Statistical Inference Part I - Exponential Distribution

Fabien Nugier

11/15/2019

This project consists of two parts, in Part I we study the properties of the exponential distribution and compare its properties with the Central Limit Theorem (CLT). In the Part II, we perform some basic inferential data analysis using the R dataset *ToothGrowth*. This is just an illustrative assignment.

This document contains part I.

Introduction

As explained on Wikipedia-Exponential_Disribution, the exponential distribution is the distribution of time intervals between two events in a Poisson point process, for which events occur continuously, independently from each other, and at a constant average rate. The probability density function of the exponential distribution is given by:

$$f(x, \lambda) = \begin{cases} \lambda \, e^{\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

with λ the "rate parameter" of the distribution. Its mean is given by $1/\lambda$ and variance is $1/\lambda^2$.

Simulations of the Exponential Distribution

Let us first illustrate the exponential distribution by one single sample taken out from this distribution. We define the parameters of the problem:

```
lambda = 0.2    # rate parameter
n = 40          # population within one sample
Nsim = 1000     # number of samples (simulations)
```

and draw one sample out of this distribution:

```
set.seed(12345)
hist(rexp(n,lambda), xlab="x", main="Histogram of x ~ Exp(0.2)", col="light blue")
```


Histogram of x ~ Exp(0.2)
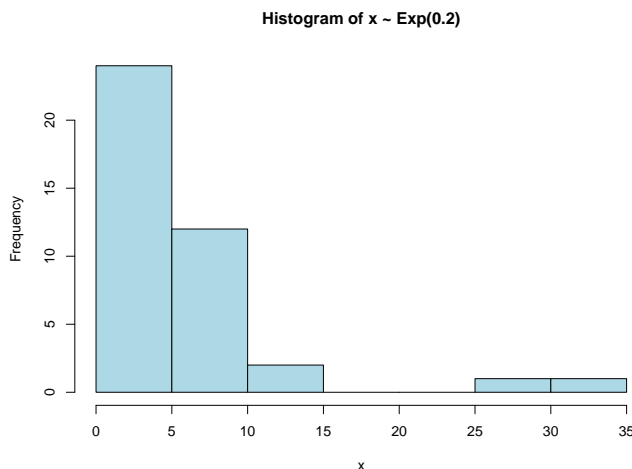
We can see that the distribution looks like a decreasing exponential with a rate of $1/0.2 \sim 5$. This value of 5 is the value we would rouhgly obtain by prolongation of the tangent at x=0.

Now sampling a thousand times, we can get the distribution of the sample mean. The sampling is done as follows:

```r
X = NULL
set.seed(123)
for (i in 1:Nsim) X = c(X, mean(rexp(n,lambda)))
m <- mean(X) # We take the average of the sample means
s <- sd(X)   # We take the standard deviation of the sample means
print(paste("mean:",round(m,3),", standard error:",round(s,3)))
```

```
## [1] "mean: 5.012 , standard error: 0.775"
```

Since we know the theoretical mean to be $1/\lambda$ and the variance to be $(1/\lambda)^2/\sqrt{(n)}$, we can plot on top of the histogram the theoretical distribution $N(x; \mu = 1/\lambda, \sigma = 1/(\lambda\sqrt{n}))$. Note that since the histogram counts the number of occurences of values within bins, we need to scale the theoretical distribution with it. We use for that a method presented on *stack overflow*.

```r
# Initializing
new.mai <- old.mai <- par("mai")
new.mai[4] <- old.mai[2]
par(mai = new.mai)
# Plot 1 preparation
h <- hist(X, breaks=50, plot=FALSE)
pos <- pretty(h$density, n = 5)
freq <- round(pos * length(X) * with(h, breaks[2] - breaks[1]))
# Plot 2 preparation
xseq <- seq(0,10,0.2)
mth <- 1/lambda # theoretical mean
sth <- (1/lambda)/sqrt(n) # theoretical standard error of the mean
# Plot 1
graphics:::plot.histogram(h, freq = FALSE, col="light blue",
                          main="Histogram of the Sample Mean",
                          xlab = "sample mean", ylab="Frequency",
                          border="black", yaxt='n')
Axis(side = 2, at = pos, labels = freq)
Axis(side = 4, at = pos, labels = pos)
mtext("Density", side = 4, line = 3)
# Plot 2
lines(xseq, dnorm(xseq,mth,sth), type='l', lwd=2)
```

Histogram of the Sample Mean
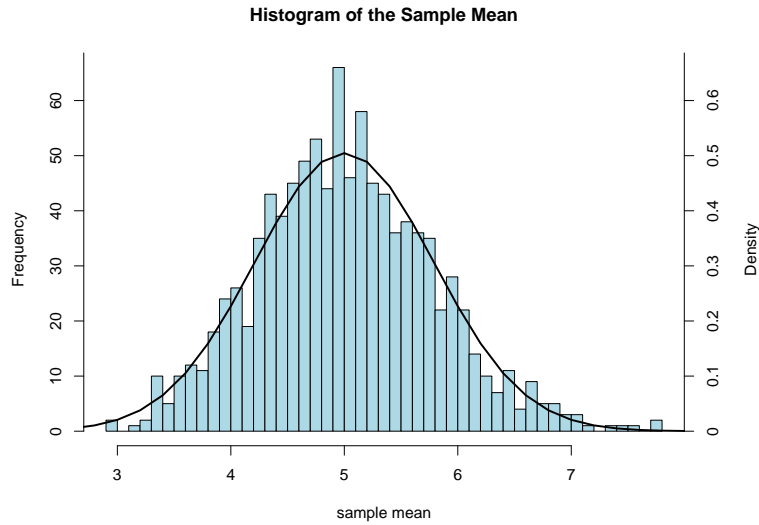
Since we are considering the sample means of idd variables here, and as attested by the plot, the sample mean distribution approaches a standard normal distribution when the size of the sample means population is large enough (here 1000). This confirms the validity of the Central Limit Theorem in the case of the exponential distribution.

Short conclusions

We have shown in this document that the sample mean of exponential distribution, like for many other distributions, converges to the normal distribution when the number of samples becomes large.