ToothGrowth Data Analysis

Pier Lorenzo Paracchini

## Overview

The goal is to analyze the ToothGrowth data in the R datasets package, performing some exploratory analysis, and then use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

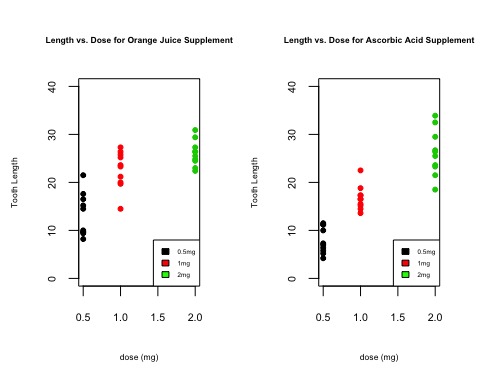
Description of the data from the documentation *"The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid)."*

## Getting and Exploring the data

require(datasets)  
rawData <- ToothGrowth  
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 ...

The rawdata includes 60 obs. of 3 features. Features are len the tooth length (numeric), supp the supplement type (factor) and finally dose, the dose in milligrams (numeric). There are 6 possible groups by supp and dose (see the plot below). For each supplement supp ("OJ" and "VC"), there are three different groups based on the dose values - 0.5mg, 1mg, 2mg.



And here some basic statics about the sample tooth length for each group.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Supp | Dose | No of Sample | Sample mean | Sample SD |
| "OJ" | 0.5 | 10 | 13.23 | 4.46 |
| "OJ" | 1.0 | 10 | 22.7 | 3.911 |
| "OJ" | 2.0 | 10 | 26.06 | 2.655 |
| --- | --- | --- | --- | --- |
| "VC" | 0.5 | 10 | 7.98 | 2.747 |
| "VC" | 1.0 | 10 | 16.77 | 2.515 |
| "VC" | 2.0 | 10 | 26.14 | 4.798 |

## Confidence Intervals & Hypothesis Testing

### Basic Assumptions

Having **each group has a limited number of sample (10 sample)** then **t distribution and t confidence intervals** are going to be used. Being the guinea pigs used for each measurement distinct from each other within a group and between the groups then the **groups can be considered independent from each other** and each guinea pig can be considered **iid**. Finally looking at the **sample standard deviation** for the different groups the assumption can be made to consider the **diffent variances as equals**.

### Independent group t confidence intervals

|  |  |  |
| --- | --- | --- |
| **Group1** | **Group2** | **t confidence interval (g1, g2)** |
| **OJ, 1.0** | **OJ, 0.5** | 5.5291857, 13.4108143 **(1)** |
| **OJ, 2.0** | **OJ, 1.0** | 0.2194983, 6.5005017 **(2)** |
| --- | --- | --- |
| **VC, 1.0** | **VC, 0.5** | 6.3156545, 11.2643455 **(3)** |
| **VC, 2.0** | **VC, 1.0** | 5.7710402, 12.9689598 **(4)** |
| --- | --- | --- |
| **OJ, all** | **VC, all** | -0.1670064, 7.5670064 **(5)** |
| --- | --- | --- |
| **OJ, 0.5** | **VC, 0.5** | 6.3156545, 11.2643455 **(6)** |
| **OJ, 1.0** | **VC, 1.0** | 14.4871567, 21.8328433 **(7)** |
| **OJ, 2.0** | **VC, 2.0** | 5.7710402, 12.9689598 **(8)** |

**Interpretation**

* **Supplement OJ only**, since the intervals (1)(2) are entirely above zero it suggest that groups with a higher dose have more tooth length than groups with a lower dose (at 95% confidence). The same cosideration can be done considering **supplement VC only** see (3)(4).
* Comparing the **supplement OJ vs. VC**
  + **overall** data, the confidence interval (5) contains zero (with most on the interval above zero) and makes it difficult to interpret the data
  + **data** for the same dose level (e.g 0.5 mg OJ vs. 0.5 mg VC), the intervals (6)(7)(8) are entirely above zero suggesting again that groups treated with OJ have more tooth length than groups treated with VC (at 95% confidence)

### Hypothesis Testing

Lets perform (two sided) hypothesis testing focusing on the comparison of the 2 different supplements OJ vs VC and considering the following hypothesis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Group1** | **Group2** | **t** | **df** | **p-value** |
| **OJ, all** | **VC, all** | 1.9152683 | 58 | 0.0603934 |
| --- | --- | --- | --- | --- |
| **OJ, 0.5** | **VC, 0.5** | 7.4634301 | 18 | 6.492264610^{-7} |
| **OJ, 1.0** | **VC, 1.0** | 10.3877952 | 18 | 4.957285710^{-9} |
| **OJ, 2.0** | **VC, 2.0** | 5.4698137 | 18 | 3.397577910^{-5} |

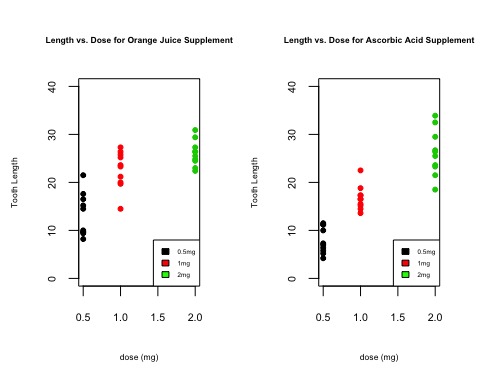
**Interpretation**  
Remember the **p-value** represents the probability of seeing evidence as extreme or more extreme than the one obtained under .

* Comparing the **supplement OJ vs. VC**
  + **overall** data, the \_\_t\_\_value is relevant so **"the hypothesis can be rejected"** with a **p-value** around 6%.
  + **data** for the same dose level (e.g 0.5 mg OJ vs. 0.5 mg VC), **t** values are quite big so it could be said that **"the hypothesis can be rejected"** and the **p-values** are very very low therefore a very unlikely event has been observed or, again, the **"the hypothesis is false"**.

## Appendix

**Code used to create the basic plot to visualize basic information of the rawData.**

require(graphics)  
rawData\_OJ <- rawData[rawData$supp == "OJ",]  
rawData\_OJ$doseF <- as.factor(rawData\_OJ$dose)  
rawData\_VC <- rawData[rawData$supp == "VC",]  
rawData\_VC$doseF <- as.factor(rawData\_VC$dose)  
  
par(ps = 8, cex.lab = 0.8, mfcol = c(1,2))  
plot(rawData\_OJ$dose, rawData\_OJ$len, col = c(rawData\_OJ$doseF), pch = 20, main ="Length vs. Dose for Orange Juice Supplement", xlab = "dose (mg)", ylab ="Tooth Length", ylim=c(0, 40), cex.main = 0.8)  
legend("bottomright",cex = 0.6, fill = c("black","red","green"), legend = c("0.5mg","1mg","2mg"))  
  
plot(rawData\_VC$dose, rawData\_VC$len, col = c(rawData\_VC$doseF), pch = 20, main ="Length vs. Dose for Ascorbic Acid Supplement", xlab = "dose (mg)", ylab ="Tooth Length", , ylim=c(0, 40), cex.main = 0.8)  
legend("bottomright",cex = 0.6, fill = c("black","red","green"), legend = c("0.5mg","1mg","2mg"))



**Code used to prepare the data of the different groups**

len\_OJ <- rawData[rawData$supp == "OJ",]$len  
len\_OJ\_05 <- rawData[rawData$supp == "OJ" & rawData$dose == 0.5,]$len  
len\_OJ\_10 <- rawData[rawData$supp == "OJ" & rawData$dose == 1.0,]$len  
len\_OJ\_20 <- rawData[rawData$supp == "OJ" & rawData$dose == 2.0,]$len  
  
len\_VC <- rawData[rawData$supp == "VC",]$len  
len\_VC\_05 <- rawData[rawData$supp == "VC" & rawData$dose == 0.5,]$len  
len\_VC\_10 <- rawData[rawData$supp == "VC" & rawData$dose == 1.0,]$len  
len\_VC\_20 <- rawData[rawData$supp == "VC" & rawData$dose == 2.0,]$len

**Code used to calculate basic statistics for the different groups**

getStatistics <- function(data){  
 return (list(n = length(data), mean = mean(data), sd = round(sd(data),digits = 3)))  
}  
  
##OJ Statistics  
stats\_OJ\_05 <- getStatistics(len\_OJ\_05)  
stats\_OJ\_10 <- getStatistics(len\_OJ\_10)  
stats\_OJ\_20 <- getStatistics(len\_OJ\_20)  
  
##VC Statistics  
stats\_VC\_05 <- getStatistics(len\_VC\_05)  
stats\_VC\_10 <- getStatistics(len\_VC\_10)  
stats\_VC\_20 <- getStatistics(len\_VC\_20)

**Code used to calculate t test data (confidence intervals and hypothesis testing) between the different groups**

t.testOJ\_10\_vs\_05 <- t.test(len\_OJ\_10, len\_OJ\_05, paired = FALSE, var.equal = TRUE)  
t.testOJ\_20\_vs\_05 <- t.test(len\_OJ\_20, len\_OJ\_05, paired = FALSE, var.equal = TRUE)  
t.testOJ\_20\_vs\_10 <- t.test(len\_OJ\_20, len\_OJ\_10, paired = FALSE, var.equal = TRUE)  
  
t.testVC\_10\_vs\_05 <- t.test(len\_VC\_10, len\_VC\_05, paired = FALSE, var.equal = TRUE)  
t.testVC\_20\_vs\_05 <-t.test(len\_VC\_20, len\_VC\_05, paired = FALSE, var.equal = TRUE)  
t.testVC\_20\_vs\_10 <-t.test(len\_VC\_20, len\_VC\_10, paired = FALSE, var.equal = TRUE)  
  
t.testOJ\_05\_vs\_VC\_05 <- t.test(len\_OJ\_05, len\_VC\_05, paired = FALSE, var.equal = TRUE)  
t.testOJ\_10\_vs\_VC\_10 <- t.test(len\_OJ\_10, len\_VC\_10, paired = FALSE, var.equal = TRUE)  
t.testOJ\_20\_vs\_VC\_20 <- t.test(len\_OJ\_20, len\_VC\_20, paired = FALSE, var.equal = TRUE)  
  
t.testOJ\_vs\_VC <- t.test(len\_OJ, len\_VC, paired = FALSE, var.equal = TRUE)

**Markdown used for the generation of the different tables**  
Note in the following markdown the **`** has been changed to **'** in the inline code, in order to be able to show the code itself.

Table of Sample statistics

Supp | Dose | No of Sample | Sample mean | Sample SD   
---- | ---- | ---- | ---- | ----  
"OJ" | 0.5 | 'r stats\_OJ\_05$n' | 'r stats\_OJ\_05$mean' | 'r stats\_OJ\_05$sd'   
...

Table of Confidence Intervals

\_\_Group1\_\_ | \_\_Group2\_\_ | \_\_t confidence interval (g1, g2)\_\_   
---- | ---- | ----  
\_\_OJ, 1.0\_\_ | \_\_OJ, 0.5\_\_ | 'r t.testOJ\_10\_vs\_05$conf' \_\_(1)\_\_   
...

Table of Hypothesis Testing

\_\_Group1\_\_ | \_\_Group2\_\_ | \_\_t\_\_ | \_\_df\_\_ | \_\_p-value\_\_  
---- | ---- | ---- | ---- | ----  
\_\_OJ, all\_\_ | \_\_VC, all\_\_ | 'r t.testOJ\_vs\_VC$statistic' | 'r t.testOJ\_vs\_VC$parameter' | 'r t.testOJ\_vs\_VC$p.value'