StatisticalInference

ArjunVenkat

October 25, 2015

Statistical Inference Course Project

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.2.2
```

## Warning: package 'ggplot2' was built under R version 3.1.1 This report deals with analysis of the ToothGrowth data in the R data sets package. The data is set of 60 observations, 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).

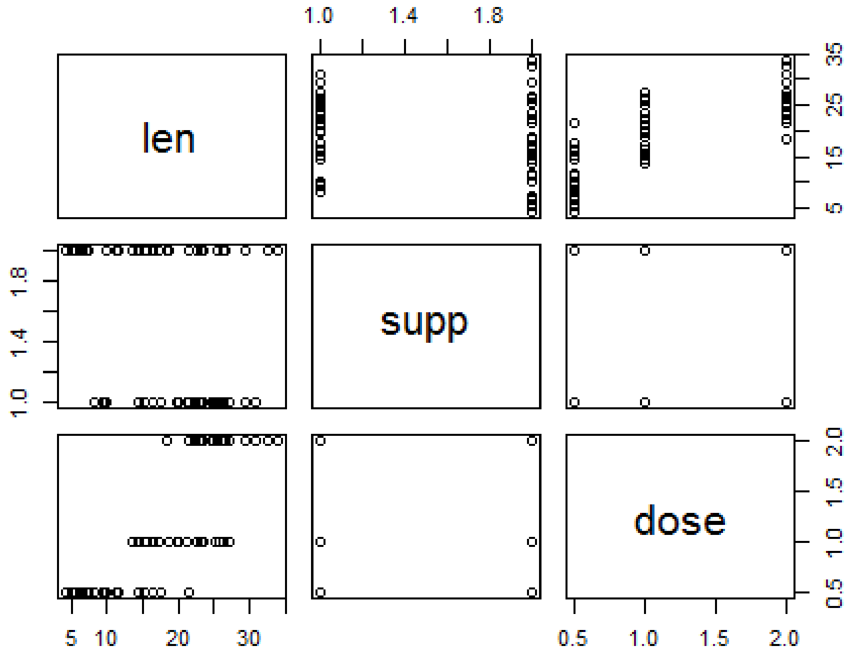
Load and visualize the data

Load the data

```
data(ToothGrowth)
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 ...

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
```

Summary of the data

```
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  

plot(ToothGrowth)
```



The structure of the data set suggests the dose variable is numeric, but in fact, we have seen that actually there are only three dose levels of Vitamin C (0.5, 1 and 2 mg). A more clear view of this, is the third row (or the third column, they are equivalent) of the scatter plot matrix, provided above. We will transform the variable dose to a factor variable

```
ToothGrowth$dose<-as.factor(ToothGrowth$dose)
summary(ToothGrowth$dose)

## 0.5   1   2 
##  20  20  20
```

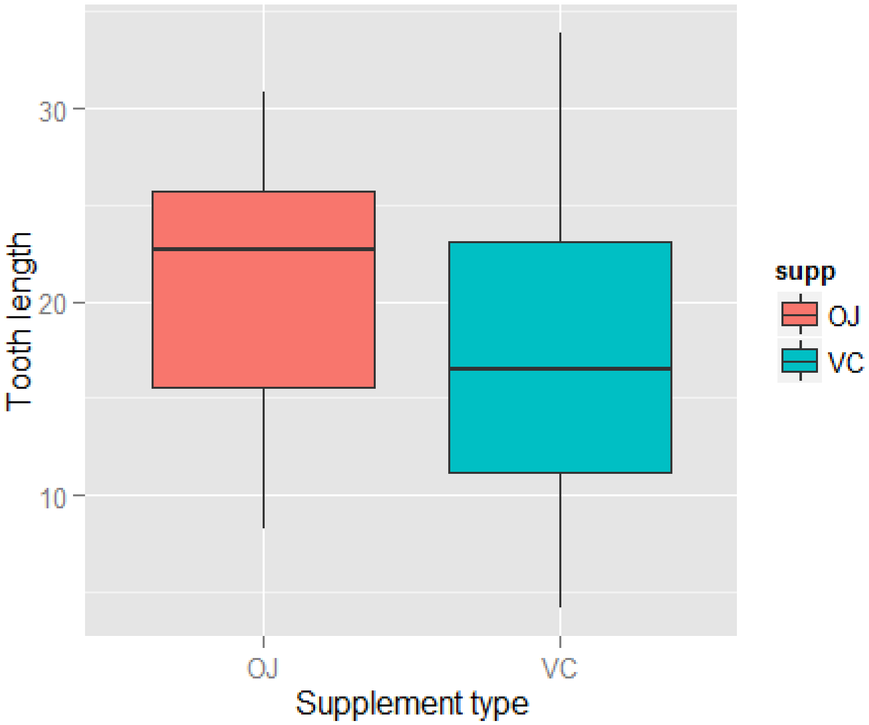
Exploratory Data analysis

So, the mean of the len variable is 18.8133, but this refers to all the 60observations of the data set. The mean for the two levels of the supply method (supp) are described as below.

```
meansupp = split(ToothGrowth$len, ToothGrowth$supp)
sapply(meansupp, mean)

##       OJ       VC 
## 20.66333 16.96333
```

Graph

```
ggplot(aes(x=supp, y=len), data=ToothGrowth) + geom_boxplot(aes(fill=supp))+
xlab("Supplement type") +ylab("Tooth length")
```

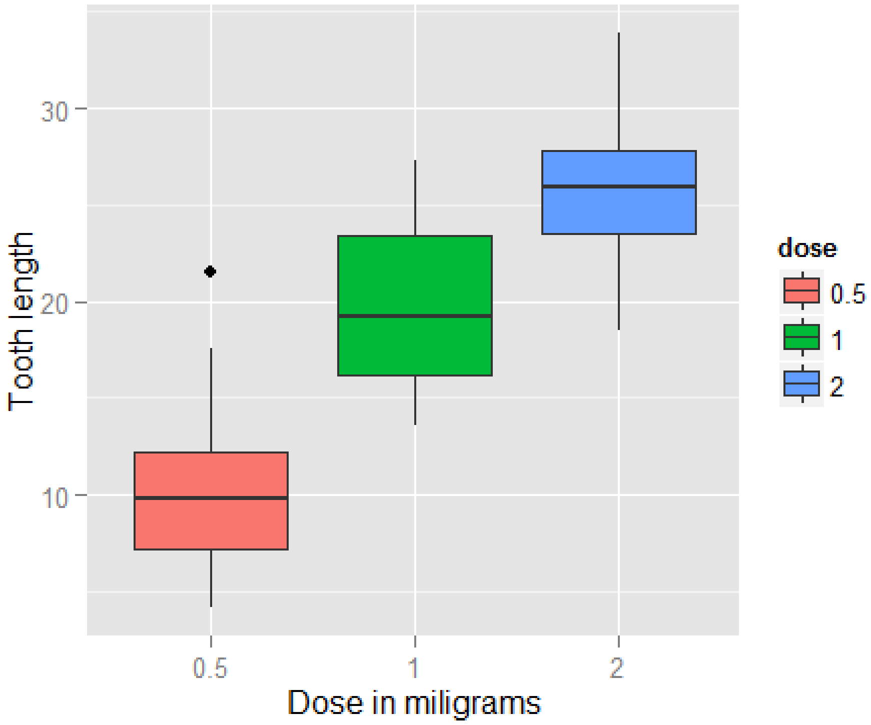


We will now check the effect of vitamin C dose on tooth length. So first we will calculate the means of the three levels of dosage.

```
meandose = split(ToothGrowth$len, ToothGrowth$dose)
sapply(meandose, mean)

##    0.5      1      2
## 10.605 19.735 26.100
```

Graphically

```
ggplot(aes(x=dose, y=len), data=ToothGrowth) + geom_boxplot(aes(fill=dose)) +
xlab("Dose in miligrams") +ylab("Tooth length")
```



##Inferential Statistics Do the tooth length of the guinea pigs depends on delivery methods? A t test for the difference will be made to test this claim

```
len<-ToothGrowth$len
supp<-ToothGrowth$supp
dose<-ToothGrowth$dose
sapply(meansupp, var)
```

```
##       OJ       VC 
## 43.63344 68.32723
```

```
t.test(len[supp=="OJ"], len[supp=="VC"], paired = FALSE, var.equal = FALSE)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  len[supp == "OJ"] and len[supp == "VC"]
## 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 of x mean of y 
##  20.66333  16.96333
```

The p-value of this test was 0.06, which is very close to the significance level of 5%. It could be interpreted as a lack of enough evidence to reject the null hypothesis, however it is paramount to account that the 0.05 value of significance is only a convenience value.

Furthermore, the confidence interval of the test contains zero (0) Now we will test the tooth length of the group with vitamin C dosage

```
t.test(len[dose==2], len[dose==1], paired = FALSE, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  len[dose == 2] and len[dose == 1]
## t = 4.9005, df = 38, p-value = 1.811e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.735613 8.994387
## sample estimates:
## mean of x mean of y
##    26.100    19.735
```

State conclusions

The dataset as presented in R has little supporting documentation, but we have assumed that the data is indeed paired.From the exploratory data analysis, we see that increased vitamin C dosages (in either orange juice or pure ascorbic acid form) is an effective promoter of tooth growth.From the T-test analysis above, we conclude from the statistically significant p-values that for dosages of 0.5 mg and 1 mg, orange juice is more effective at promoting tooth growth than just ascorbic acid. From the p-value for the 2 mg, we cannot conclude that orange juice promotes tooth growth more effectively than just ascorbic acid.