Statistical Inference Course Project Part 2

Hernan Solano

March 05 2018

Overview

In this report, we will analyze the ToothGrowth data in the R datasets package. This data set will be explored to understand its structure using some exploratory visualizations, then some hypothesis tests will be performe in order to draw some conclusions.

The description of the data set can be found here: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ToothGrowth.html

```
library(datasets)
library(ggplot2)
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
```

```
library(tibble)
library(ggthemes)
dataset <- as.tibble(ToothGrowth)
```

Summary of the data

According to the datasets package description the tooth growth dataset shows measures of tooth length for 60 guinea pigs. They recieved one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) administrated by one of two delivery methods: orange juice (OJ) or vitamin C (VC). Therefore we have 3 variables: - len: a numeric measure of the tooth length - supp: a categoric/factor variable that indicates the supplement type OJ or VC - dose: a numeric variable indicating the level of vitamin C administrated in milligrams/day

```
##General infor about the dataset
str(dataset)
```

```
## Classes 'tbl_df', 'tbl' and '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 ...
```

In the following lines one can find the summary statistics as well as the first and last rows of the dataset:

```
##Summary statistics of the variables
summary(dataset)
```

```
##       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
```

```
##First 10 rows of the dataset
head(dataset)
```

```
## # A tibble: 6 x 3
##     len supp   dose
##   <dbl> <fct> <dbl>
## 1  4.20 VC    0.500
## 2 11.5  VC    0.500
## 3  7.30 VC    0.500
## 4  5.80 VC    0.500
## 5  6.40 VC    0.500
## 6 10.0  VC    0.500
```

```
##Last 10 rows of the dataset
tail(dataset)
```

```
## # A tibble: 6 x 3
##     len supp   dose
##   <dbl> <fct> <dbl>
## 1  24.8 OJ       2.
## 2  30.9 OJ       2.
## 3  26.4 OJ       2.
## 4  27.3 OJ       2.
## 5  29.4 OJ       2.
## 6  23.0 OJ       2.
```

In the following table count of the guinea pigs by any possible combination of the supplement and dose:

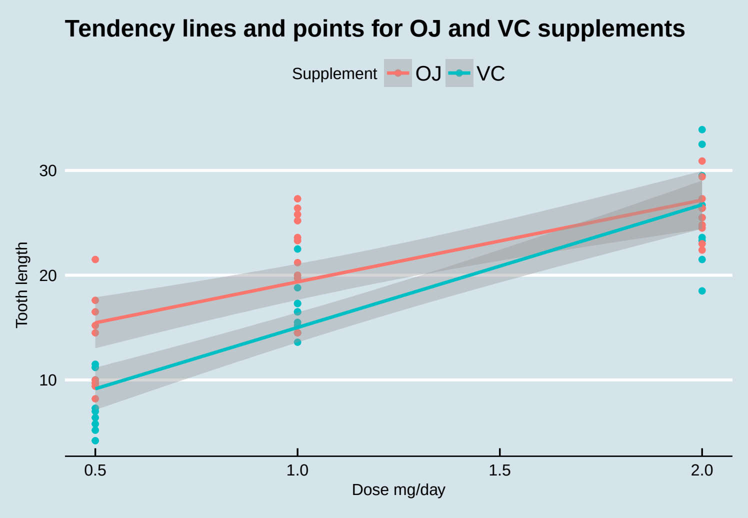
```
##Number of guinea pigs by combination of supplement and dose
table(select(dataset,c("supp","dose")))
```

```
##     dose
## supp 0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

Now it is time for some visualizations:

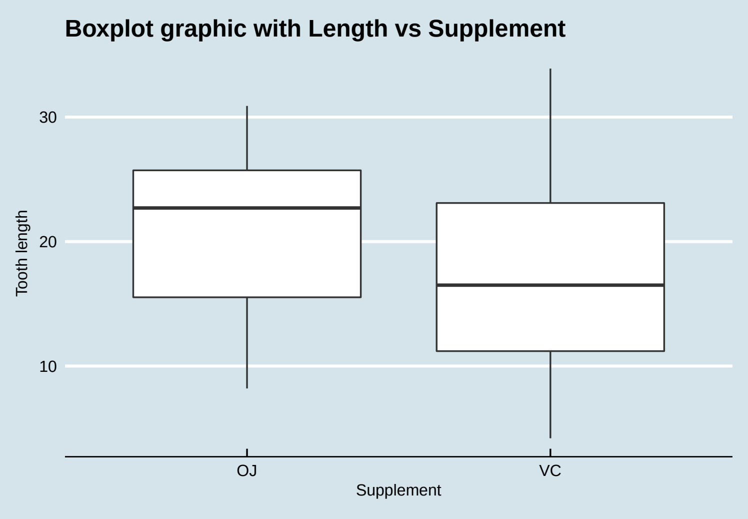
- In this figure one can see clearly that the bigger the dose on vitamin C the , no matter what suplement is used the

```
g <- ggplot(data = dataset)
##Point plot with linear tendenci lines
g + geom_point(aes(x = dose,y=len,colour = supp)) +
    geom_smooth(aes(x = dose,y=len,colour = supp),method = "lm") +
    theme_economist() +
    labs(title = "Tendency lines and points for OJ and VC supplements",x = "Dose mg/day",
         y = "Tooth length",colour= "Supplement")
```



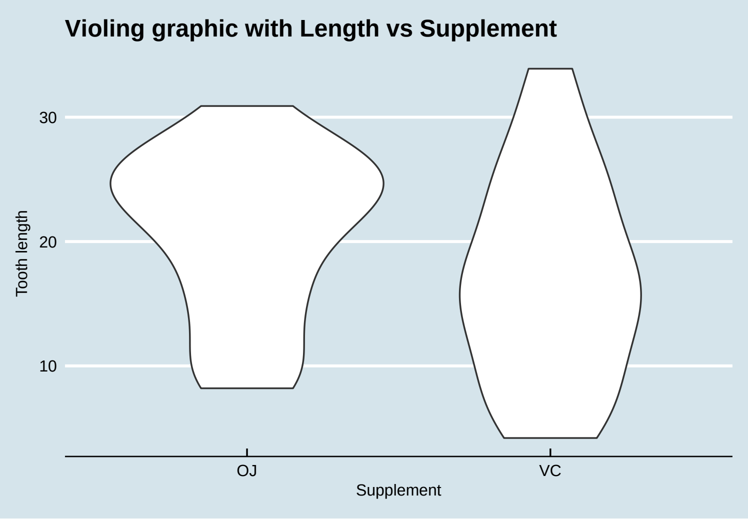
- The figure bellow another exploratory visualization that makes me think, in complement with the figure above, that the Orange Juice supplement could yield to best results than the Vitamin C supplement.

```
g + geom_boxplot(aes(x = supp,y = len)) + theme_economist() +
    labs(title = "Boxplot graphic with Length vs Supplement",y = "Tooth length",
         x = "Supplement")
```



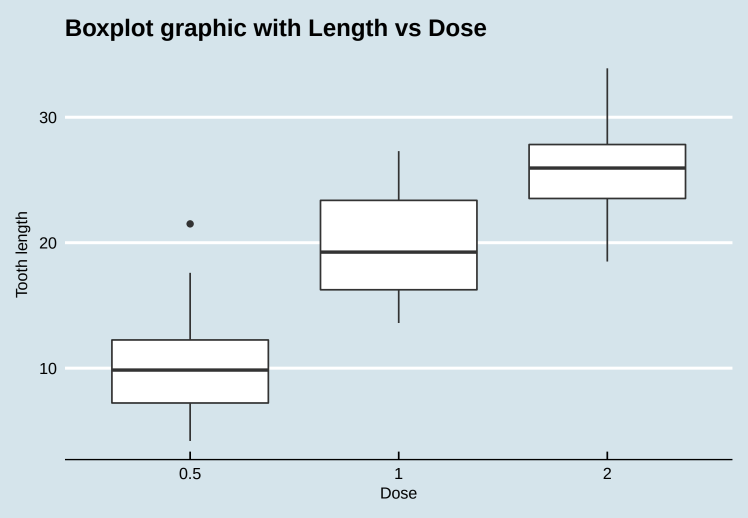
- The violin graph also show us some interesting shapes that makes think that the OJ gives better results than the VC.

```
g + geom_violin(aes(x=supp,y= len)) + theme_economist() +
    labs(title = "Violing graphic with Length vs Supplement", x = "Supplement",
         y = "Tooth length")
```

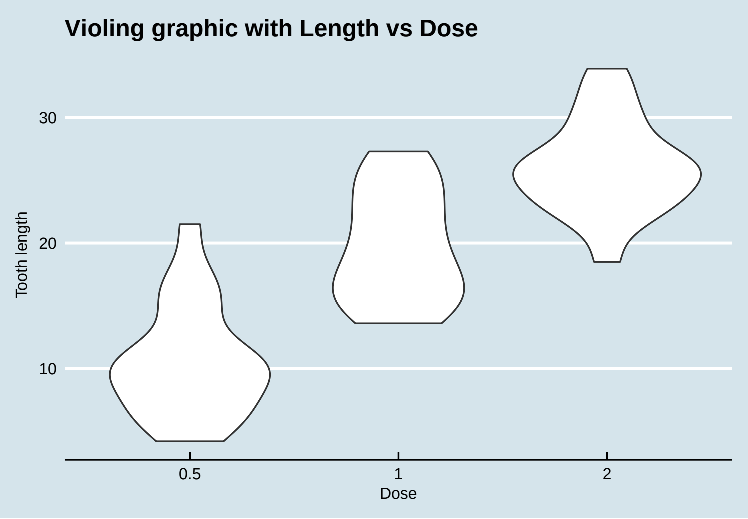


- By running the same comparisons with the Length vs Dose graphs, one may think that with a higher dose longer teeth will be attained.

```
g + geom_boxplot(aes(x = as.factor(dose),y = len)) + theme_economist() +
    labs(title = "Boxplot graphic with Length vs Dose",y = "Tooth length",x = "Dose")
```



```
g + geom_violin(aes(x= as.factor(dose),y= len)) + theme_economist() +
    labs(title = "Violing graphic with Length vs Dose", x = "Dose",
         y = "Tooth length")
```



- To wrap up, so far we have this hypothesis to test:
  - OJ supplement gives longer teeth than VC supplement does.
  - The bigger the dose, the longer the teeth.

Hypothesis tests

First, we are going to compare tooth growth by supplement, in order to do this we are going to perform a two sided, two sampled t test for the mean. In this case we are assuming different variances and our hypothesis could be stated as follow:

- $H_0 : \bar{X}_{OJ} = \bar{X}_{VC}$
- $H_a : \bar{X}_{OJ} \neq \bar{X}_{VC}$

```
test_supp <- t.test(data = dataset,len ~ supp,paired = FALSE, var.equal = FALSE)
print(test_supp)
```

```
##
##  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
```

With this p-value equal to 0.0606 we fail to reject the null hypothesis at a 5% level of significance ($\alpha = 0.05$). Therefore, we can safely assume that the means are equivalent eventhough at first sight it does not seem like this.

Now we are going to conduct 3 hypothesis tests, one for every combination of dose groups and length. The following code perform all the tests:

```
dose05 <- filter(dataset,dose == 0.5) %>% select(len)
dose1 <- filter(dataset,dose == 1) %>% select(len)
dose2 <- filter(dataset,dose == 2) %>% select(len)
```

```
test_1_05 <- t.test(dose1- dose05,paired = FALSE, var.equal = FALSE,
                    alternative = "greater")
test_2_05 <- t.test(dose2- dose05,paired = FALSE, var.equal = FALSE,
                    alternative = "greater")
test_2_1 <- t.test(dose2 - dose1,paired = FALSE, var.equal = FALSE,
                   alternative = "greater")
```

- Our first case is the hypothesis test of the dose of 1 mg/day and the 0.5 mg/day:
  - $H_0 : \bar{X}_1 - \bar{X}_{0.5} < 0$
  - $H_a : \bar{X}_1 - \bar{X}_{0.5} > 0$

Note: we are assuming that the variances are different and that the guinea pigs are different.

```
print(test_1_05)
```

```
##
##  One Sample t-test
##
## data:  dose1 - dose05
## t = 6.9669, df = 19, p-value = 6.127e-07
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  6.863996      Inf
## sample estimates:
## mean of x
##      9.13
```

- The second case is the hypothesis test of the dose of 2 mg/day and the 0.5 mg/day:
  - $H_0 : \bar{X}_2 - \bar{X}_{0.5} < 0$
  - $H_a : \bar{X}_2 - \bar{X}_{0.5} > 0$

Note: we are assuming that the variances are different and that the guinea pigs are different.

```
print(test_2_05)
```

```
##
##  One Sample t-test
##
## data:  dose2 - dose05
## t = 11.291, df = 19, p-value = 3.595e-10
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  13.12216      Inf
## sample estimates:
## mean of x
##    15.495
```

- The third case is the hypothesis test of the dose of 2 mg/day and the 1 mg/day:
  - $H_0 : \bar{X}_2 - \bar{X}_1 < 0$
  - $H_a : \bar{X}_2 - \bar{X}_1 > 0$

Note:we are assuming that the variances are different and that the guinea pigs are different.

```
print(test_2_1)
```

```
##
##  One Sample t-test
##
```

```
## data:  dose2 - dose1
## t = 4.6046, df = 19, p-value = 9.671e-05
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  3.974821      Inf
## sample estimates:
## mean of x
##     6.365
```

In this three cases the p-value was less than 0.05, therefore we are able reject all individual null hypothesis at a 5% level of significance ($\alpha = 0.05$).

Just to be sure, but I have to admit I do not know if this is used in this cases, I will use the bonferroni correction. So if a p-value calculated is less than 0.017, it will be called significant so the null hypothesis will be rejected:

```
p_values <- c(test_1_05$p.value,test_2_05$p.value,test_2_1$p.value)
##With the Bonferroni correction alpha_corrected = 0.017
p_values < 0.017
```

```
## [1] TRUE TRUE TRUE
```

Conclusions

To sum up I am assuming that the variances of the compared groups are different, also that I can use the Bonferroni correction in this case. With the analysis performed I can conclude the following:

- There is no difference in the tooth length between the OJ and VC supplement groups, but there is enough evidence to support that a higher dose yields to a greater length of the guinea pigs' teeth.