# Statistical Inference Project Part 2

Analyzing the Tooth Growth Data set from R

*Allan R Brewer Cappellin*

*May 19, 2015*

## Introduction

In this, the second part of the project, we will analyze the tooth growth data set from R. We will do a Exploratory Data Analysis of it, create a summary of the data, and do a hypothesis test to compare the means of two data sets in the data frame. This analysis will be done using the tools learned in the Statistical Inference Course from Coursera. And the main objective is to understand the data and compare if there is a statistical significant difference between the Supplement and Dose Tooth Growth Data.

---

## Data Analysis and Summary

In the first section we will create a plot to do an exploratory data analysis and a figure of the summary of the data. For this analysis we will load the packages ggplot2 and dplyr

Our first steps in the Exploratory Data Analysis (EDA) will be to load the data sets into the R environment and transform into a table data frame to manipulate with dplyr, this R code can be seen in the Appendix as "Load Data R Code". Before we do anything we must understand were the data comes from. The length measurement comes from the odontoblast cells from the incisors of a population of 60 guinea pigs. The different measurer are done by Supplement type divided in Orange Juice and pure Vitamin C ingestion and the doses vary from 0.5, 1 and 2 milligrams per day, the experiment was done for 42 days.

To do a proper EDA we must print a summary and structure of the data to understand the the variables involved and their values. The R code for this section is under the appendix "Summary R Code"

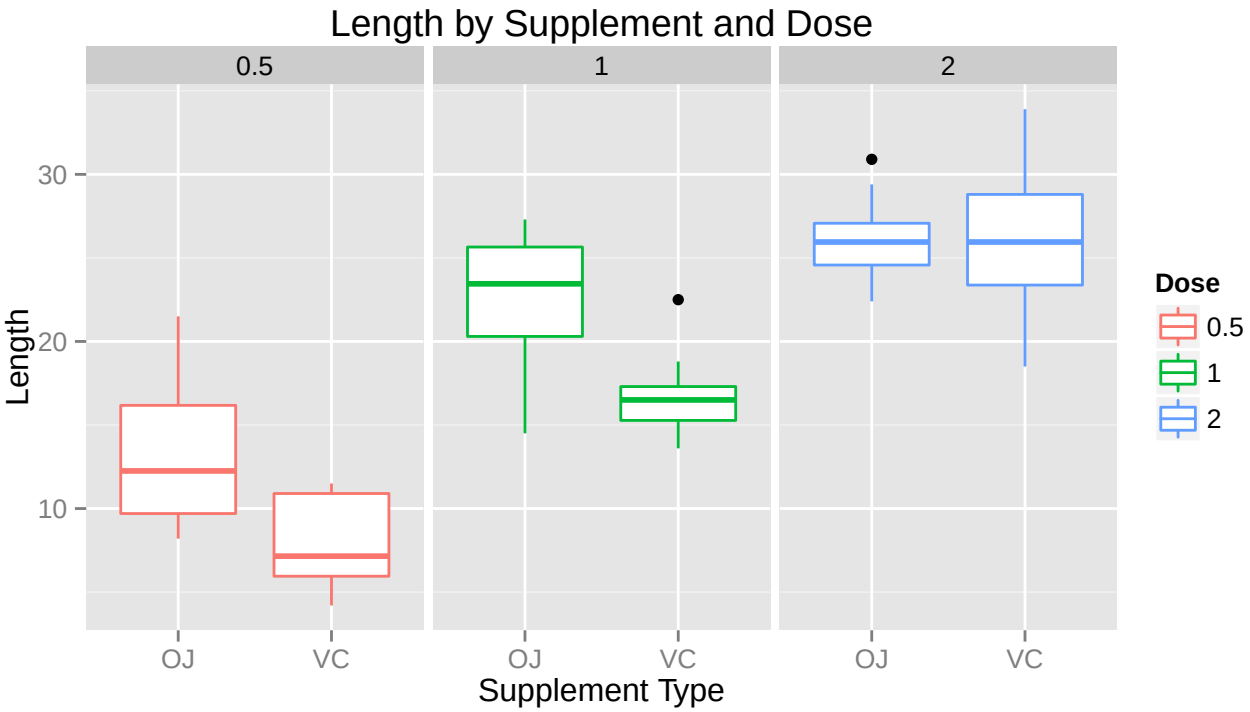
- Summary and Structure*

```
## 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 ...
```

```
##       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
```

From the information above we can see the data has 3 variables composed by length of the tooth in numbers, supplement type by two levels OJ and VC and dose amount in three levels 0.5, 1 and 2 milligrams per day. From the summary of the data we appreciate that there ate no missing values or NAs and that is well

organized as a tidy data set. The data is divided in 30 observations for Oj and 30 for VC and in this sections there are ten for each dose. After this summary analysis of the data we can create a plot for a proper EDA and better understand the behavior of the variables. The following figure presents a box plot faceted by Dose and its R code is in the Appendix under "Box Plot R Code"




From the Box Plot by Dose and Supplement we can make some assumptions about the data that we will test in the following section using a t test. First we can appreciate how the Length seems to get larger with the increase in dosage and that it appears for the 0.5 an 1 milligram that the Orange Juice has a better result in Teeth Growth than the pure Vitamin C. This Assumptions will be tested next.

## Hypothesis Test

For this section we will perform various Hypothesis test to understand if the difference in length between Supplement Type and Dose is statistically significant. We will perform 5 different Hypothesis Test using the T Test. The Test will be done with the following characterisitcs. Assume unequal variance, unpaired data and we will do a two sided test for two different variables to determine if the means difference is statistically significant. The R code for the test can be found in the appendix under "Hypothesis Test R Code"

## Conclusions

For this test we made some assumptions, first there are 60 guinea pigs that has no connection between one and other. The t test were done as a 2 sample test that have unequal variance, since when one has no idea if variance is equal or not it is recommended to do it as unequal. The next assumption is that the tests are done as 2 tailed since the hypothesis is in every case to determine if the means are equal or not.

## Appendix

In this section we will show all the R code, figures and graphs that support the document and conclusions drawn in it.

### R code

In this first section of the appendix you can find all of the R code that is not shown in the main body of the document.

*R Code #1 : Load Data R Code*

```
data(ToothGrowth)
x <- tbl_df(ToothGrowth); rm(ToothGrowth)
```

*R Code #2 : Summary R Code*

```
str(x)
summary(x)
```

*R Code #3 : Box Plot R Code*

```
plot1 <- ggplot(x, aes(x = supp, y = len, colour = factor(dose)))
plot1 <- plot1 + labs(title = "Length by Supplement and Dose", x = "Supplement Type", y = "Length")
plot1 <- plot1 + geom_boxplot() + scale_color_discrete(name = "Dose") + facet_grid(~dose)
plot1
```

*R Code #4 : Hypothesis Test R Code*

```
oj05 <- filter(x, supp == "OJ", dose == 0.5)
oj10 <- filter(x, supp == "OJ", dose == 1)
oj20 <- filter(x, supp == "OJ", dose == 2)
vc05 <- filter(x, supp == "VC", dose == 0.5)
vc10 <- filter(x, supp == "VC", dose == 1)
vc20 <- filter(x, supp == "VC", dose == 2)
d05 <- filter(x, dose == 0.5)
d10 <- filter(x, dose == 1)
d20 <- filter(x, dose == 2)
hypo1 <- t.test(oj05$len, vc05$len, paired=FALSE, var.equal=FALSE, alternative="two.sided")
hypo2 <- t.test(oj10$len, vc10$len, paired=FALSE, var.equal=FALSE, alternative="two.sided")
hypo3 <- t.test(oj20$len, vc20$len, paired=FALSE, var.equal=FALSE, alternative="two.sided")
hypo4 <- t.test(d05$len, d10$len, paired=FALSE, var.equal=FALSE, alternative="two.sided")
hypo5 <- t.test(d10$len, d20$len, paired=FALSE, var.equal=FALSE, alternative="two.sided")
```

### Results

*Result for Hypothesis Test 1 : Compare OJ vs VC for 0.5 mg/d*

```
##
##  Welch Two Sample t-test
##
## data:  oj05$len and vc05$len
## 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 of x mean of y
##     13.23      7.98
```

*Result for Hypothesis Test 2 : Compare OJ vs VC for 1.0 mg/d*

```
##
##  Welch Two Sample t-test
##
## data:  oj10$len and vc10$len
## 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 of x mean of y
##     22.70     16.77
```

*Result for Hypothesis Test 3 : Compare OJ vs VC for 3.0 mg/d*

```
##
##  Welch Two Sample t-test
##
## data:  oj20$len and vc20$len
## 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 of x mean of y
##     26.06     26.14
```

*Result for Hypothesis Test 4 : Compare 0.5 mg/d vs 1.0 mg/d for all Supp Types*

```
##
##  Welch Two Sample t-test
##
## data:  d05$len and d10$len
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

*Result for Hypothesis Test 5 : Compare 1.0 mg/d vs 2.0 mg/d for all Supp Types*

```
##
##  Welch Two Sample t-test
##
## data:  d10$len and d20$len
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

---