# Inference Statistics Course Project 3

Mathias Barat

25/09/2020

## PART 1 - SIMULATION EXERCISE

### Overview

Investigate the exponential distribution in R and compare it with the Central limit theorem

### Simulation

```
# Set-up my inputs:
lambda <- 0.2
mu <- 1/lambda
sd <- 1/lambda
n <- 40
n_simu <- 1000


# data simulation
set.seed(12345)
matrix_sample <- matrix(rexp(n*n_simu,lambda), n_simu, n)
mean_sample <- rowMeans(matrix_sample)


# summary
summary(mean_sample)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.032   4.424   4.938   4.972   5.492   8.380
```

### Sample Mean

```
mua <- mean(mean_sample)

deltamu <- mua - mu
deltamu
```

```
## [1] -0.02802804
```

```r
round (deltamu/mu*100,1)
```

```
## [1] -0.6
```

The sample mean (4.971972) is 0.6% lower than the theoretical mean (5). The difference is really minimal.

### Sample Variance

```r
Var_theo <- sd^2 / n
Var_sample <- var(mean_sample)

Var_theo
```

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

```r
Var_sample
```

```
## [1] 0.6157926
```

```r
deltavar = Var_sample - Var_theo
var_perc <- round(deltavar/Var_theo*100,2)
```

The Theoretical Variance (0.625) and is -1.47% lower than the Sample Variance is (0.6157926).

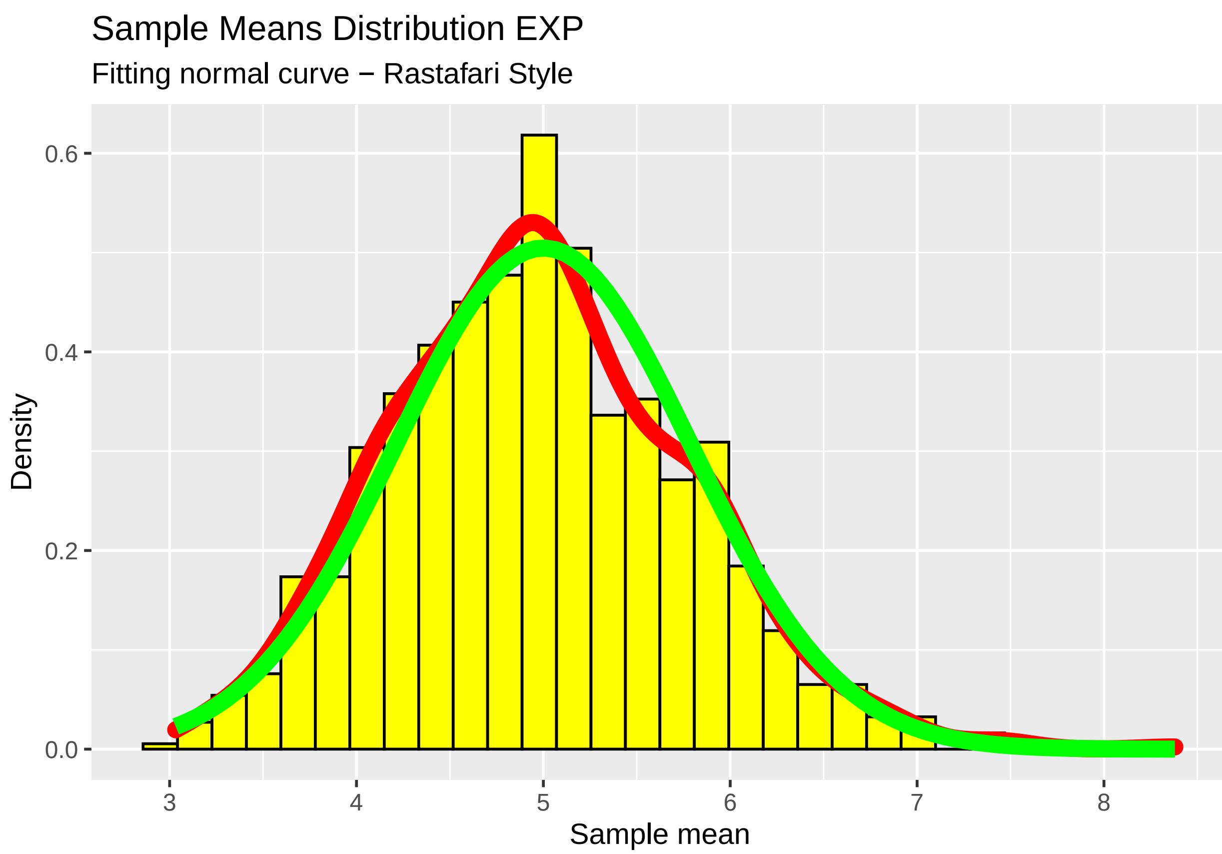
### Distribution

```r
library(ggplot2)

sd <- sqrt(Var_theo)

ggplot(as.data.frame(mean_sample), aes(as.data.frame(mean_sample)[,1]))+
        geom_histogram(aes(y=..density..), position="identity", fill="yellow", col="black")+
        geom_density(colour="red", size=3)+
        stat_function(fun = dnorm, colour = "green", args = list(mean = mu, sd = sd), size=3)+
        ggtitle ("Sample Means Distribution EXP")+
        labs(subtitle = "Fitting normal curve - Rastafari Style")+
        xlab("Sample mean")+
        ylab("Density")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




As highlighted by the plot, we can see a small discrepancy between the sample distribution and the theoretical one.

### Conclusion

We can conclude that the simulation performed with the rexp R function has produced a random dataset globally normal.

## PART 2

### Overview

Analyse the ToothGrowth dataset :

The tooth growth data set is the length of the odontoblasts (teeth) in each of 10 guinea pigs at different Vitamin C dosage levels with two delivery methods.

The procedure will consists in: - Doing Exploratory Data Analyses - Provide a summary - Perform confidence interval - State some conclusions.

### Load the Dataset

```r
library(datasets)
mydata <- ToothGrowth
```

## EDA

### Structure of the dataset

```r
str(mydata)
```

```
## '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 ...
```

The dataframe has 60 observations and 3 columns.

- len : Tooth length
- supp : Supplement Type -> "OJ" for orange Juice -> "VC" for Ascorbic Acid (???)
- dose : 3 levels of Vitamin C dosage (0.5, 1, 2 mg)

```r
sum(is.na(mydata))
```

```
## [1] 0
```

There is no NA in the dataset.

```r
summary(mydata$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   13.07   19.25   18.81   25.27   33.90
```

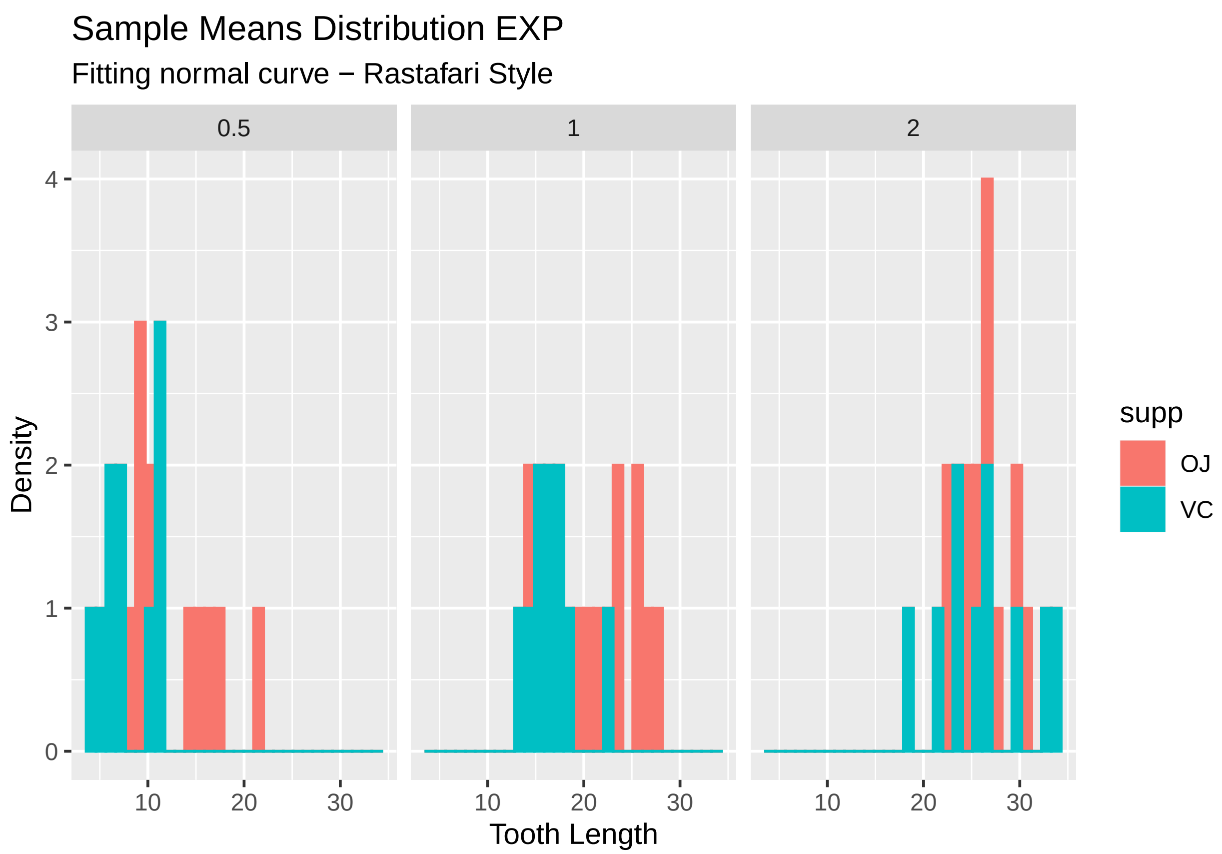
### Plot

A first blind graph to show roughly to have an idea of the content.

```r
qplot(data = mydata, x = mydata$len , facets = . ~ mydata$dose, )+
        aes(color = supp, fill= supp) +
        ggtitle ("Sample Means Distribution EXP")+
        labs(subtitle = "Fitting normal curve - Rastafari Style")+
        xlab("Tooth Length")+
        ylab("Density")
```

```
## Warning: Use of `mydata$len` is discouraged. Use `len` instead.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




We can directly observed that when the doses of Vitamin C are increasing the teeth are increasing also. Let's summarize closely the dataset:

### Summarize the Tooth Length by Dose and Supp

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
sum_tot <- mydata %>%
    group_by(supp,dose) %>%
    summarize(mean_len_tooth=mean(len), sd_len_tooth=sd(len), count = n())
```

```
## `summarise()` regrouping output by 'supp' (override with `.groups` argument)
```

```r
print(sum_tot)
```

```
## # A tibble: 6 x 5
## # Groups:   supp [2]
##   supp   dose mean_len_tooth sd_len_tooth count
##   <fct> <dbl>          <dbl>        <dbl> <int>
## 1 OJ      0.5          13.2          4.46    10
## 2 OJ      1            22.7          3.91    10
## 3 OJ      2            26.1          2.66    10
## 4 VC      0.5           7.98         2.75    10
## 5 VC      1            16.8          2.52    10
## 6 VC      2            26.1          4.80    10
```

### Summarize the Tooth Length by Supp only

```r
library(dplyr)

sum_supp <- mydata %>%
    group_by(supp) %>%
    summarize(mean_len_tooth=mean(len), sd_len_tooth=sd(len), count = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
print(sum_supp)
```

```
## # A tibble: 2 x 4
##   supp  mean_len_tooth sd_len_tooth count
##   <fct>          <dbl>        <dbl> <int>
## 1 OJ              20.7         6.61    30
## 2 VC              17.0         8.27    30
```

### Summarize by Dosage level

```r
sum_dose <- mydata %>%
    group_by(dose) %>%
    summarize(mean_len_tooth=mean(len), sd_len_tooth=sd(len), count = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
print(sum_dose)
```

```
## # A tibble: 3 x 4
##    dose mean_len_tooth sd_len_tooth count
##   <dbl>          <dbl>        <dbl> <int>
## 1   0.5           10.6         4.50    20
## 2   1             19.7         4.42    20
## 3   2             26.1         3.77    20
```

Clearly, the teeth length means are greater when the doses of vitamin C increase. Same observation when we administrate the treatment with the Orange Juice.

## Confidence Interval/Hypothesis

Let's run t.test for the different configurations possible of the data:

### Supplement Method Comparison

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

**At all dosage levels:**

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

```
t.test(len ~ supp, paired=FALSE, var.equal=FALSE, data=mydata[mydata$dose==0.5,])
```

**At 0.5mg dosage level:**

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

```
t.test(len ~ supp, paired=FALSE, var.equal=FALSE, data=mydata[mydata$dose==1,])
```

**At 1mg dosage level:**

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
```

```
t.test(len ~ supp, paired=FALSE, var.equal=FALSE, data=mydata[mydata$dose==2,])
```

**At 2mg dosage level:**

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

We can find a significant difference between the 2 supplement methods for the 0.5 and 1mg dosage. No significant difference at 2mg.

### Dosage Comparison

So we will compare the different dosage with OJ:

```
t.test(len ~ dose, paired=FALSE, var.equal=FALSE, data=mydata[mydata$dose<2 & mydata$supp=="OJ",])
```

**Compare 0.5 to 1 for OJ:**

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -5.0486, df = 17.698, p-value = 8.785e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13.415634  -5.524366
```

```
## sample estimates:
## mean in group 0.5   mean in group 1
##             13.23             22.70
```

```r
t.test(len ~ dose, paired=FALSE, var.equal=FALSE, data=mydata[mydata$dose>0.5 & mydata$supp=="OJ",])
```

**Compare 1 to 2 for OJ:**

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -2.2478, df = 15.842, p-value = 0.0392
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.5314425 -0.1885575
## sample estimates:
## mean in group 1 mean in group 2
##           22.70           26.06
```

```r
t.test(len ~ dose, paired=FALSE, var.equal=FALSE, data=mydata[mydata$dose<2 & mydata$supp=="VC",])
```

**Compare 0.5 to 1 for VC:**

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -7.4634, df = 17.862, p-value = 6.811e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.265712  -6.314288
## sample estimates:
## mean in group 0.5   mean in group 1
##              7.98             16.77
```

```r
t.test(len ~ dose, paired=FALSE, var.equal=FALSE, data=mydata[mydata$dose>0.5 & mydata$supp=="OJ",])
```

**Compare 1 to 2 for VC:**

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
```

```
## t = -2.2478, df = 15.842, p-value = 0.0392
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.5314425 -0.1885575
## sample estimates:
## mean in group 1 mean in group 2
##           22.70           26.06
```

For all the tests comparing the dosage, the confidence interval is always excluding 0. The differences between the dosage levels are significant.

## Statements of the study

### Conclusions

- The Vitamin C is correlated to the tooth growth with high confidence (95%) and this whatever the supplement method.
- The Orange Juice is providing better tooth growth at low dosage (=<1mg) than Ascorbic Acid. There is no significant difference at 2mg.

### Assumptions:

- The measurement are not paired
- The variances are not equal
- The test subjects were randomly selected and independants.