.3 Normalization of means

The final graphic compare the distribution of means with the normal distribution using $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

```
cfunc <- function(x, n) sqrt(n) * (mean(x) - 1/lambda) / (1/lambda)

dat <- data.frame(
  x = c(apply(matrix(data1, nosim), 1, cfunc, n1),
        apply(matrix(data2, nosim), 1, cfunc, n2),
        apply(matrix(data3, nosim), 1, cfunc, n3)
  ),
  size = factor(rep(c(n1, n2, n3), rep(nosim, 3))))


g <- ggplot(dat, aes(x = x, fill = size)) + geom_histogram(alpha = .20, binwidth=.3, colour = "black", a
g <- g + stat_function(fun = dnorm, size = 2)
g + facet_grid(. ~ size)
```



2.4 Theoretical Sample variance vs. Experimental sample variance

As previous observed at graphics of means, the mean converges to its theoretical value and the variance reduce with increasing of samples.

The interval for means at 95% of confidence is:

Size

Min. value

Max. value

10

4.912307

5.109486

20

4.920897

5.061432

40

4.920897

5.061432

The table below show the theoretical variance of samples in comparison to measured value.

Size

Theor. Variance

Sample Variance

10

2.500

2.5241199

20

1.250

1.2822132

40

0.625

0.6602726

3. Conclusion

Some simulation where provided to shown this very useful characteristic of mean of distribution that tends to be a normal distribution according to the CLT. It´s also viewed variance reduction as the sample increases. That said, even with no idea of the distribution itself, take the mean of samples lead to *good estimatives of mean and variance of population.*

References

Source used in this lecture can be found at:

[1] https://github.com/bcaffo/courses
[2] https://www.calvin.edu/~rpruim/courses/s341/S17/from-class/MathinRmd.html