Statistical Inference - Basic Inferential Data Analysis

H.Harvey

25 December 2015

## Introduction

The purpose of this project is to perform basic exploratory and inferential data analysis using the ToothGrowth data in the R datasets package.

This project should cover the follwing task: - perform some basic exploratory data analyses - provide a basic summary of the data. - use confidence intervals and hypothesis tests to compare tooth growth by supp and dose. - state your conclusions and the assumptions needed for your conclusions

## Setting the environment

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(ggplot2)  
data(ToothGrowth)

## Data structure:

A data frame with 60 observations on 3 variables.

[,1] "len" numeric Tooth length [,2] "supp" factor Supplement type (VC-Vitamin C or OJ-Orange Juice). [,3] "dose" numeric Dose in milligrams/day (0.5 / 1.0 / 2.0)

## Initial data exploration

# Looking at the dataset  
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 ...

# Validating the structure  
head(ToothGrowth)

## len supp dose  
## 1 4.2 VC 0.5  
## 2 11.5 VC 0.5  
## 3 7.3 VC 0.5  
## 4 5.8 VC 0.5  
## 5 6.4 VC 0.5  
## 6 10.0 VC 0.5

# Validating the number of rows  
nrow(ToothGrowth)

## [1] 60

# Getting a statistical summary of the variables  
summary(ToothGrowth)

## 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

# Checking that no missing data exist  
sum(is.na(ToothGrowth))

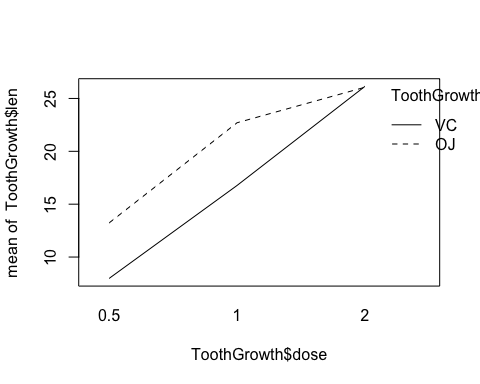
## [1] 0

# Representing the numbers of case per dose & supplement type to validation # the distribution of cases.  
table(ToothGrowth$supp,ToothGrowth$dose)

##   
## 0.5 1 2  
## OJ 10 10 10  
## VC 10 10 10

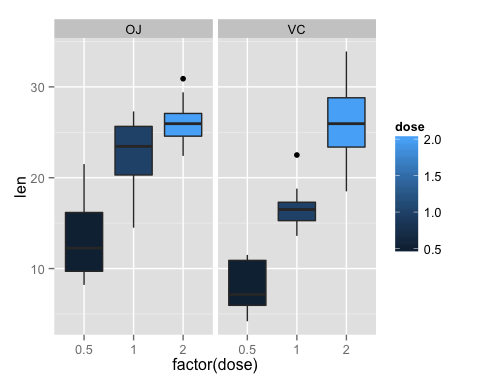
The dataset is split equally within the supplement type (30 observations each) and dose (10 observations each). For each supplement type, and a given dose, 10 subjects received Orange Juice and another received Vitamin C.

# Visualzating the variable interaction between lenght and dose/supplement # type  
interaction.plot(ToothGrowth$dose, ToothGrowth$supp, ToothGrowth$len)



## Graphical representation of the data

q <- ggplot(ToothGrowth,aes(factor(dose),len)) +  
 geom\_boxplot(aes(fill=dose)) + facet\_grid(.~supp)  
print(q)

 From the above graph, Orange Juice in doses of 0.5 and 1.0mg seems to support tooth growth over Vitamin C. However, at higher dosage (2.0mg), there doesn't seems to be one supplement better than the other.

## Hypotheses

Null Hypothesis (H0): Orange Juice and Vitamin C have the same effect on tooth growth.

Alternative Hypothesis (Ha): Orange Juice and Vitamin C do not have the same effect on tooth growth.

A p-value of 0.05 or 5% was used to determine whether the null hypothesis (H0) is rejected or not. If p<0.05, the (H0) was rejected.

## Data preparation

# Calculating the mean by supplement type and doses for t.test reference  
stats\_df <-ToothGrowth %>%  
 group\_by(supp,dose) %>%  
 summarise(mean\_Len = mean(len), stddev\_Len = sd(len), n = n())  
# Calculating the mean by dose  
stats\_df %>% group\_by(dose) %>% summarise(mean\_d=mean(mean\_Len))

## Source: local data frame [3 x 2]  
##   
## dose mean\_d  
## 1 0.5 10.605  
## 2 1.0 19.735  
## 3 2.0 26.100

# Calculating the mean by supplement type  
stats\_df %>% group\_by(supp) %>% summarise(mean\_d=mean(mean\_Len))

## Source: local data frame [2 x 2]  
##   
## supp mean\_d  
## 1 OJ 20.66333  
## 2 VC 16.96333

## Hypotheses calculation

# t.test of by dose & supplement type  
t\_d0 <- ToothGrowth %>% filter(dose == 0.5)  
t\_d1 <- ToothGrowth %>% filter(dose == 1)  
t\_d2 <- ToothGrowth %>% filter(dose == 2)  
  
t.test(filter(t\_d0,supp=="OJ")$len,filter(t\_d0,supp=="VC")$len ,paired = FALSE, var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: filter(t\_d0, supp == "OJ")$len and filter(t\_d0, supp == "VC")$len  
## t = 3.1697, df = 18, p-value = 0.005304  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 1.770262 8.729738  
## sample estimates:  
## mean of x mean of y   
## 13.23 7.98

t.test(filter(t\_d1,supp=="OJ")$len,filter(t\_d1,supp=="VC")$len ,paired = FALSE, var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: filter(t\_d1, supp == "OJ")$len and filter(t\_d1, supp == "VC")$len  
## t = 4.0328, df = 18, p-value = 0.0007807  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 2.840692 9.019308  
## sample estimates:  
## mean of x mean of y   
## 22.70 16.77

t.test(filter(t\_d2,supp=="OJ")$len,filter(t\_d2,supp=="VC")$len ,paired = FALSE, var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: filter(t\_d2, supp == "OJ")$len and filter(t\_d2, supp == "VC")$len  
## t = -0.046136, df = 18, p-value = 0.9637  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3.722999 3.562999  
## sample estimates:  
## mean of x mean of y   
## 26.06 26.14

Dose (OJ vs VC) Lower Endpoint Upper Endpoint P-Value H 0.5 1.770 8.730 0.0053 reject 1.0 2.841 9.019 0.0008 reject 2.0 -3.723 3.563 0.964 do not reject

The p-value of the 0.5 and 1.0 dosages demonstrate that there is a difference in the tooth growth for each supplement type. However, at 2.0mg dosage,the p-value doesn't allow one to claim that a differene exist, or is not due to chance.

The confidence interval of the 0.5 and 1.0 dosages indicate that the Orange Juice supplement is more effective than the Vitamin C (positive interval). However, at higher dosage (2.0mg), the effect between supplement type seems to be equivalent due to the symmetrical value of the interval.

## Conclusion

The results of both the null hypothesis tests and confidence intervals demonstrate that doses of 0.5 and 1.0mg of Orange Juice is more effective in tooth growth. However, at higher dosage,2.0mg, neither supplement type is more effective.

Finally, it can be concluded that tooth growth increase with doses.