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 Theorm (CLT). First we compare Sample Mean Versus Theorical Mean, then the Sample Variance Versus Theorical Variance and at end the difference between our simulation and the CLT.

Requirement and reproductibility

Library : ggplot2 (If need install it : install.packages("ggplot2")) and for reproducability, we fixe the seed :

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

Constants of the simulation

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

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

```
theoricalMean <- (1 / lambda)
theoricalStandardDeviation <- (1 / lambda)*(1/sqrt(n))
theoricalVariance <- ((1 / lambda)*(1/sqrt(n)))^2
```

Results :
- theoricalMean = 5
- theoricalStandardDeviation = 0.7905694
- theoricalVariance = 0.625

The data frame of the simulation

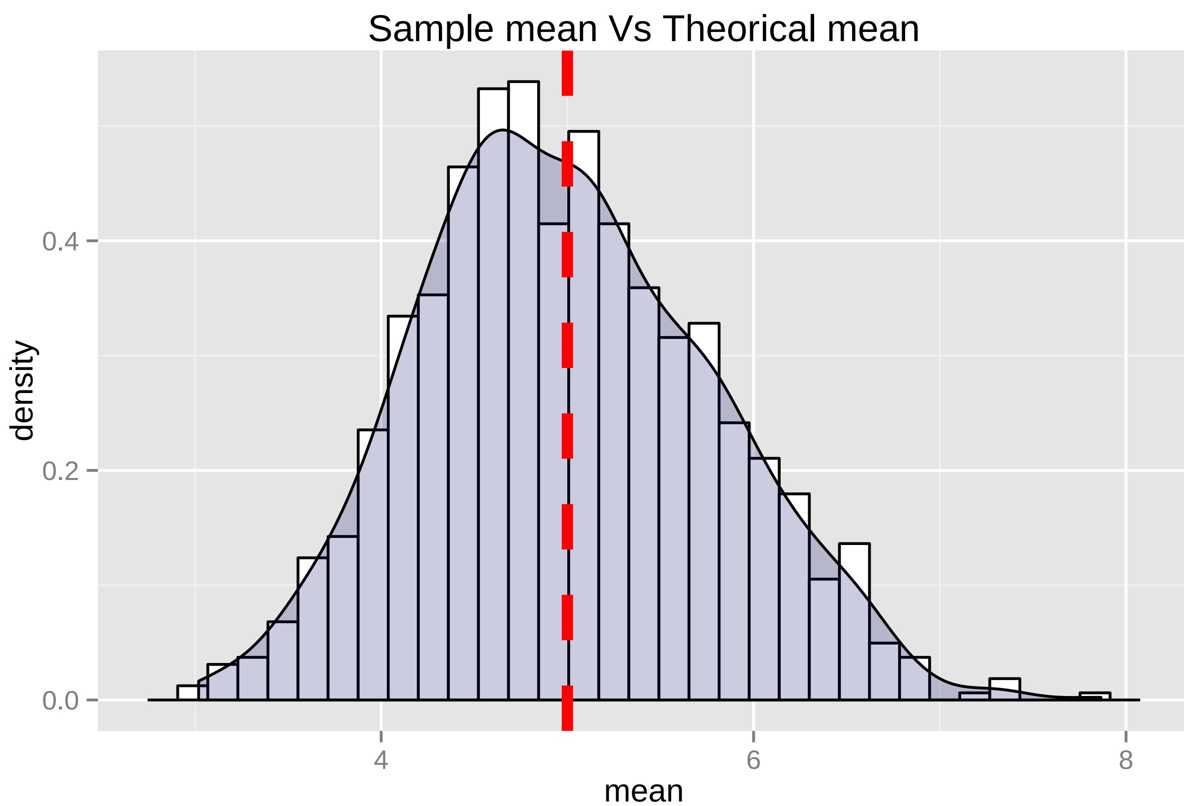
Here we are going to make a data frames with the result of nbSimul=1000 means of n=40 exponencial distribution

```
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 Theorical Mean

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

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



And we can compare sample mean and theorical mean :

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

Results :
- sampleMean = 4.9913111
- theoricalMean = 5

$\Rightarrow$ 0.17% of difference, so it is a good estimator.

2. Sample Variance Versus Theorical Variance

Here we must compare the therical variuance and the variance of our simulation :

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

Results :
- simulationVariance = 0.6172176
- theoricalVariance = 0.625

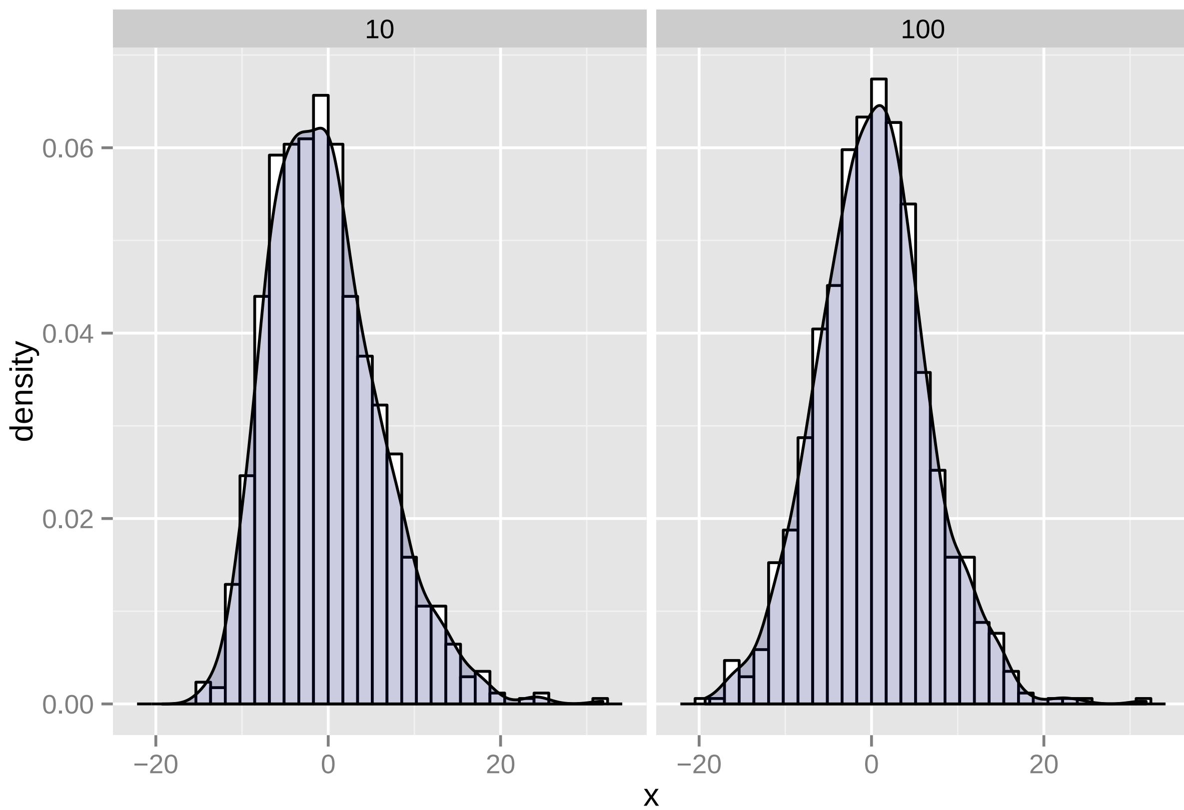
$\Rightarrow$ 1.25% of difference, so it is a good estimator.

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 :

```r
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 - theoricalMean)/ theoricalStandardDeviation,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 theorical mean and variancee are quite near the simulmation mean and variance
2. With the formula of the CLT we can see a cuvre

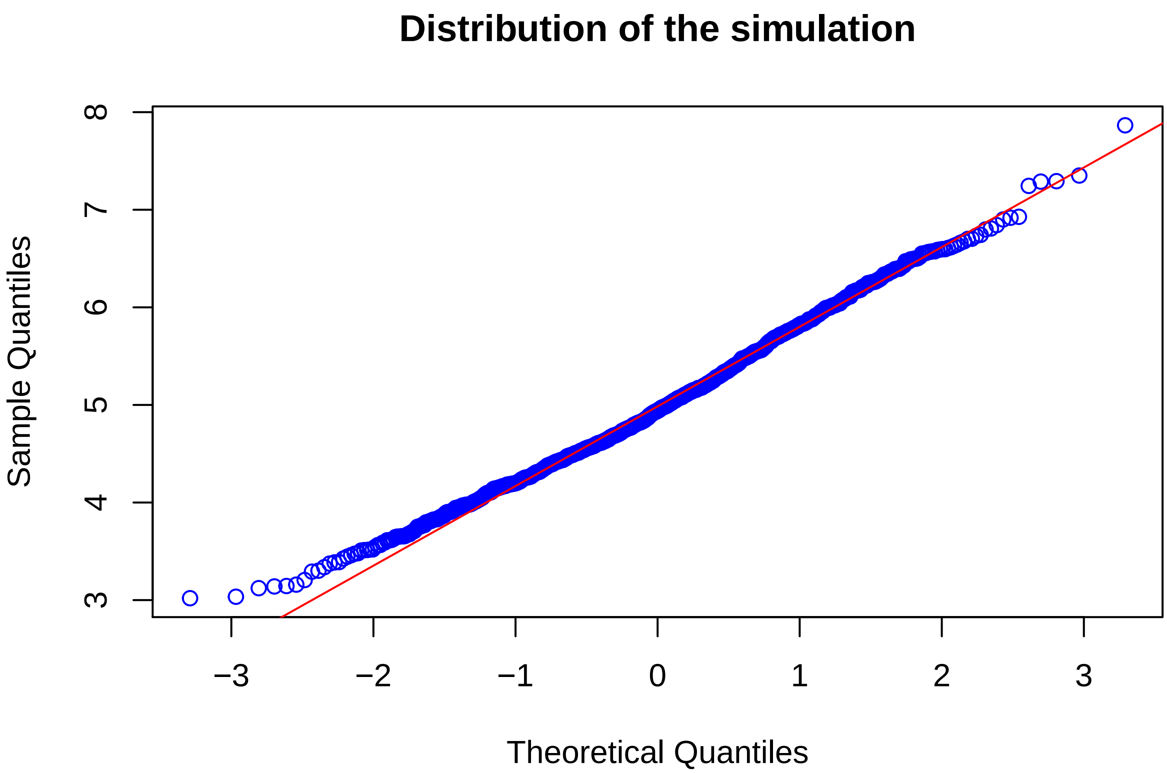
⇒ So this distribution is approximately normal.

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 extrems but follow it on the middle ⇒ 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 wilconox test is less than 5% ⇒ So this distribution is approximately normal.

End of the document