ToothGrowth data analysis

Javier Angoy

Jun 4th, 2017

Synopsis

This document contains the second part of the assignment for the Statistical Inference Course of Coursera. The objective is to analyze the ToothGrowth dataset and try to answer the question "Do delivery method and/or Dosage affect tooth growth in guinea pigs?". It consists of an analysis of the ToothGrowth dataset. ## Exploratory Data Analysis From the R datasets package we obtained the ToothGrowth dataset, entitled as "The Effect of Vitamin C on Tooth Growth in Guinea Pigs". It contains data on the response in the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice or ascorbic acid (a form of vitamin C and coded as VC). More info on this dataset can be found via ?ToothGrowth and also on https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ToothGrowth.html.

First we must load data. Later we explore the data

```
data(ToothGrowth)           #Load data
str(ToothGrowth)            #Structure of the data
```

```
## '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)           #Explore first rows of the data frame
```

```
##    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
```

```
unique(ToothGrowth$dose)    #Three different doses to administer
```

```
## [1] 0.5 1.0 2.0
```

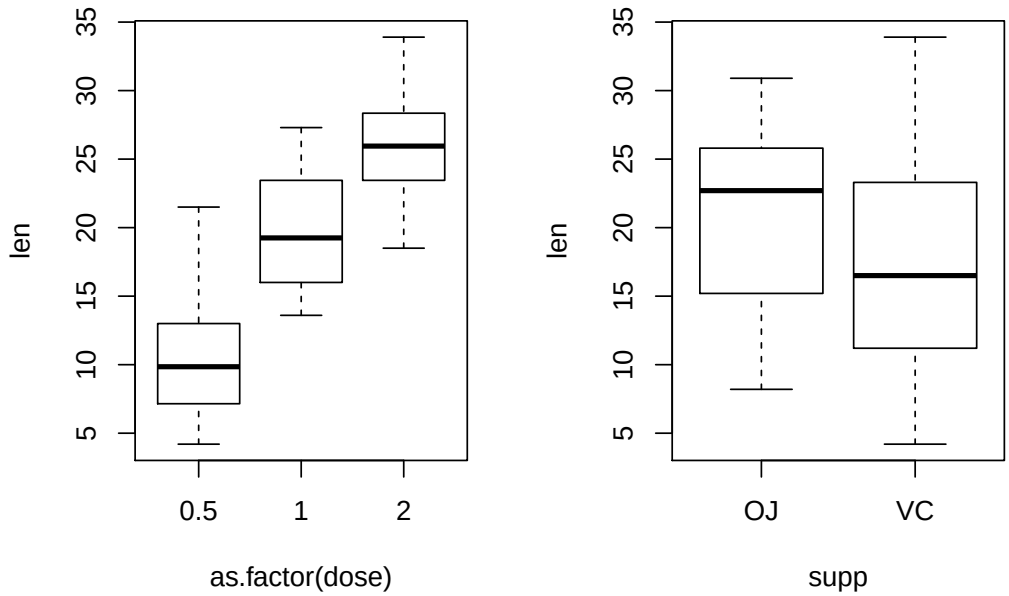
```
levels(ToothGrowth$supp)    #Two different ways of dose administering
```

```
## [1] "OJ" "VC"
```

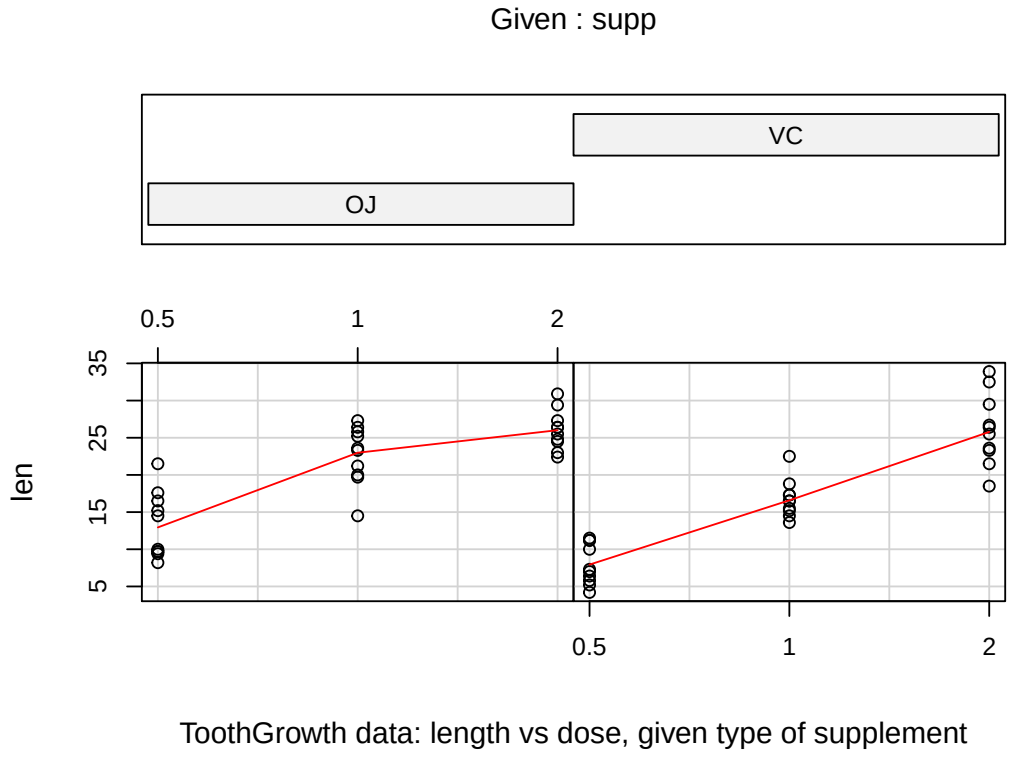
So we observe that the ToothGrowth dataset has 3 columns. The first column "len" contains the response in the length of odontoblasts in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) according to "dose" with each of two delivery methods (orange juice or ascorbic acid, OJ or VC respectively) stored on "supp".

Now we plot the data and continue to analyze.

```
par(mfrow=c(1,2))
with(ToothGrowth, plot(len ~ as.factor(dose) + supp))
```




```
coplot(len ~ as.factor(dose) | supp, data = ToothGrowth, panel = panel.smooth,
       xlab = "ToothGrowth data: length vs dose, given type of supplement")
```



Key Assumptions

For this analysis, we supplose that the sample of guinea pigs are representative of the universe of guinea pigs and the 60 animals were randomly assigned to the given doses and delivery methods. The observations are statistically independent. The two populations being compared have the same variance.

We have to keep in mind that we want to answer if delivery method and/or dosage affect tooth growth in

guinea pigs. Watching the previous plot we observe a clear positive relation between dose and lenght, that is to say that the bigger the dose is, the larger the lenght for both delivery methods. Our null hypothesis is

$$H_0 : \mu_1 = \mu_2$$

being μ_1, μ_2 the average growth for each sample group.

Hypothesis Tests

To determine the relation between dose and length, we found that there are three dose levels to analyze. This leads us to perform the study in pairs of values, so we have three different tests: For pairs (0.5 - 1), (1 - 2), and (0.5 - 2).

First pair (0.5 - 1)

```
t.test(ToothGrowth[which(ToothGrowth$dose==0.5),]$len,
       ToothGrowth[which(ToothGrowth$dose==1),]$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth[which(ToothGrowth$dose == 0.5), ]$len and ToothGrowth[which(ToothGrowth$dose == 1
## 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
```

The 95% confidence interval doesn't contain 0 and p-value is below 0.5 so reject the null hypothesis.

Second pair (1 - 2)

```
t.test(ToothGrowth[which(ToothGrowth$dose==1),]$len,
       ToothGrowth[which(ToothGrowth$dose==2),]$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth[which(ToothGrowth$dose == 1), ]$len and ToothGrowth[which(ToothGrowth$dose == 2),
## 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
```

The 95% confidence interval doesn't contain 0 and p-value is below 0.5 so reject the null hypothesis.

Third pair (0.5 - 2)

```
t.test(ToothGrowth[which(ToothGrowth$dose==0.5),]$len,
       ToothGrowth[which(ToothGrowth$dose==2),]$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth[which(ToothGrowth$dose == 0.5), ]$len and ToothGrowth[which(ToothGrowth$dose == 2
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean of x mean of y
##    10.605    26.100
```

The 95% confidence interval doesn't contain 0 and p-value is below 0.5 so reject the null hypothesis.

Conclusions

Having rejected the null hypothesis for the three pairs of values, it can be affirmed that there is difference between the means of the groups of values (0.5, 1 and 2). Thus we can conclude that there is a relationship between the supply of vitamin C and the variation in odontoblasts length.