Statistical Inference Course - Study of the 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.

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 $1/\lambda^2$.

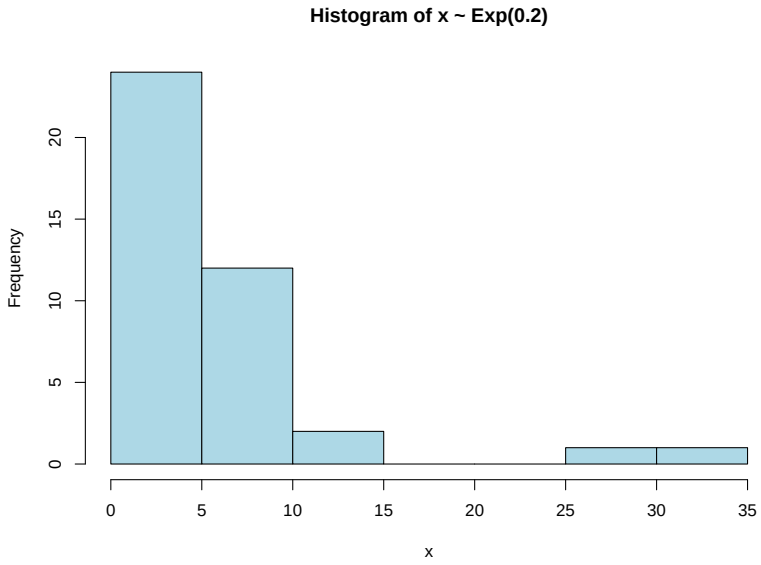
Part I - Simulations of the Exponential Distribution

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")
```



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:

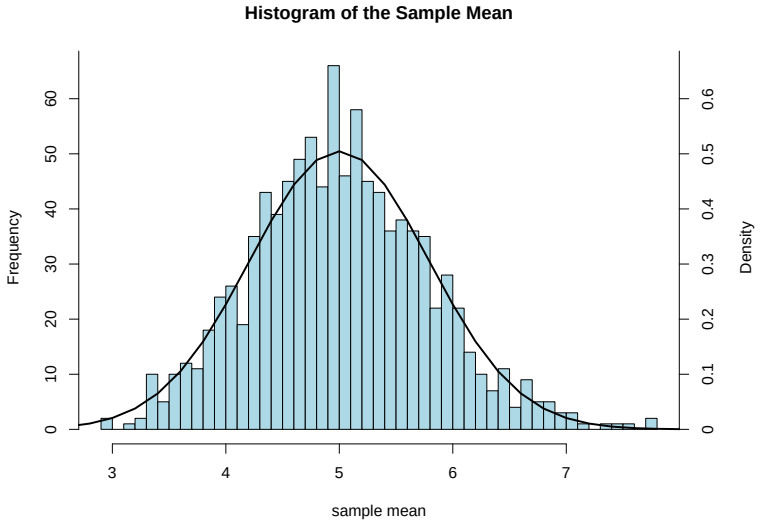
```
X = NULL
set.seed(123)
for (i in 1:Nsim) X = c(X, mean(rexp(n,lambda)))
m <- mean(X) # We take the average of the sample means
s <- sd(X)   # We take the standard deviation of the sample means
print(paste("mean:",round(m,3),", standard error:",round(s,3)))
```

```
## [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*.

```
# 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)
```



Since we are considering the sample means of idd variables here, and as attested by the plot, the sample means 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.

Part II - Inferential Data Analysis on the ToothGrowth dataset

In this part we perform a simple exploratory analysis on the R dataset `ToothGrowth`. According to the description of the dataset, the data reports the study of the effect of Vitamin C on tooth growth (the tooth being used is the incisor tooth) in guinea pigs, more precisely:

> The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC).

We load the package as:

```
data(ToothGrowth)
```

and check information about the dataset:

```
data(ToothGrowth)
head(ToothGrowth,n=3)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
```

```
dim(ToothGrowth)
```

```
## [1] 60  3
```

```
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
summary(ToothGrowth$supp)
```

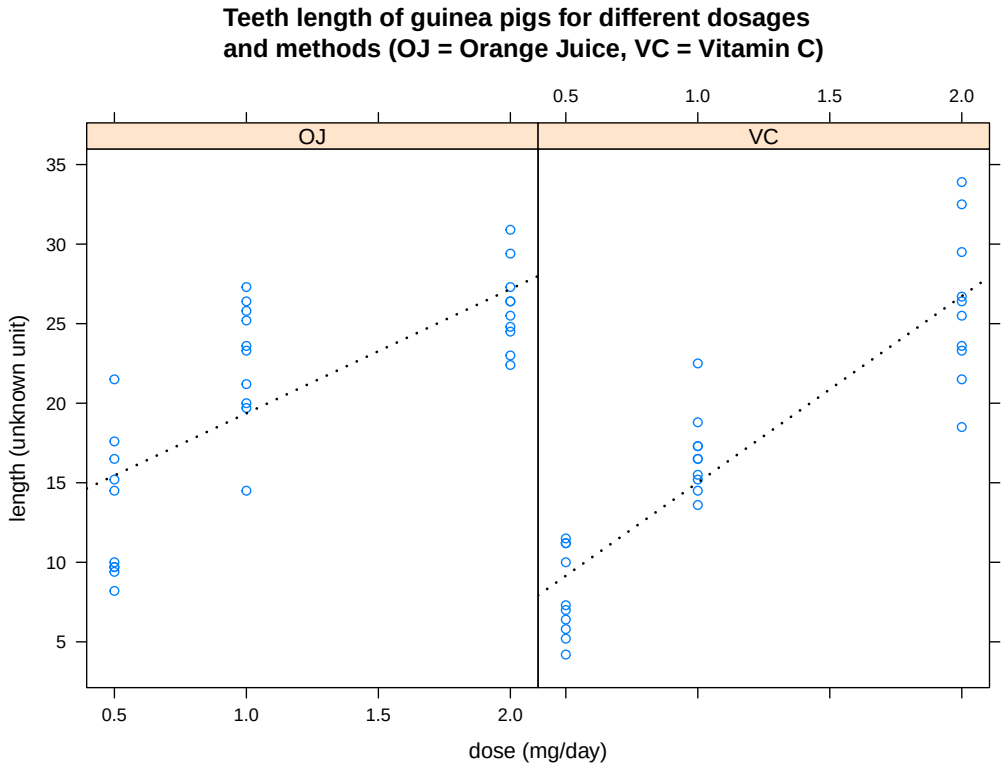
```
## OJ VC
## 30 30
```

```
unique(ToothGrowth$dose)
```

```
## [1] 0.5 1.0 2.0
```

To get a better grasp on the data, we do a scatter plot separating the 2 delivery methods (OJ for "Orange Juice" and VC for "Ascorbic acid", a form of "Vitamin C") and different dosages:

```
library(lattice)
pltt <- "Teeth length of guinea pigs for different dosages \n"
pltt <- paste(pltt,"and methods (OJ = Orange Juice, VC = Vitamin C)")
xl <- "dose (mg/day)"
yl <- "length (unknown unit)"
pan <- function(x,y,...) {
    panel.xyplot(x, y, ...)
    panel.lmline(x, y,type='l',lty=3,lwd=2, ...)
    }
xyplot(len~dose|supp,data=ToothGrowth,layout=c(2,1),xlab=xl,ylab=yl,main=pltt,panel=pan)
```



We perform a T-test on the data, assuming that the two guinea pigs test groups with OJ and VC treatments have the same underlying distribution and not pairing elements of the groups. We obtain the following results:

```r
t.test(len ~ supp, paired=FALSE, var.equal=TRUE, data=ToothGrowth)
```

```
##
##  Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 58, p-value = 0.06039
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1670064  7.5670064
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

The null hypothesis H_0 is that both test groups have the same tooth length in the OJ and VC groups. Considering the p-value which is 0.06039, we can see that a 5% p-value would not reject the null hypothesis. However, a 1% p-value would reject $H - 0$, meaning the treatment has been more efficient on the OJ group, for which the average tooth length is longer, than in the VC group.

The t-test also returns to us the 95% confidence interval of the difference between the two groups. As we can see the value 0 is not excluded, which confirms that the null hypothesis is not excluded a 5% confidence level.

Looking at the figures, we can also see that the dosage has a significant importance in the result. The more the dosage, the longer the tooth. This seems even more true for the VC group than it is for the OJ group.

Short conclusions

We have presented in this assignment two things. In the first part, we have shown that the sample mean of exponential distribution, like for many other distributions, converges to the normal distribution when the number of samples becomes large. In the second part we have studied a dataset of two different treatments and dosages applied to guinea pigs, with the associated measurement of their tooth length. It was shown that one treatment is better than the other, but a 5% confidence level is not enough to reject the null hypothesis of equal mean between the two groups. However, dosage is clearly correlated with the tooth lenth.