Course Project - Statistical Inference

Jean-Luc BELLIER

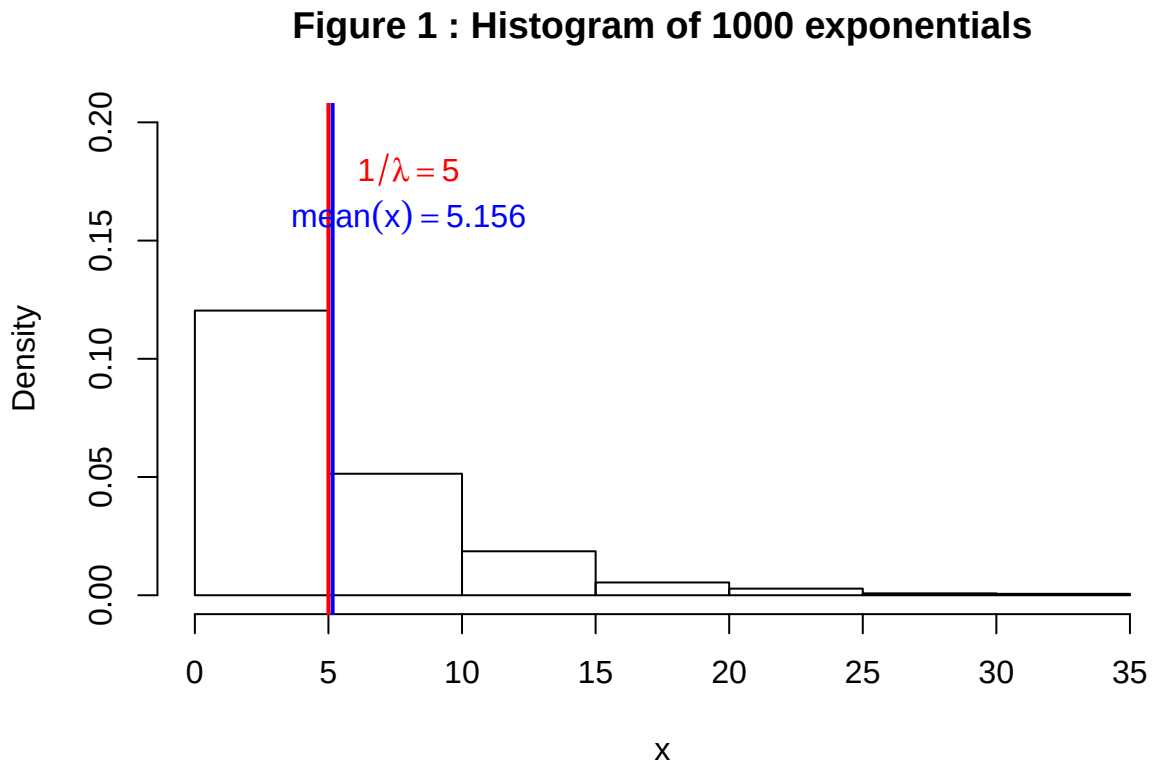
13 novembre 2016

Part 1: Simulation Exercise Instructions

In the first part of the project, we will investigate the behaviour of means for exponential distribution in R and compare it with the expected distribution obtained by Central Limit Theorem. In this part, we will set the rate λ to 0.2. First we draw the histogram of 1000 random exponential values. Let us call this distribution X. For the mean distribution, we will consider a mean of n=40 exponential random values. Let us call Y this distribution of means.

Sample distribution of exponential values

```
set.seed(1)
lambda = 0.2
x <-rexp(1000, rate =lambda)
hist(x,main="Figure 1 : Histogram of 1000 exponentials",xlim=c(0,35),ylim=c(0,0.2),freq=FALSE)
abline(v=mean(x),col="blue", lwd=2)
abline(v=1/lambda,col="red", lwd=2)
text(8,0.18,expression(1/lambda== 5), col="red")
text(8,0.16,expression(mean(x) ==5.156), col="blue")
```



```r
print(mean(x))
```

```
## [1] 5.156513
```

```r
print(sd(x))
```

```
## [1] 4.946295
```

```r
print(var(x))
```

```
## [1] 24.46583
```

Let us compare the theoritical values and the real values obtained for this sample :

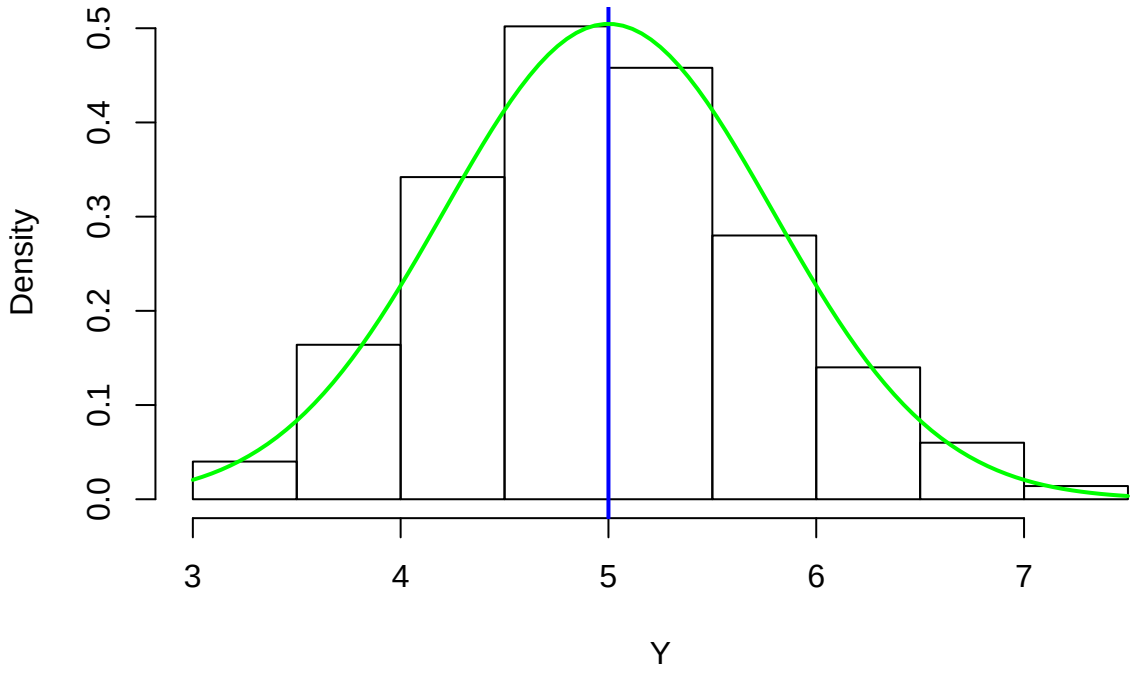
- Mean : theoritical = $1/\lambda = 5$, real value = mean(x) = 5.156513, disparity = 3.13%
- Standard deviation : theoritical = $1/\lambda = 5$, real value = sd(x) = 4.94629, disparity = -1.07%
- Variance : theoritical = $1/\lambda^2 = 25$, real value = var(x) = 24.464583, disparity = -2.14%

Sample distribution of mean of 40 exponential values

Below we draw a distibution of 1000 means of random exponential values, each mean built on 40 random exponential values. The blue thick line represents the mean of the exponential distribution. The green curve is the corresponding theoritical limit, which is a normal distribution.

```r
rate = 0.2
mns = NULL
size=40
for (i in 1 : 1000) mns = c(mns, mean(rexp(size,rate=lambda)))
hist(mns, main ="Figure 2 : Histogram of 1000 means of 40 exponentials", xlab="Y",freq=FALSE) # histogr
abline(v=5,col="blue", lwd=2)
curve(dnorm(x, mean=1/lambda, sd=1/lambda/sqrt(size)), col="green", lwd=2, add=TRUE, yaxt="n")
```

Figure 2 : Histogram of 1000 means of 40 exponentials




```
moy <- mean(mns)
sd_mns <- sd(mns)
print(moy)
```

```
## [1] 4.988882
```

```
print(sd_mns)
```

```
## [1] 0.7822446
```

Let us have a look on the results. We give first the means of the exponential distribution. We know that, for a random variable X following the exponential distribution of rate λ :

$$\bar{X} = 1/\lambda$$

and

$$Var(X) = \sigma^2 = 1/\lambda^2$$

Let us focus now on the distribution of means. We know, by the Central Limit Theorem, that, the distribution of means follows in theory a normal distribution :

$$\bar{X} \sim \mathcal{N}(1/\lambda, 1/(\lambda^2 * n)$$

- theoritical value of mean : $1/\lambda = 5$, real value $= 4.9889$

- theoritical value of standard deviation : $\sigma/\sqrt{n} = 0.79057$, real value = 0.7901.
- The % of values that contain the mean are the following :
  - $P(-\sigma \leqq X \leqq \sigma) = 68.27\%$
  - $P(-2\sigma \leqq X \leqq 2\sigma) = 95.45\%$
  - $P(-3\sigma \leqq X \leqq 3\sigma) = 99.75\%$

For the real value, we get the same set of percentages. This can be obtained by computing the quantiles using mean(x) and sd(x).

Discussion

When we look at the two figures (Figure 1 and Figure 2), we can see that :

- Theoritical Mean vs Sample Mean : Although the two figures are really different, we can notice that both means are very similar (around 5) : the first figure gives a histogram of frequencies, and the second a histogram of densities. Furthermore, in Figure 2, the repartition of values is not the same on both sides of the mean, but the mean is very similar.
- Theoritical variance vs Sample Variance : We can notice that the larger n is, the more the variance of the mean distribution is low, that means that the sample is more concentrated arouns the mean. This can be easily deducted from figures 1 and 2. In figure 1, we have $n = 1$, and the dispersion of the values is very large, whereas in figure 2 ($n = 40$), the dispersion is much lower.
- Standard exponential distribution vs mean distribution : we can also notice, from figures 1 and 2, that the distributions are very different : the density of exponential is a decreasing function, that has 0 as a limit as x goes to infinity, whereas the density of mean is a bell curve, which is dense around the mean. This confirms our expectations on the fact that the larger the sample is (the value of n), the more the behaviour is converging to a normal distribution, and that the distribution of means converges to a normal distribution.