JHU Data Science Spcecialization

Statistical Inference Course Project Part 2

The Effect of Vitamin C on Tooth Growth in Guinea Pigs

Author: Rafael Marino

Date: April 25st, 2015

Overview

In this project we will analyze the ToothGrowth data from the datasets package in R. The first step is to load the data in R, to explain what the variables are and in which units they are expressed. Then, explore it using functions such as *summary()* and *str()*; as well as exploratory graphs using the *ggplot2* package. Based upon suggestions provided by the exploratory analysis, assumptions will be made and tested in a formal confidence interval and/or hypothesis testing framework leading to the final conclusions.

Data Pre-processing and package loading

```r
library(ggplot2)
library(datasets) #loads the datasets package
data(ToothGrowth) #loads the ToothGrowth dataset
data <- ToothGrowth
colnames(data) <- c("Length", "Supplement", "Dose") #Making clearer names
```

Variable Explanation

According to R's own documentation, the ToothGrowth dataset contains an experiment on "The Effect of Vitamin C on Tooth Growth in Guinea Pigs". "The response variable 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)."

Therefore:

1.**Length**= length of the teeth.

2.**Supplement**= supplement, delivery method, orange juice (OJ) or ascorbic acid (VC).

3.**Dose**= dosage, measured in mg.

Exploratory Data Analysis

From this initial exploration of the data, we can see that it's a data frame containing 3 features, with 60 observations each. The **Length** variable is numerical and takes on many values, the **Supplement** variable is categorical, with 2 levels while the **Dose** variable is numerical and can only take on 3 values.

Analyzing teeth length by supplement taken only (*ignoring the differences in dosage*), it would appear that Orange Juice produces a larger mean tooth length (20.66) vs vitamin C (16.96). It is easier to notice this by looking at the boxplots in **Figure 1**, on the Appendix section. The variability also appears to be lower on the OJ sample vs the VC sample (43.63 vs 68.32, respectively).

Now, isolating the effect of Dose on tooth Length (*ignoring the method of delivery*), the results appear consistent, showing higher *mean* tooth length as Dose increases. The mean tooth length for all 3 levels of studied dosage (0.5, 1 and 2) are: 10.60, 19.73, 26.1 Please check **Figure 2** to see the comparison.

Finally, in order to take into consideration how both variables affect tooth length (please refer to **Figure 3**), a panel figure was constructed, comparing the effect of each supplement at each dosage level. From Figure 3, it would appear the OJ and VC cause quite different mean tooth length at the lower dosages (0.5 and 1 mg), while very similar mean length at the higher dosage of 2 mg (though at this dosage, OJ shows much smaller variability).

Hypothesis Testing

Statistical assumptions of the test (important):

This hypothesis test is the case of a 2 group comparison. As seen in class the t-student distribution is quite appropriate for this case, so it will be the one used. T tests can be carried out in R using the *t.test()* function. The parameter "paired" will be set to FALSE, because on each sample (of the same dose), the guinea pigs whose teeth are measured are different, and one does not have anything to do with the other. The "var.equal" parameter will be set to FALSE, because of two things: 1. Prof. Caffo mentioned in class "when in doubt always go with the unequal" and 2. because if we use the sample variance as estimate of the population, it is in fact quite different between samples (as can be seen in Figure3).

All of the tests carried out will be of the form:

Null Hypothesis: $\mu_1 = \mu_2$ or equivalently μ_1 - μ_2 = 0 #means are equal

Alternative Hypothesis: μ_1 - μ_2 != 0

Test statistic: μ_1 - μ_2 - 0 / (sqrt (S_1^2/n1)+ (S_2^2/n2)) ; t~ (n1+n2-2)

Which Hypotheses will be tested?

After an initial exploration of the data, some hypothesis will be discarded, for instance, comparing the mean tooth length in the OJ sample with the mean tooth length in the VC sample, wouldn't make much sense *without* discriminating between different dosages. Because as seen in Figure 2, the Dosage variable is very significant, for both methods of delivery.

Therefore 3 hypotheses were selected to be tested (a the 95% confidence level):

1. The mean tooth length caused by Orange Juice is equal to that caused by VC, at the 0.5 mg level.
2. The mean tooth length caused by Orange Juice is equal to that caused by VC, at the 1 mg level.
3. The mean tooth length caused by Orange Juice is equal to that caused by VC, at the 2 mg level.

NOTE: Keep in mind that in order to run these 3 tests, 6 samples are required. For instance, for the first hypothesis test, the two required samples would be: 1. the sample of the guinea pigs that were supplemented with OJ at 0.5 mg and the sample of guinea pigs that were supplemented with VC at 0.5 mg. Analogous subsetting has to be done for the 2 other tests. This subsetting process is done in the **Appendix**, in a section titled "Subsetting the Samples".

Conclusions from Hypothesis Testing.

NOTE: The test for each hypothesis can be seen in the Appendix, **Hypothesis Testing** section.

1. The mean tooth length caused by Orange Juice is equal to that caused by VC, at the 0.5 mg level.

The p-value for this hypothesis is 0.006, which means there's a 0.06% probability of observing such exteme values under the null. We therefore reject the null hypothesis.

2. The mean tooth length caused by Orange Juice is equal to that caused by VC, at the 1 mg level.

Similarly, the p-value for the equality of means at the 1mg level is low, at 0.01% probability of observing such extreme values under the null. Which leads as to reject the null of equal means in favor of the alternative hypothesis.

3. The mean tooth length caused by Orange Juice is equal to that caused by VC, at the 2 mg level.

Unlike the 2 previous hypotheses, the p-value here is quite high, with a 96.4% probability of observing such extreme values, under the null hypothesis, which leads us not to reject, or "accept" the null of equal means.

To conclude, dosage levels are statistically significant and positively correlated with tooth length, regardless of the delivery method. At lower dosages (0.5mg and 1mg) the delivery method causes significant differences in tooth growth, with OJ generating far higher tooth length than VC. However, at a higher dosage (2mg) the delivery method becomes much less important, and the mean length from OJ and VC are equal at the 95% confidence.

Appendix

Exploratory Data Analysis

```r
str(data)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ Length    : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ Supplement: 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 ...
```

```r
summary(data)
```

```
##      Length      Supplement      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
```

Calculating mean and variance of tooth length of the OJ sample and the VC sample.

Mean

```
## [1] 20.66333 16.96333
```

Variance

```
## [1] 43.63344 68.32723
```

Calculating the mean tooth length for each level of dosage:

```
##       0.5mg    1mg  2mg
## [1,] 10.605 19.735 26.1
```

Graphs

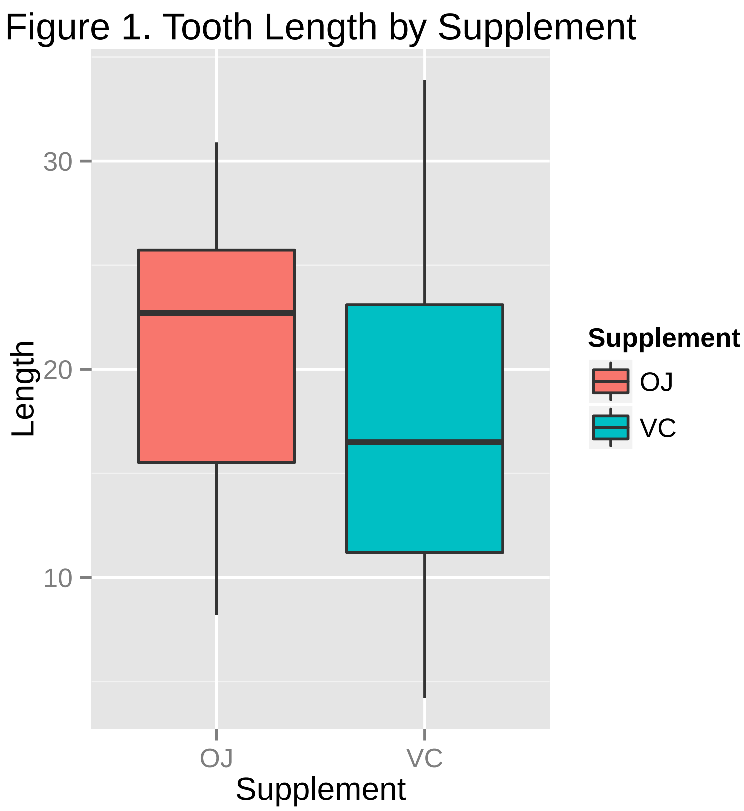


Figure 1. Tooth Length by Supplement
30
20
10
Length
OJ
VC
Supplement
Supplement
OJ
VC

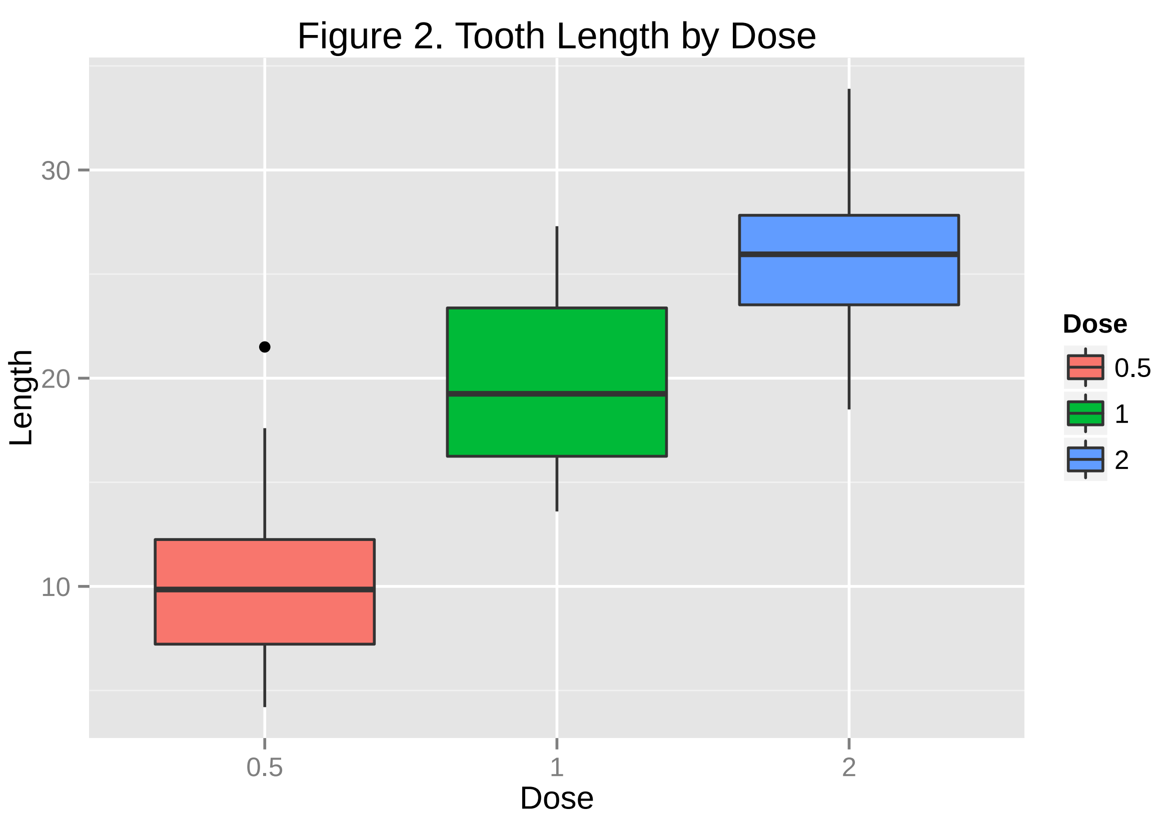
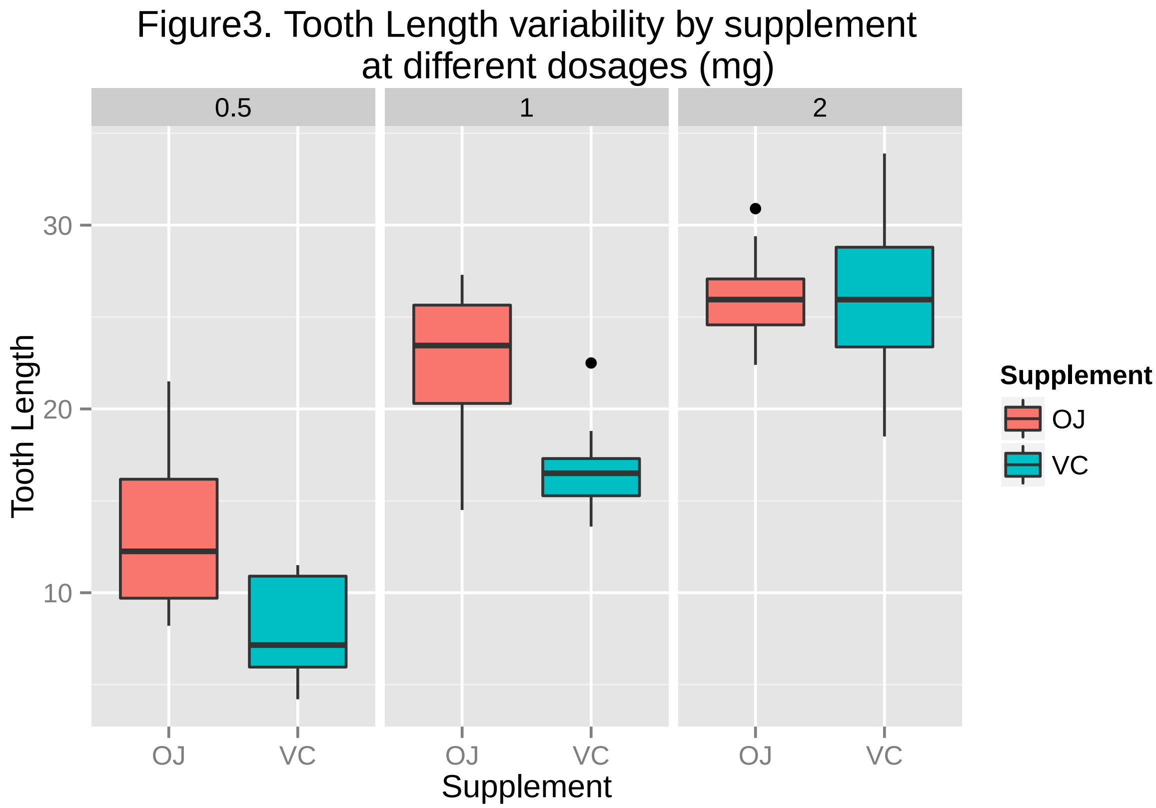


Figure 2. Tooth Length by Dose
30
20
10
Length
0.5
1
2
Dose
Dose
0.5
1
2



Subsetting the Samples

```
ojpoint5 <- data[data$Dose==0.5 & data$Supplement=="OJ",1]
vcpoint5 <- data[data$Dose==0.5 & data$Supplement=="VC",1]
oj1 <- data[data$Dose==1 & data$Supplement=="OJ",1]
vc1 <- data[data$Dose==1 & data$Supplement=="VC",1]
oj2 <- data[data$Dose==2 & data$Supplement=="OJ",1]
vc2 <- data[data$Dose==2 & data$Supplement=="VC",1]
```

Hypothesis testing.

1. The mean tooth length caused by Orange Juice is equal to that that caused by VC, at the 0.5 mg level.

```
##
##  Welch Two Sample t-test
##
## data:  ojpoint5 and vcpoint5
## 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
```

2. The mean tooth length caused by Orange Juice is equal to that caused by VC, at the 1 mg level.

```
##
##  Welch Two Sample t-test
##
## data:  oj1 and vc1
## 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
```

3. The mean tooth length caused by Orange Juice is equal to that caused by VC, at the 2 mg level.

```
##
##  Welch Two Sample t-test
##
## data:  oj2 and vc2
## t = -0.0461, 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
```