

Statistical Inference Project 1

Eric VACHON

May 2015

Overview

In this paper we explain and compare the distribution of 1000 simulations of the mean of 40 exponential distributions and the Central Limit Theorem (CLT). First we compare Sample Mean Versus Theoretical Mean, then the Sample Variance Versus Theoretical Variance and at the end the difference between our simulation and the CLT.

Requirement and reproducibility

Library : ggplot2 (If need install it : `install.packages("ggplot2")`) and for reproducibility, we fix the seed :

```
library(ggplot2)
set.seed(19)
```

Constants of the simulation

```
lambda <- 0.2
n <- 40
nbSimul <- 1000
```

Results : - $\lambda = 0.2$ - $n = 40$ - $\text{nbSimul} = 1000$ ## Theoretical values Mean of exponential distribution = $\frac{1}{\lambda}$
Standard deviation of the mean of n exponential distribution = $\frac{1}{\lambda} * \frac{1}{\sqrt{n}}$
The variance of the mean of n exponential distribution = $(\frac{1}{\lambda} * \frac{1}{\sqrt{n}})^2$

```
theoreticalMean <- (1 / lambda)
theoreticalStandardDeviation <- (1 / lambda)*(1/sqrt(n))
theoreticalVariance <- ((1 / lambda)*(1/sqrt(n)))^2
```

Results :

- theoreticalMean = 5
- theoreticalStandardDeviation = 0.7905694
- theoreticalVariance = 0.625

The data frame of the simulation

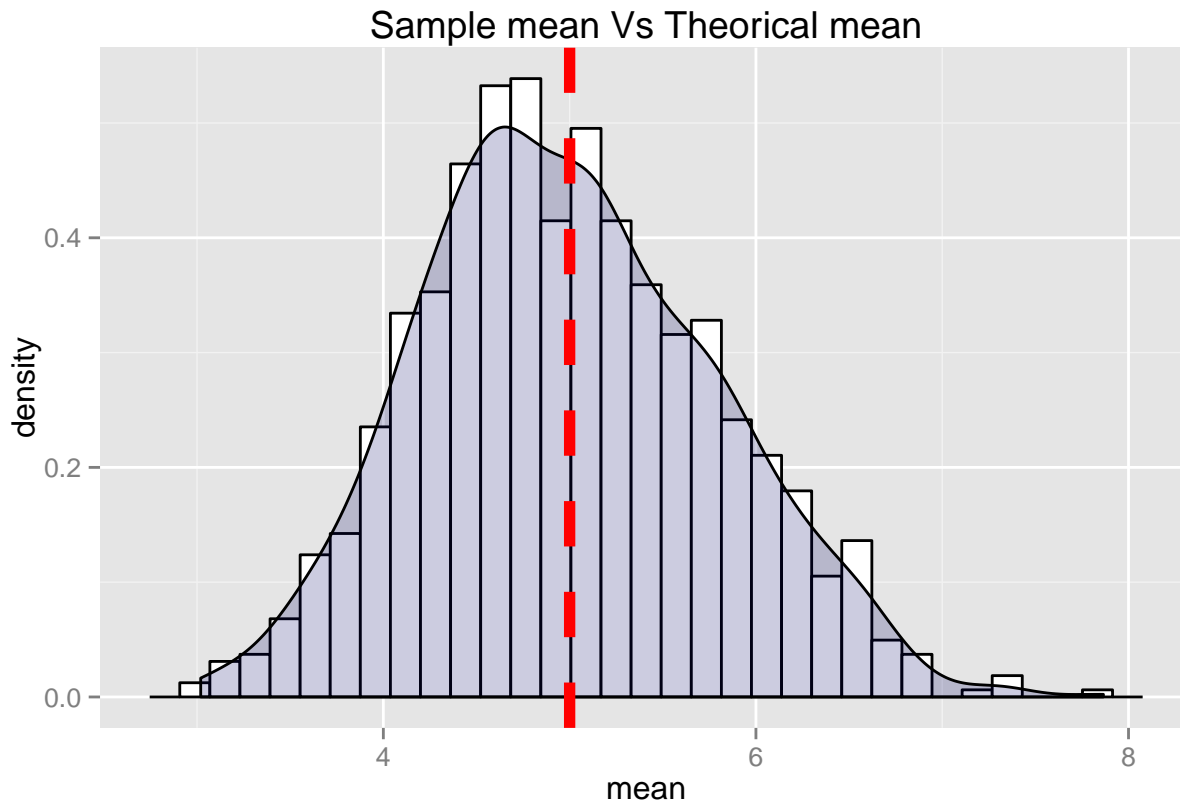
Here we are going to make a data frames with the result of $\text{nbSimul} = 1000$ means of $n = 40$ exponential distribution

```
mySimul <- data.frame()
for (i in 1 : nbSimul)
{ mySimul <- rbind(mySimul, c(i, mean(rexp(n, lambda))))}
names(mySimul) <- c('numSimul', 'valueSimul')
```

1. Sample Mean Versus Theoretical Mean

Now we make the histogram of this 1000 simulations and put a line of the theoretical mean

```
ggplot(mySimul, aes(x=valueSimul))+
  geom_histogram(aes(y=..density..), colour="black", fill="white")+
  geom_density(alpha=.2, fill="#000066")+
  geom_vline(xintercept=theoreticalMean,color="red",linetype="dashed",size=2)+
  ggtitle("Sample mean Vs Theoretical mean")+labs(x = "mean")
```



And we can compare sample mean and theoretical mean :

```
sampleMean <- mean(mySimul$valueSimul)
```

Results :

- sampleMean = 4.9913111

- theoreticalMean = 5

⇒ 0.17% of difference, so it is a good estimator.

2. Sample Variance Versus Theoretical Variance

Here we must compare the theoretical variance and the variance of our simulation :

```
simulationVariance <- var(mySimul$valueSimul)
```

Results :

- simulationVariance = 0.6172176

- theoreticalVariance = 0.625

⇒ 1.25% of difference, so it is a good estimator.

3. Comparaison with CLT

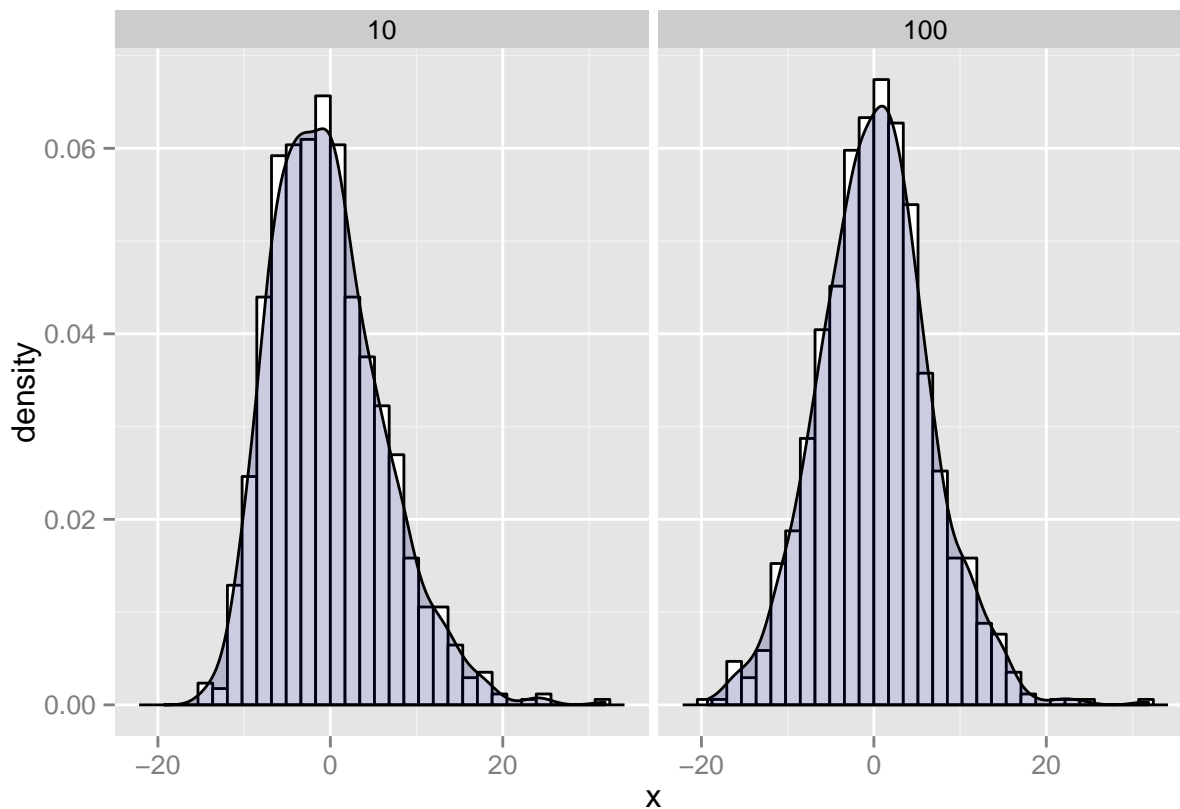
Now let's use the formula of the CLT : $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ and use n=10 and n=100 :

```

set.seed(17)
cfunc <- function(n)
{
  mySimul <- data.frame()
  for (i in 1 : nbSimul)
  {
    theMean <- mean(rexp(n, lambda))
    mySimul <- rbind(mySimul, c(sqrt(n)*(theMean - theoreticalMean)/ theoreticalStandardDeviation, n))
  }
  names(mySimul) <- c('x', 'size')
  return(mySimul)
}
dat <- data.frame( rbind(cfunc(10), cfunc(100)))

ggplot(dat, aes(x=x))+
  geom_histogram(aes(y=..density..), colour="black", fill="white")+
  geom_density(alpha=.2, fill="#000066")+facet_grid(. ~ size)

```



We can see a nice curve center on 0 and become more normal with n greater

Conclusion :

1. The theoretical mean and variance are quite near the simulation mean and variance
2. With the formula of the CLT we can see a curve

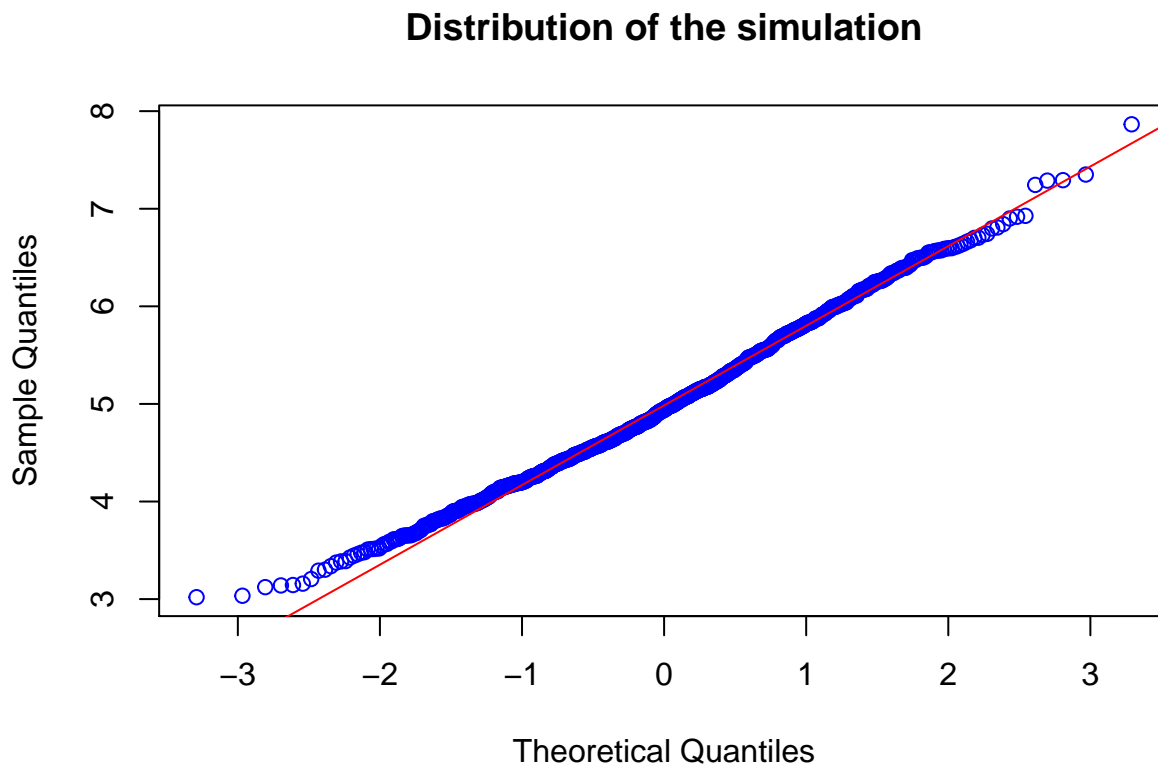
⇒ So this distribution is approximately normal.

Appendix

Another ways to know if the distribution is approximately normal we can also draw a normal Quantile-Quantile plot or use the wilcoxon test :

Quantile-Quantile plot

```
qqnorm(mySimul$valueSim, col = 'blue', main = 'Distribution of the simulation')
qqline(mySimul$valueSim, col = 'red')
```



With the normal Quantile-Quantile plot the blue circles follow the red line more or less at the extremis but follow it on the middle \Rightarrow So this distribution is approximately normal.

wilcoxon test

```
wilcox.test(mySimul$valueSim)

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  mySimul$valueSim
## V = 500500, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 0
```

the p-value of the wilcoxon test is less than 5% \Rightarrow So this distribution is approximately normal.

End of the document