The Effect of Vitamin C on Tooth Growth in Guinea Pigs

A basic data analysis

Diego Rabatone Oliveira

Overview

This is the part 2 of the final project for the Statistical Inference Course, offered by Johns Hopkins University at Coursera, as part of the Data Science Specialization. This project consists a basic exploratory data analysis of the ToothGrowth dataset (from R datasets package).

Loading the data and doing some exploratory analysis

Let's start by loading our dataset:

```
## Observations: 60
## Variables: 3
## $ len  <dbl> 4.2, 11.5, 7.3, 5.8, 6.4, 10.0, 11.2, 11.2, 5.2, 7.0, 16....
## $ supp <fctr> VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, ...
## $ dose <dbl> 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 1.0, 1....
```

As we can see, this dataset contains 60 observations with 3 variables (**len** (dbl), **sup** (fact: *OJ* and *VC*) and **dose** (dbl)). Looking at the `help(ToothGrowth)`, we can see that **dose** is a measure in **mg**, growth does not present a unit measure, and *OJ* stands for *OrangeJuice* and *VC* for *Ascorbic Acid*, both being delivery methods from Vitamin C. And all these data came from a study of the "*Effect of Vitamin C on Tooth Growth in Guinea Pigs*".

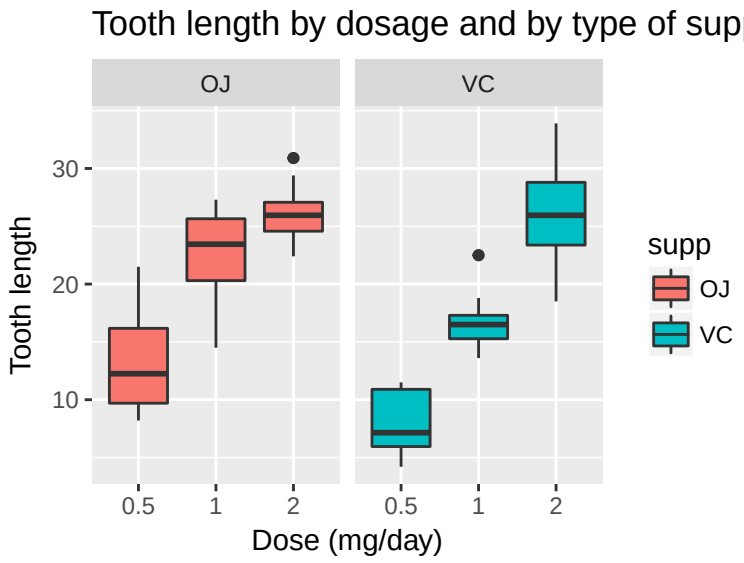
Dataset Summary

Let's now do some summaries from our dataset:

```
##       len       supp         dose
##  Min.   : 4.2   OJ:30   Min.   :0.50
##  1st Qu.:13.1   VC:30   1st Qu.:0.50
##  Median :19.2           Median :1.00
##  Mean   :18.8           Mean   :1.17
##  3rd Qu.:25.3           3rd Qu.:2.00
##  Max.   :33.9           Max.   :2.00
```

The **dose** variable, despite being numeric, only presents the values (0.5, 1 and 2). Thus, we will convert it to a factor variable in order to threat them as groups.

Now let's do some boxplotting to observe the overall behaviour of these three variables togheter, and we are also going to calculate the mean and standard deviation of the tooth *length*, grouped by *dose* and *supp*, only



dose and only *supp*:

```
## Source: local data frame [11 x 4]
## Groups: dose [4]
##
##      dose   supp  mean    sd
##    <fctr> <fctr> <dbl> <dbl>
## 1     0.5     OJ 13.23  4.46
## 2     0.5     VC  7.98  2.75
## 3     0.5     NA 10.61  4.50
## 4       1     OJ 22.70  3.91
## 5       1     VC 16.77  2.52
## 6       1     NA 19.73  4.42
## 7       2     OJ 26.06  2.66
## 8       2     VC 26.14  4.80
## 9       2     NA 26.10  3.77
## 10     NA     OJ 20.66  6.61
## 11     NA     VC 16.96  8.27
```

From the boxplot above and the table, it looks like there is a tendency that the higher the dosage, the larger the tooth length. But let's do some tests to verify or refute this hypothesis

Confidence Interval and Hypothesis Testing

Our dataset is not too large, n=60, and when grouping the observations it will be even smaller (`n/2` on the best case, grouping only by *supp*). So, our standard errors tends to be bigger than what we would want. Therefore, we will need to use the `T distribution` for our hypotesis tests. We are assuming that the given data respects the `Independent and identically distributed random variables (i.i.d)` criteria, despite no information about this was given.

We are going to start with two tests, and, depending on its results we will do another complementary test.

Supplement types (supp)

Here our hypothesis is that there are differences between the mean tooth length among the groups that received different supplements: $H_0 : \mu_{OJ} \neq \mu_{VC}$. The test `t.test(formula=len ~ supp, data=TG, var.equal =`

`FALSE)` results in a `p-value` of 0.061, which is larger than the significance value of 0.05. Being so, we *fail to reject the null hypothesis* that there are differences on the means between the *OJ* and the *VC* groups.

Dosage (dose)

Now, our hypothesis is that there are differences between the mean tooth length among the groups that received different dosages, independently of the supplement type. Here, as we have three different levels of dosage, we will need to do three t-tests to cover the possible combinations. The results of these tests are below:

```
##   null hypothesis              p.value          conf.low         conf.high
## 1     u_0.5-u_1=0 1.26830072017385e-07 -11.9837812579016 -6.27621874209841
## 2     u_0.5-u_2=0 4.39752495936323e-14 -18.1561665388306 -12.8338334611694
## 3       u_1-u_2=0  1.9064295136718e-05 -8.99648051689202 -3.73351948310799
```

So, on the three tests done the `p-values` where smalled than the significance level of 0.05, which means that we **reject the null hypothesis** of not existing difference between the observations while varying the dosage. So, our data have strong evidence that the vitamin C dosage influences the average tooth length on guinea pigs with 95% of confidence.

Conclusions

Considering the tests done and graphs evaluated, we can conclude that, statistically speaking, the Guinea Pigs tooth length presents a strong relationship with the vitamin C consumption in terms of dose levels, with a 95% confidence, and does not present relationship with how this vitamin is ingested.

The source code used to procude this report can be found at: https://github.com/diraol/CourseraProjects

Appendix (Codes) Below are the codes used to buid this report:

```
options(digits = 3)
library(tidyr)
library(dplyr)
library(ggplot2)
library(datasets)
data("ToothGrowth")
TG <- ToothGrowth
glimpse(TG)
summary(TG)
TG$dose <- as.factor(TG$dose)
ggplot(TG, aes(x=factor(dose), y=len)) +
  facet_grid(.~supp) +
  geom_boxplot(aes(fill = supp)) +
  labs(title="Tooth length by dosage and by type of supplement",
    x="Dose (mg/day)",
    y="Tooth length")
head(bind_rows(TG %>% group_by(dose, supp) %>% summarize(mean=mean(len), sd=sd(len)),
      TG %>% group_by(dose) %>% summarize(supp=NA, mean=mean(len), sd=sd(len)),
      TG %>% group_by(supp) %>% summarize(dose=NA, mean=mean(len), sd=sd(len))) %>%
  arrange(dose, supp), n=11)
test1 <- t.test(formula=len ~ supp, data=TG, var.equal = FALSE)
test2.a <- TG %>% filter(dose==0.5 | dose==1) %>% t.test(formula=len ~ dose, data=., var.equal=FALSE)
test2.b <- TG %>% filter(dose==0.5 | dose==2) %>% t.test(formula=len ~ dose, data=., var.equal=FALSE)
test2.c <- TG %>% filter(dose==1 | dose==2) %>% t.test(formula=len ~ dose, data=., var.equal=FALSE)
ta <- c('u_0.5-u_1=0', test2.a$p.value, test2.a$conf.int[1], test2.a$conf.int[2])
tb <- c('u_0.5-u_2=0', test2.b$p.value, test2.b$conf.int[1], test2.b$conf.int[2])
tc <- c('u_1-u_2=0', test2.c$p.value, test2.c$conf.int[1], test2.c$conf.int[2])
cnames <- c('null hypothesis', 'p.value', 'conf.low', 'conf.high')
result <- data.frame(0,0,0,0)
colnames(result) <- cnames
result[1,] <- ta
result[2,] <- tb
result[3,] <- tc
result
```