Statistical Inference - Project (Part II)

Benny96

August 12th 2016

Basic Inferential Data Analysis.

Statement:

Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package.

1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.
3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)
4. State your conclusions and the assumptions needed for your conclusions.

1. Basic exploratory data analyses:

First, as we did before, let's load the required libraries and the "ToothGrowth" data:

```
library(ggplot2)
data(ToothGrowth)
```

Let's have a quick look at the data. First, let's look the first and last 6 observations:

```
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
```

```
tail(ToothGrowth)
```

```
##     len supp dose
## 55 24.8   OJ    2
## 56 30.9   OJ    2
## 57 26.4   OJ    2
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

And now, let's see how many supplements and dose amounts appear:

```r
table(ToothGrowth$supp)
```

```
##
## OJ VC
## 30 30
```

```r
table(ToothGrowth$dose)
```

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

2. Summary of the data:

Let's print the summary of the data contained in ToothGrowth:

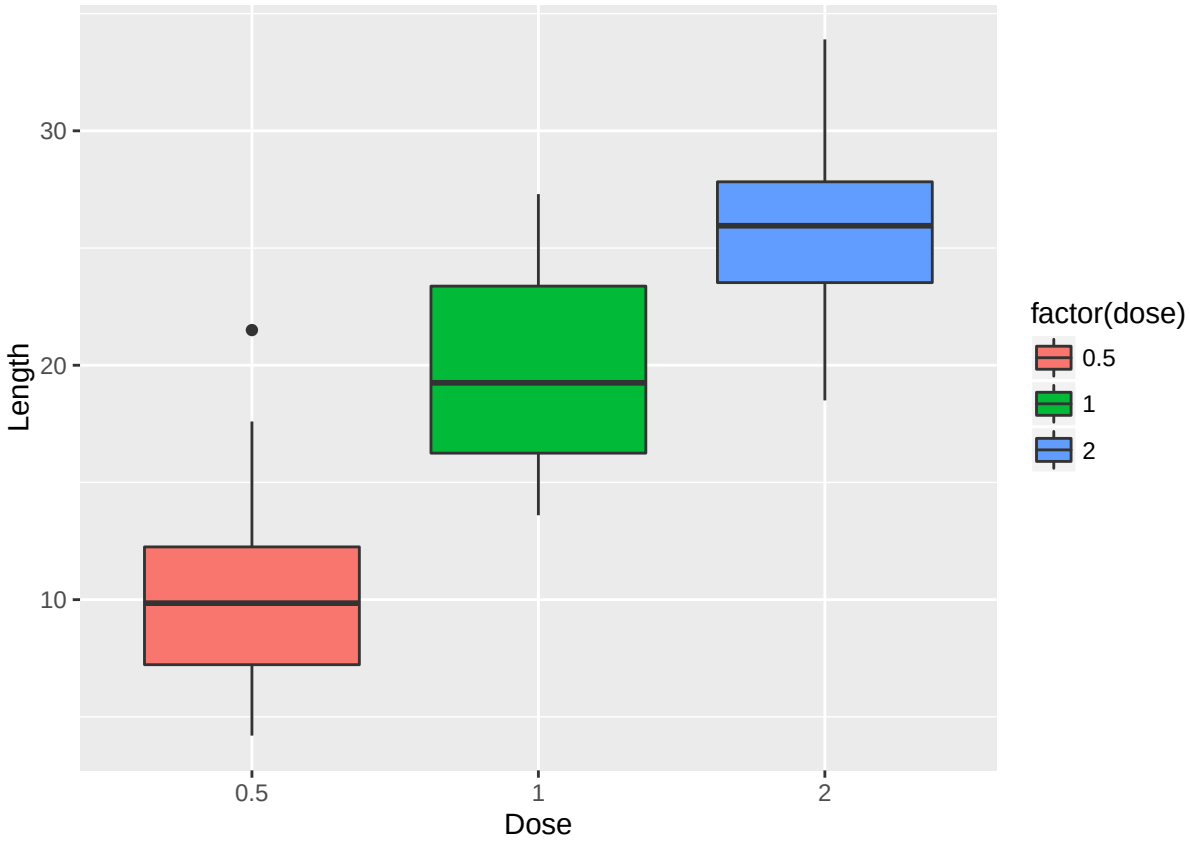
```r
summary(ToothGrowth)
```

```
##       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
```

We can see 3 variables here:

- len: Tooth length.
- supp: The supplement types given. (OJ and VC).
- dose: The dose (quantity) of the aforementioned supplements (mg/day).

To make the range of our data clear, let's plot the given information in various boxes, considering as a distinguishing factor the dose given to the test subjects.



So, we can see at a glance that the given dose and the tooth growth are correlated, as the higher the value of the dose gets, the bigger the growth of the teeth.

3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp or dose.

The hypothesis test we are doing is:

- H0: There is no relation between tooth growth and supplements or doses.
- Ha: There is a factual relation between tooth growth and supplements or doses.

First, we'll consider a confidence interval of 95% looking to the tooth growth in the data. We'll understand the distribution as normal.

```
tglength <- ToothGrowth$len
interval <- (mean(tglength) + c(-1, 1) * qnorm(0.975) * sd(tglength)/sqrt(length(tglength)))
interval
```

```
## [1] 16.87783 20.74884
```

Now, we'll do a t.test (which is applicable as Student's T distribution tends to the Normal distribution in a big amount of observations).

```
t.test(len~supp, data=ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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 in group OJ mean in group VC
##         20.66333         16.96333
```

As we see, the means obtained in the OJ group and the VC group are inside the interval, so we fail to reject the Null Hypothesis.

Now, we will subset the observations, classifying them by the dose (0.5|1|2):

```
dose0 <- subset(ToothGrowth, dose == 0.5)
dose1 <- subset(ToothGrowth, dose == 1)
dose2 <- subset(ToothGrowth, dose == 2)
```

And now, we'll perform a t.test on each of the subsets:

- Dose: 0.5. As the p-value is small enough, we can reject the null hypothesis.

```
t.test(len~supp, data=dose0)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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 in group OJ mean in group VC
##            13.23             7.98
```

- Dose: 1. As the p-value is small enough, we can also reject the null hypothesis in this case.

```
t.test(len~supp, data=dose1)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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 in group OJ mean in group VC
##            22.70            16.77
```

- Dose: 2. Now, the p-value is quite high, and it also goes out of the bounds of the interval, so we can also reject the null hypothesis.

```
t.test(len~supp, data=dose2)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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 in group OJ mean in group VC
##            26.06            26.14
```

4. Conclusions:

According to the obtained data from the t.tests, we can state that:

- The Tooth Length Growth is quite related to the Supplement.
- Orange Juice has notably a higher effect in small doses (dose = 0.5|1; due to a higher group mean), than the Vitamin C application.
- However, the case of "big doses (dose = 2)" isn't so clear; it seems that both supplements have a similar effect in that quantity.