# Statistical Inference Project - Data Analysis

ElisaRMA

5/6/2021

## Overview

The objective of this second part of the project is to load and analyze the ToothGrowth data, a dataset available in R. This dataset describes the treatment of 60 ginea pigs with ascorbic acid and orange juice (`supp`), at different doses (`dose`), monitoring the lenght of the odontoblasts (cells responsible for tooth growth) (`len`).

All the code used in this project is described in the **Appendix**

### Loading ToothGrowth data

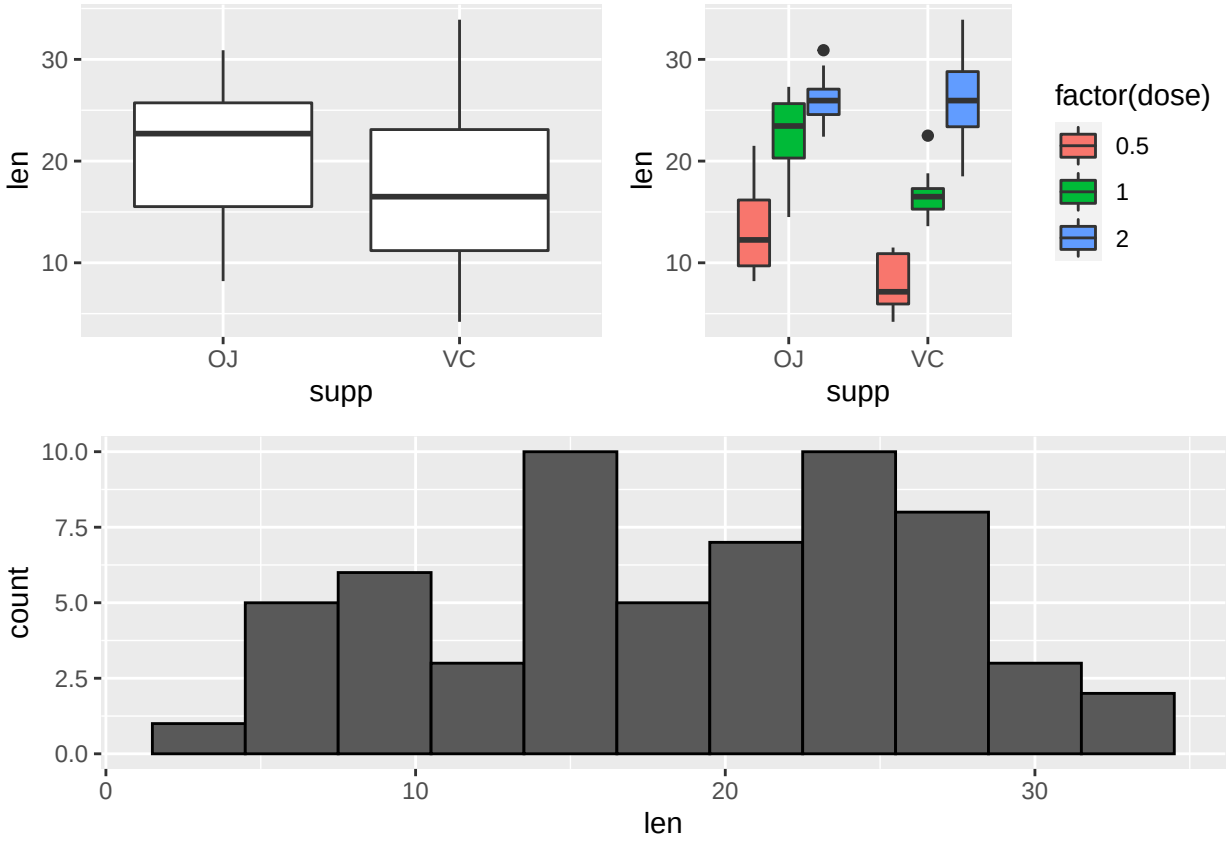
To load the ToothGrowth data the function `data()` was used. Then, to get a first overview of the data, the `head` function was used.

### Exploratory data analyses and summary of ToothGrowth data

After loading the data, some exploratory analysis, as instructed, were performed.The first function used was `summary()`. With this function we could get a overview of each column, and the structure of the data.

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

Then, the data was visualized using two boxplots and a histogram, to understand how the treatment and doses affected tooth length and how the data was distributed:




## Comparing tooth growth by supp and dose

Based on the data, two important questions arise: 'Is there a difference between treatments?' and 'Is there a difference between doses in each treatment?'

To answer such questions, simple t-tests can be applied. For the first question, our hypotheses could be:

- H0:$\mu VC = \mu OJ$
- Ha: $\mu OJ \neq \mu VC$

As observed, the p-value was 0.0606345 and the confidence interval was -0.1710156, 7.5710156

For the second question, our hypothesis could be, for each supp:

- H0: all $\mu$ are equal
- Ha1 and 4:$\mu dose1 \neq \mu dose0.5$
- Ha2 and 5:$\mu dose2 \neq \mu dose0.5$
- Ha3 and 6:$\mu dose2 \neq \mu dose1$

```
##        test supp           id                 value
## 1  conf.int   OJ mu1 != mu0.5          8.784919e-05
## 2   p.value   OJ mu1 != mu0.5  -13.415634, -5.524366
## 3  conf.int   OJ mu2 != mu0.5          1.323784e-06
## 4   p.value   VC mu2 != mu0.5  -16.335241, -9.324759
## 5  conf.int   VC   mu2 != mu1            0.03919514
```

```
## 6   p.value   VC   mu2 != mu1 -6.5314425, -0.1885575
## 7  conf.int   OJ mu1 != mu0.5           6.811018e-07
## 8   p.value   OJ mu1 != mu0.5  -11.265712, -6.314288
## 9  conf.int   OJ mu2 != mu0.5           4.681577e-08
## 10  p.value   VC mu2 != mu0.5   -21.90151, -14.41849
## 11 conf.int   VC   mu2 != mu1           9.155603e-05
## 12  p.value   VC   mu2 != mu1  -13.054267, -5.685733
```

### Conclusions and Assumptions

Based on the p-values and the confidence intervals observed, it was not possible reject the null hypothesis, given that the p value is $> 0.05$ and the confidence interval crossed 0. Therefore, there is no difference between supplements (Orange Juice or Arcorbic Acid). However, taking into consideration the doses used, both p value and confidence intervals allow us to reject the null hypothesis, suggesting that the higher the dose, the larger the effect, for both supplements.

To draw such conclusions, the assumptions made were:

- The treatment was randomized on the 60 guinea pigs
- The n used in this study was enough to represent the entire population of guinea pigs, allowing us to generalize the results of the study.
- The variance was assumed to be unequal (var.equal = FALSE was kept as default)
- The data are iid Gaussian and as it increases, it gets closer to a normal distribution.

## Appendix

```r
# load the data
data("ToothGrowth")

#summary of ToothGrowth
summary(ToothGrowth)

# PLOT

# Loads ggplot2 and ggpubr
library(ggplot2)
library(ggpubr)
# Creates two boxplots, one by supp and other by supp and dose.
bx1 <- ggplot(ToothGrowth, aes(y=len, x=supp))+ geom_boxplot()
bx2 <- ggplot(ToothGrowth, aes(y=len, x=supp))+ geom_boxplot(aes(fill=factor(dose)))
# Creates a histogram to show the distribution of the data
ht <- ggplot(ToothGrowth, aes(len)) + geom_histogram(binwidth=3, colour='black')
# Arranges the data into one single plot, with the histogram below the two boxplots:
# So the boxplots are arranged into a plot with two columns, side by side.
# Then, this plot is arranged with the histogram, one below the other, so, 2 rows.
ggarrange(ggarrange(bx1, bx2, ncol = 2), ht, nrow = 2)

# T-test between the supplements
ttest <-t.test(len~supp, ToothGrowth)

# Extracts both pvalue and confidence interval
```

```
ttest$p.value
ttest$conf.int

# Loads tidyr and dplyr
library(tidyr); library(dplyr)

# Subsets the ToothGrowth into the doses, two by two, as the ttest will be performed
ha14 <- subset(ToothGrowth, dose %in% c(0.5, 1))
ha25 <- subset(ToothGrowth, dose %in% c(0.5, 2))
ha36<- subset(ToothGrowth, dose %in% c(1, 2))

# Within that subset, OJ and VC are separated.
ha1OJ <- subset(ha14, supp=='OJ')
ha2OJ <- subset(ha25, supp=='OJ')
ha3OJ <- subset(ha36, supp=='OJ')
ha4VC <- subset(ha14, supp=='VC')
ha5VC <- subset(ha25, supp=='VC')
ha6VC<- subset(ha36, supp=='VC')

# T-test for each combination of dose, within each supp
t1 <- t.test(len~dose,ha1OJ)
t2 <- t.test(len~dose,ha2OJ)
t3 <- t.test(len~dose,ha3OJ)
t4 <- t.test(len~dose,ha4VC)
t5 <- t.test(len~dose,ha5VC)
t6 <- t.test(len~dose,ha6VC)

# Creates a data frame to display all pvalues and conf.intervals
test <- data.frame(test=c("conf.int", "p.value"),
             supp=c("OJ", "OJ", "OJ", "VC", "VC", "VC"),
             id = c('mu1 != mu0.5', 'mu1 != mu0.5','mu2 != mu0.5', 'mu2 != mu0.5',
                    'mu2 != mu1','mu2 != mu1','mu1 != mu0.5', 'mu1 != mu0.5',
                    'mu2 != mu0.5', 'mu2 != mu0.5', 'mu2 != mu1', 'mu2 != mu1'))
test$value <- list(t1$p.value, t1$conf.int, t2$p.value, t2$conf.int, t3$p.value,
                   t3$conf.int, t4$p.value, t4$conf.int, t5$p.value, t5$conf.int,
                   t6$p.value, t6$conf.int )
```