Coursera Statistical Inference Project: Part I

A large sampling simulation of the exponential distribution

Giuseppe Di Bernardo

February 29, 2016

This is the Assignment Project part I of the Coursera Statistical Inference course. We will explore the properties of the **exponential distribution** in `R`, and compare this one with the results of the **Central Limit Theorem**.

From the theory, we know that the **PDF** (probability density finction) corresponding to the exponential distribution is $f_X(x) = \lambda e^{-\lambda x}, x > 0$. In the statistical software package `R` we can get it by invoking the command `rexp(sample_size, rate = lambda)`, where `sample_size` is the size of our sample, and `lambda` the rate parameter. The goal is to investigate the distribution of the averages (sample means) of $N = 40$ exponential distributions, by running a large number of simulations (e.g 10^3).

Before showing the analysis performed - and the results obtained as well - we remind to the reader that the exponential distribution with parameter λ has the mean value $\mu_X = 1/\lambda$, and the variance σ_X^2 given by $1/\lambda^2$. Through all our simulations, we set $\lambda = 0.2$. Therefore, theoretically we expect a mean value of 5 and a standard deviation sd $= \sqrt{\sigma_X^2}$ of 5 as well.

1. Simulations

Below, it is the chunck code used to perform a thousand of simulations of exponential distributions.

```
set.seed(12121)
lambda <- 0.2
sample_size <- 40
nsim <- 1.0e3
exp_sample <- rexp(sample_size, rate = lambda)
```

Here, `exp_sample` is the vector corresponding to one sample realization of the exponential distribution, of size 40 outcomes. Here is a brief insights into the data:

```
head(exp_sample)
```

```
## [1] 5.425193 7.367347 5.222159 6.458976 2.041147 3.917055
```

```
tail(exp_sample)
```

```
## [1] 1.26051362 7.06139768 0.06522054 0.66544598 4.89078803 3.10026186
```

```
summary(exp_sample)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.06522  1.28000  4.58000  4.91500  6.04800 27.77000
```

```r
sd(exp_sample)
```

```
## [1] 5.128182
```

The funtion `replicate()` allows us to repeat many times $(1,000)$ the indipendent identically distributed (**i.i.d**) random variables $X_1, ..., X_{1e3}$, to get an idea of the distribution of the averages quantities.

```r
repeated_exp <- replicate(nsim, mean(rexp(n = sample_size, rate = lambda)))
```

Here, `repeated_exp` is the vector containing the means of a thousand samples sized 40, drawn without replacement from an exponential distribution.

```r
summary(repeated_exp)
```

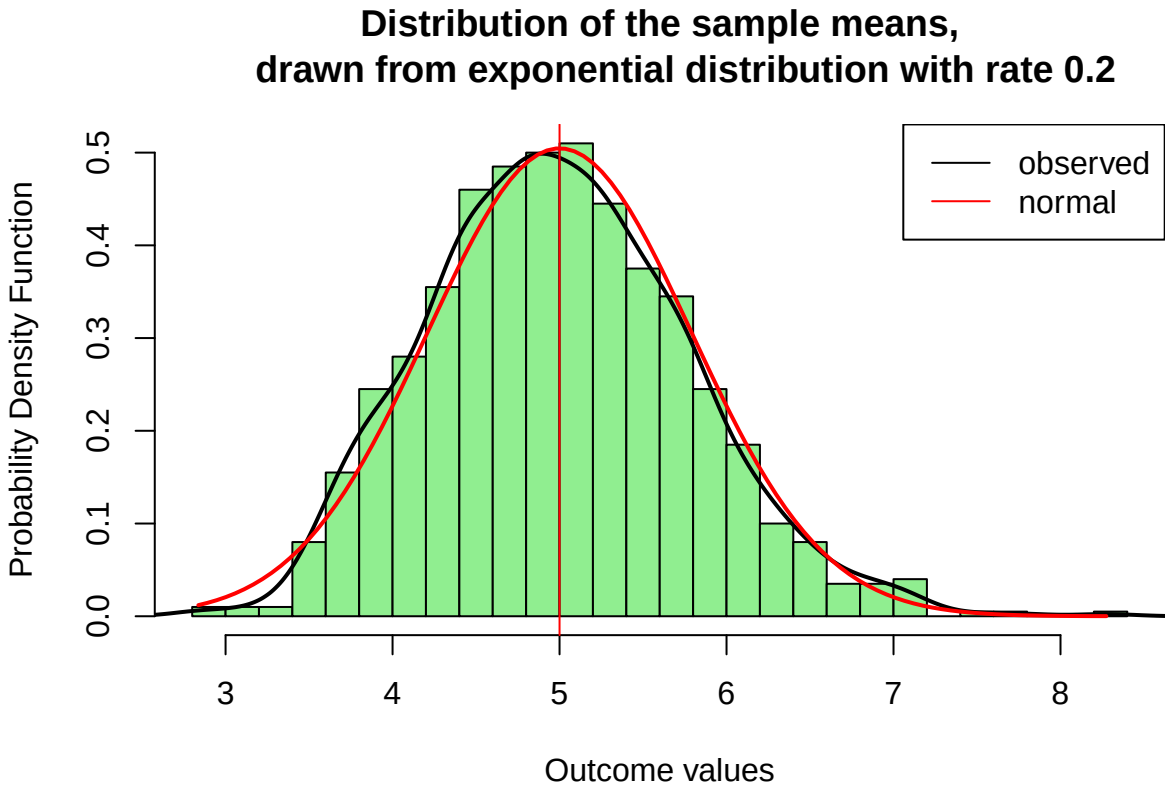
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.837   4.431   4.954   4.994   5.504   8.275
```

```r
sdeviation <- sd(repeated_exp, na.rm = FALSE)
sdeviation
```

```
## [1] 0.7768964
```

2. Results and Plots

The Central Limit Theorem (**CLT**) assures us that the so-called **sample mean**, $\bar{X}_{1e3} = \frac{1}{N}\sum_{i=1}^{10^3} X_i$, has a distribution which is approximately Normal $N(\mu, \sigma^2/n)$, with mean μ_X and variance σ_X^2/N. The observed mean of the large sampling distribution is 4.994, perfectly in agreement with the theoretical expectation value of 5, as previously stated. And the standard deviation value got from the simulations is 0.777, which agrees with what is predicted from the **CLT**, i.e. $(1/\lambda)/\sqrt{N} = 0.791$. The two numbers would be even closer with larger samples. We show in the following plot what we have argued so far in our analysis:



Clearly we observe that the averages of samples (black line) follow closely a theoretical normal distribution (red line). The normal behaviour of our sampling distribution is evident also by looking at the quantile - quantile plot: