Part I: Simulation Exercise

Maurício Collaça

October 24 2017

Introduction

This report investigates an simulation experiment of the exponential distribution and compare it with the Central Limit Theorem (CLT).

Wikipedia (2017): The exponential distribution is the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate.

Its probability density function (PDF) is $\lambda e^{-\lambda x}$ where $\lambda > 0$ and $x \epsilon [0, \infty)$.

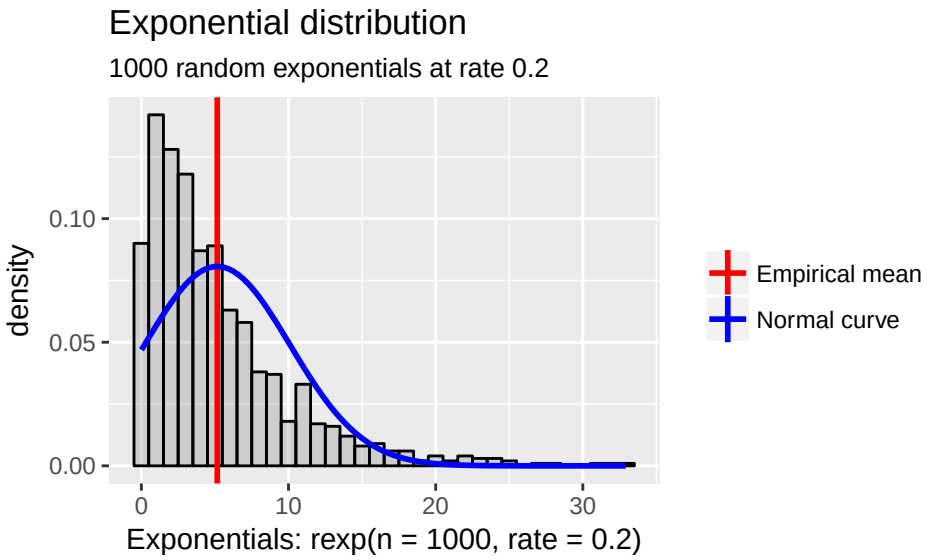
The exponential distribution can be simulated in R with `rexp(n, lambda)` where `lambda` is the rate parameter. The mean of the exponential distribution is 1/lambda and the standard deviation is also 1/lambda.

Simulations

The exponential distribution

The exponential distribution can be visualized by plotting a histogram of relatively large, e.g. 1000, random exponentials with a rate parameter λ (lambda) equals to, for instance, 0.2. A random number generation seed is specified in order to ensure the reproducibility of the code results. Overlaying the histogram and the normal curve of the data demonstrates the exponential distribution is not normally distributed, i.e., it's not close to a symmetric bell-shaped curve.

```
library(ggplot2)
set.seed(1)
nosim <- 1000; lambda <- 0.2
exponentials <- data.frame(x = rexp(nosim, lambda))
ggplot(exponentials, aes(x = x)) + xlab("Exponentials: rexp(n = 1000, rate = 0.2)") +
    geom_histogram(aes(y = ..density..), alpha = .2, binwidth = 1, color = "black") +
    geom_vline(aes(xintercept = mean(exponentials$x), color = "Empirical mean"), size=1) +
    stat_function(fun=dnorm, args=list(mean=mean(exponentials$x),
                                       sd=sd(exponentials$x)),size=1,aes(color="Normal curve")) +
    scale_color_manual(NULL, values = c("Empirical mean"="red", "Normal curve"="blue")) +
    ggtitle("Exponential distribution", paste(nosim, "random exponentials at rate", lambda))
```



The empirical or sample mean of a relatively large sample approximately estimates the theoretical or population mean `1/lambda` which is 5.

```
mean(exponentials$x)
```

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

The empirical or sample variance of a relatively large sample approximately estimates the theoretical or population variance `(1/lambda)^2` which is 25..

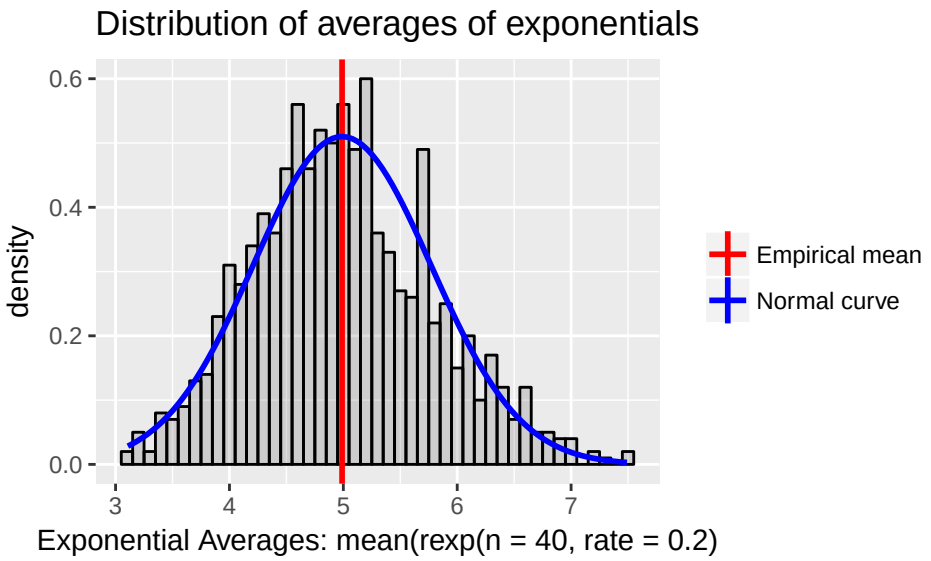
```
var(exponentials$x)
```

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

The distribution of averages of random exponentials

Random variables are said to be independent and identically distributed (IID) if they are independent and all are drawn from the same population, therefore, random exponentials are a case of IID variables. The Central Limit Theorem (CLT) states that the distribution of averages of IID variables becomes that of a standard normal as the sample size increases. This theorem can be visually demonstrated by plotting a histogram of a reasonably large (e.g. 1000) number of averages of a reasonably large (e.g. 40) sample size of random exponentials with a rate parameter λ (lambda), for instance, equals to 0.2. Overlaying the histogram and the normal curve of the data demonstrates the CLT: the distribution of averages of exponentials is normally distributed even when the population of exponentials is not normally distributed, as shown in the previous plot.

```
n <- 40
avgs <- data.frame(x = sapply(1:nosim, function(x) mean(rexp(n,lambda))))
library(ggplot2)
g2 <- ggplot(avgs, aes(x = x)) +
    geom_histogram(aes(y = ..density..), alpha = .20, binwidth = .1, color = "black") +
    geom_vline(aes(xintercept = mean(avgs$x), color = "Empirical mean"), size=1) +
    stat_function(fun = dnorm,
                  args = list(mean = mean(avgs$x), sd = sd(avgs$x)),
                  aes(color = "Normal curve"), size = 1) +
    scale_color_manual(NULL, values = c("Empirical mean"="red", "Normal curve"="blue")) +
    xlab("Exponential Averages: mean(rexp(n = 40, rate = 0.2)") +
    ggtitle("Distribution of averages of exponentials")
g2
```



The sampling distribution of the mean is centered around the population mean `1/lambda` which is 5.

```
mean(avgs$x)
```

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

The variance of the sampling distribution of the mean is σ^2/n, where σ^2 is the variance of the population of being sampled from. In the case of the exponential distribution population, $\sigma = (1/\lambda)$.

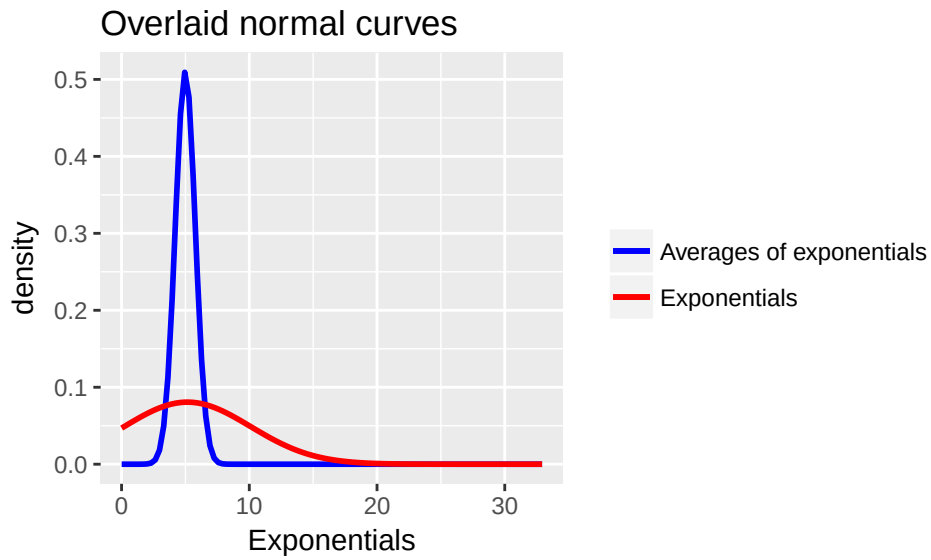
Therefore, the variance of the sampling distribution of the mean is much lower than the theoretical variance of the population `1/lambda` which is 5.

```
((1/lambda)^2)/n
```

```
## [1] 0.625
```

The following plot shows overlaid normal curves of 1000 exponentials and 1000 averages of 40 exponentials. One can see how less spread is the distribution of averages and how both curves are centered at 1/lambda. It also demonstrates the CLT: how the sampling distribution of the mean of IID exponential variables approximates the normal distribution.

```
ggplot(data.frame(x=range(exponentials,avgs)), aes(x = x)) +
    stat_function(fun=dnorm, args=list(mean=mean(avgs$x), sd=sd(avgs$x)),
                  size=1, aes(color="Averages of exponentials")) +
    stat_function(fun=dnorm, args=list(mean=mean(exponentials$x),
                                       sd = sd(exponentials$x)),
                  size=1, aes(color="Exponentials")) +
    scale_color_manual(NULL, values = c("Exponentials"="red","Averages of exponentials"="blue")) +
    ylab("density") + xlab("Exponentials") + ggtitle("Overlaid normal curves")
```



Conclusion

As demonstrated in the previous section, the conclusion is that the exponential distribution population is not normally distributed but as per stated by the Central Limit Theorem, their sampling distribution of the mean approximates a normal distribution because of the number of simulations (1000) and the random exponential variable is independent and identically distributed (IID).