Analysis of the ToothGrowth data in R

Pedro Magalhães Bernardo

Saturday, March 26, 2016

Overview

This report is part of a course project within the Statistical Inference course on the Data Science Specialization by Johns Hopkins University on Coursera.

On this report we will analyze the ToothGrowth data in the R datasets package

Exploratory Analysis

The ToothGrowth dataset shows the length of odontoblasts in 60 guinea pigs. Each animal received one of three dose levels of vitamin C by one of two delivery methods.

First let's load the dataset and take a look at some information it contains.

```r
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 ...
```

```r
head(ToothGrowth)
```

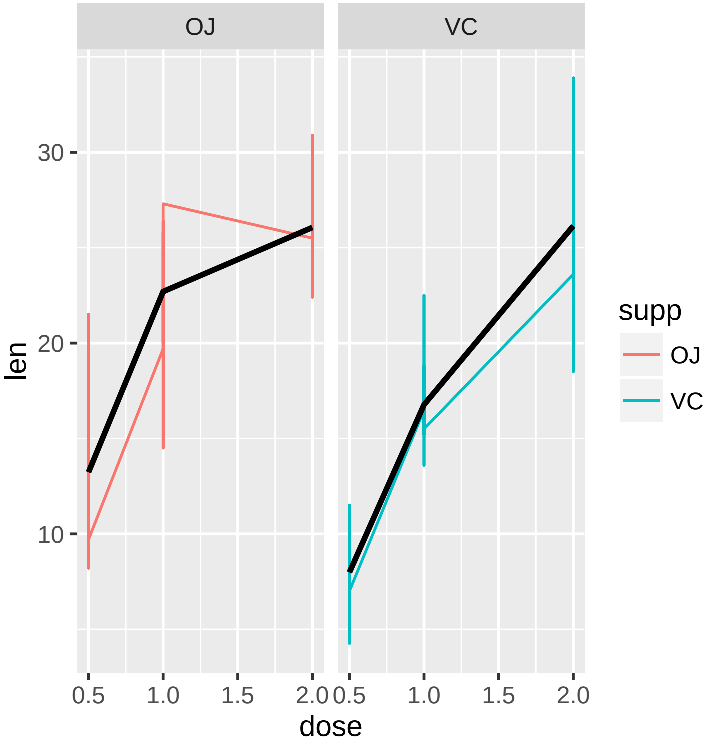
```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```r
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  
```

Let's take a look at how the length of odontoblasts varies with dosage for each delivery method.

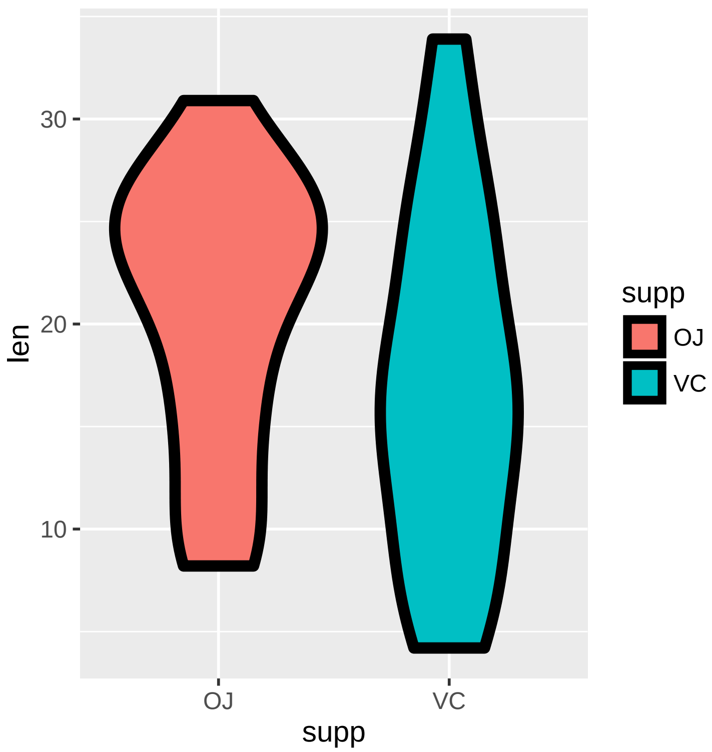
```
library(ggplot2)
raw_data <- ggplot(ToothGrowth, aes(x=dose,y=len,colour=supp)) +
      geom_line() +
      stat_summary(aes(group = 1), geom = "line", fun.y = mean, size = 1, col = "black") +
      facet_grid(. ~ supp)
raw_data
```



We can see that as the dosage increases, the mean of length of odontoblasts (black line) also increases, for both methods. Also, we can see, that for lower dosages the mean of the length of odontoblasts for the OJ method is bigger.

Now let's take a look at the length of odontoblasts by delivery method.

```
violin <- ggplot(ToothGrowth, aes(x=supp,y=len,fill=supp)) +
      geom_violin(col="black",size=2)
violin
```



From this violin chart we see that the VC delivery method has a bigger variance compared to the OJ method. Also the mean of length of odontoblasts of the pigs that received the vitamin through the OJ method is bigger.

Comparing tooth growth.

Now we will compare the two delivery methods to see if one is significantly better than the other when we look at the length of the odontoblasts.

For that we will make three different comparisons, one for each dosage (0.5, 1.0, 2.0)

First let's subset our datatset.

```
oj1 <- subset(ToothGrowth, supp=='OJ' & dose==1.0)$len
oj2 <- subset(ToothGrowth, supp=='OJ' & dose==2.0)$len
oj5 <- subset(ToothGrowth, supp=='OJ' & dose==0.5)$len
vc1 <- subset(ToothGrowth, supp=='VC' & dose==1.0)$len
vc2 <- subset(ToothGrowth, supp=='VC' & dose==2.0)$len
vc5 <- subset(ToothGrowth, supp=='VC' & dose==0.5)$len
```

Now let's run a **t test** for each pair and discuss the results.

```
t.test(oj5,vc5)
```

```
##
##  Welch Two Sample t-test
##
## data:  oj5 and vc5
## 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
```

```
t.test(oj1,vc1)
```

```
##
##  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
```

```
t.test(oj2,vc2)
```

```
##
##  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
```

We do not consider the data as paired, since the individuals (pigs), are not the same for each delivery method.

We can see that for the dosages of 0.5 and 1, the OJ delivery methods produces a higher mean for the length of odontoblasts. For both cases our confidence interval of 95% is quite far from 0, and our p-value is quite small (smaller than 0.05), that means that we reject the null hypothesis. This shows us that the OJ method is better than the VC method for this dosages.

On the other hand, for the dosage of 2.0, our confidence interval contains 0, and our p-value is quite large (almost 1). Therefore we fail to reject the null hypothesis, and we can not say if one method is better than other, in fact, they behave very similar for this dosage.