Statistical Inference Project Part 2: Data Analysis

## Overview

In this project we are going to explore and do some basic analysis of the ToothGrowth data set available in R.

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(datasets)  
RawData <- ToothGrowth

## Basic Summaries

We explore the structure of the data and the overall summary.

head(RawData)

## 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

summary(RawData)

## 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

str(RawData)

## '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 ...

Summary\_1 <- RawData %>%   
 group\_by(supp,dose) %>%  
 summarize(lenmean=mean(len), lensd=sd(len), count = n())  
print(Summary\_1)

## # A tibble: 6 x 5  
## # Groups: supp [?]  
## supp dose lenmean lensd 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

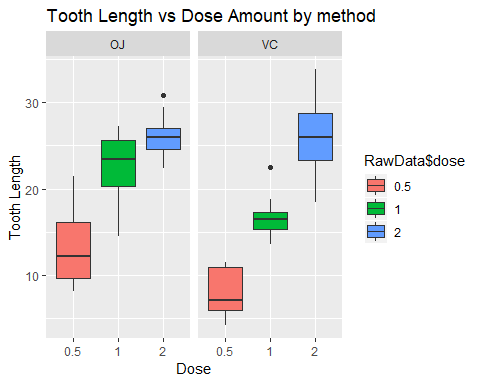
## Plots

First we transform dose to factor and then we make some box plots to see the difference in means for each supplement and dosage.

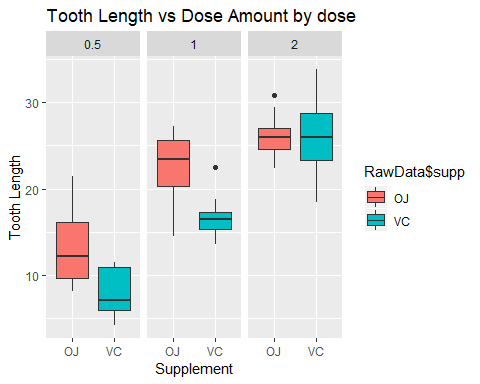
RawData$dose <- as.factor(RawData$dose)

Box plot showing the difference by methods

ggplot(data = RawData, aes(x=RawData$dose, y = RawData$len)) + geom\_boxplot(aes(fill = RawData$dose)) + labs(title = "Tooth Length vs Dose Amount by method", x = "Dose", y = "Tooth Length") + facet\_grid(~RawData$supp)

 Box plot showing the difference by dose

ggplot(data = RawData, aes(x=RawData$supp, y = RawData$len)) + geom\_boxplot(aes(fill = RawData$supp)) + labs(title = "Tooth Length vs Dose Amount by dose", x = "Supplement", y = "Tooth Length") + facet\_grid(~RawData$dose)



## 3. Using confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

1. Comparing if there is a difference between the supply method and length

t.test(RawData$len ~ RawData$supp, paired = FALSE,var.equal = FALSE, data = RawData)

##   
## Welch Two Sample t-test  
##   
## data: RawData$len by RawData$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

In this case, the confidence interval includes 0 and the p-value is 0.06 which is greater than 0.05. So we fail to reject the Null Hypothesis (there is no difference in means).

1. We can subdivide the data and do the analysis for each level of dosage (0.5,1 and 2)

2.1 Testing dosage = 0.5 (lowdose) for OJ and VC

lowdose <- RawData[RawData$dose == 0.5,]  
t.test(lowdose$len ~ lowdose$supp, paired = FALSE, var.equal = FALSE, data = lowdose)

##   
## Welch Two Sample t-test  
##   
## data: lowdose$len by lowdose$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

The confidence interval does not include 0 and the p-value is less than 0.05 so we can reject the Null hypothesis. Therefore there is a difference between the methods with a 0.5 dose.

2.2 Testing dosage = 1 (middose) for OJ and VC

middose <- RawData[RawData$dose == 1,]  
t.test(middose$len ~ middose$supp, paired = FALSE, var.equal = FALSE, data = middose)

##   
## Welch Two Sample t-test  
##   
## data: middose$len by middose$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

The confidence interval does not include 0 and the p-value is less than 0.05 so we can reject the Null hypothesis. Therefore there is a difference between the methods with a 1 dose.

2.3 Testing dosage = 2 (highdose) for OJ and VC

highdose <- RawData[RawData$dose == 2,]  
t.test(highdose$len ~ highdose$supp, paired = FALSE, var.equal = FALSE, data = highdose)

##   
## Welch Two Sample t-test  
##   
## data: highdose$len by highdose$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

The confidence interval includes 0 and the p-value is greater than 0.05 so in this case we fail to reject the Null hypothesis. Therefore there is no difference in means with a 2 level dose.

## 4. Conclussions

1. If we use the just the supply method, ignoring the level of dosage we conclude that there is no evidence to reject the Null hypothesis, ie no difference in means
2. If we desagregate the supply method for each level of dosage we can conclude that in 2 out of 3 we fail to reject the null hypothesis. This occurs in dosage levels of 0.5 and 1. For the last dosage level (highest), we fail to reject the null hypothesis, so there is no difference between OJ and VC methods.

## 5. Assumptions

1. Members of the sample population are representative of the entire population.
2. The variance between groups are different (var.equal = FALSE).
3. Population are independent there are no mixture between them.
4. The measures are not paired (paired = FALSE).