Statistical Inference course project part 2: Basic Inferential Data Analysis

Mahmoud Nabegh

December 15, 2018

Contents

Overview

We use the ToothGrowth data to look at the tooth growth for 10 subjects when taking Orange Juice and Vitamin C with different doses.

Load the ToothGrowth data

```
library(datasets)
```

Exploratory analysis

Data structure

First step is to take a quick look at the data and its structure

```
dim(ToothGrowth)
```

```
## [1] 60  3
```

```
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
```

```
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 ...
```

Data properties

Next we look at some of the data properties such as: 1. Missing data

```
dim(ToothGrowth)[1]
```

```
## [1] 60
```

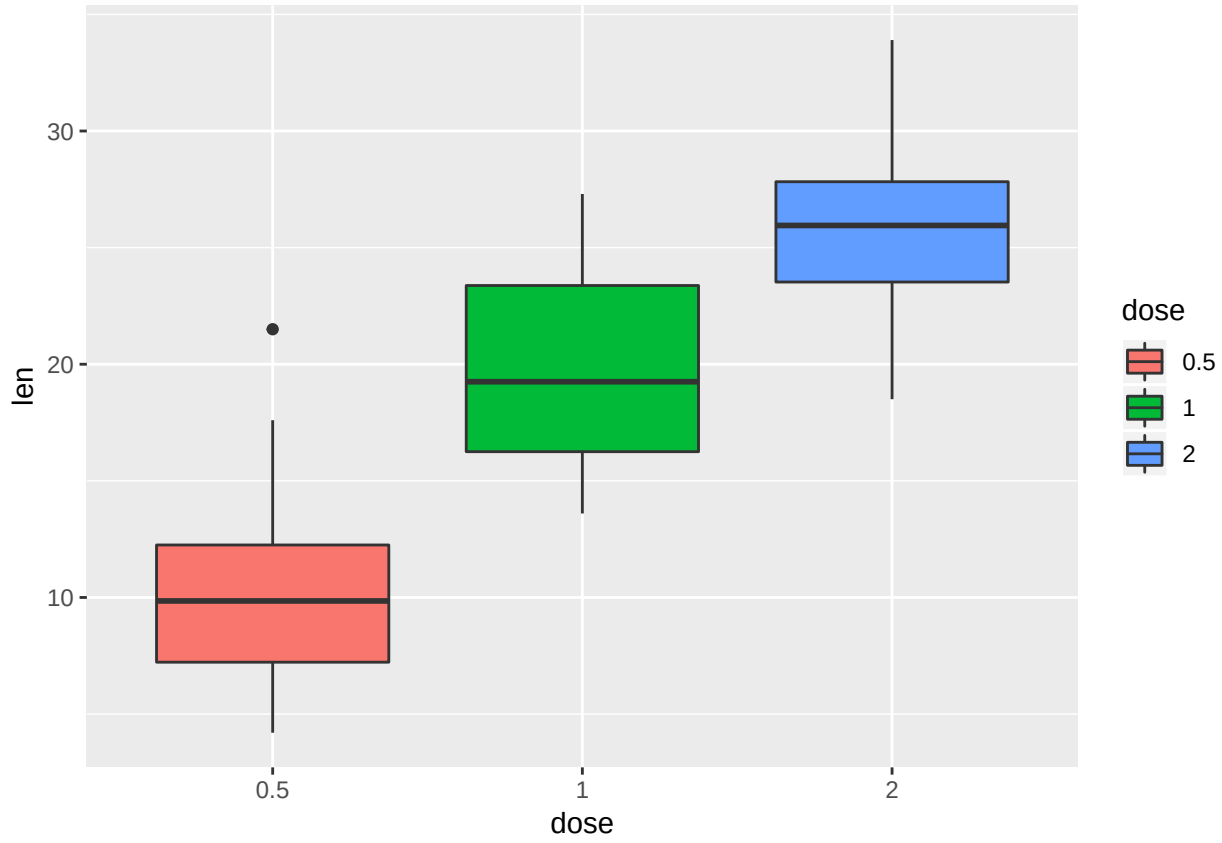
```
sum(complete.cases(ToothGrowth))
```

```
## [1] 60
```

The data contains no missing values

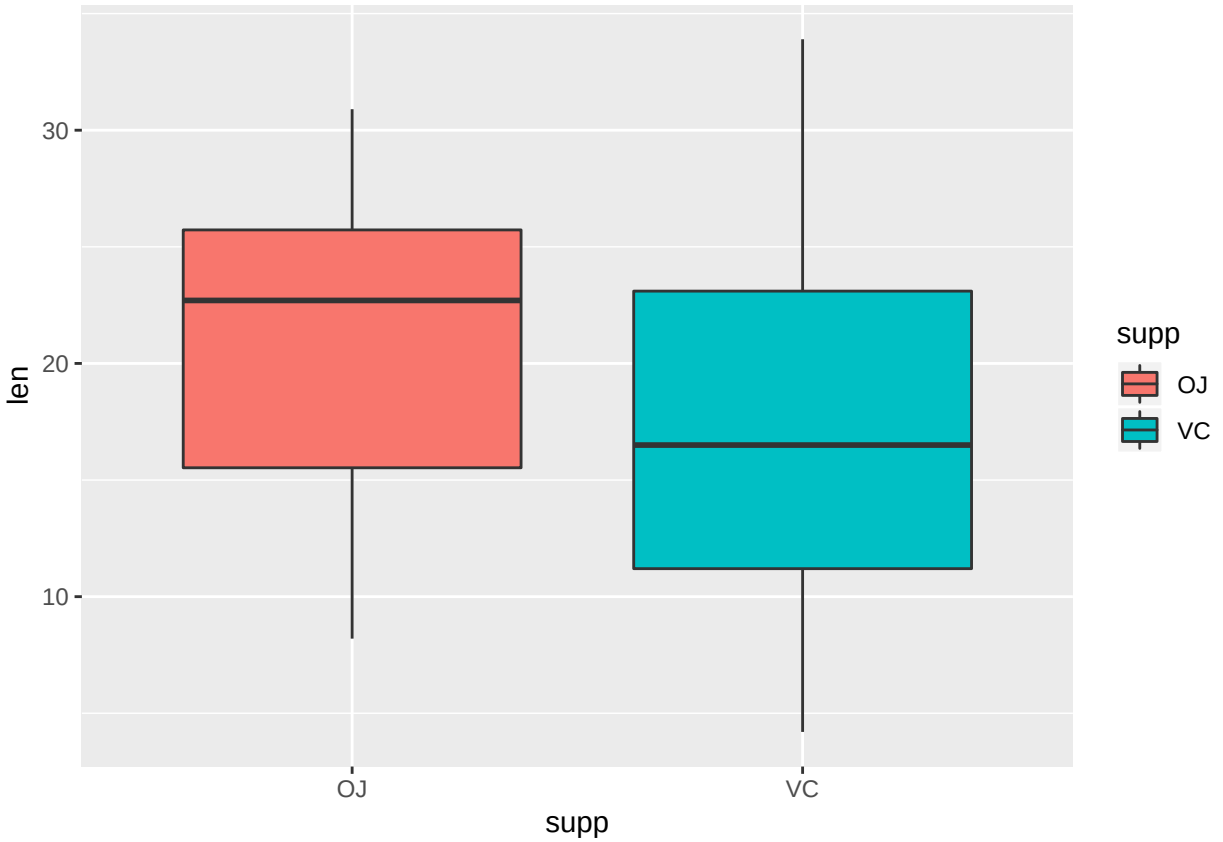
2. Boxplots to represent the 5 number summary for each group by dose

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
ggplot(ToothGrowth, aes(x = dose, y = len, fill = dose)) + geom_boxplot()
```



So there is very obvious increase in tooth length with increasing the dose.

```
ggplot(ToothGrowth, aes(x = supp, y = len, fill = supp)) + geom_boxplot()
```



Orange juice treatment appears to give a more predictable and better result by having a higher median growth in length and a smaller range, but it is interesting to look at the extreme results of the Vitamin C in case the treatment works better for these cases.

Hypothesis testing

Here we test the effect of the dose and the method of intaking Vitamin C on the teeth growth

Dose effect

Our null hypothesis will be that different doses, do not affect the growth length, the data is not paired as they are for different subjects.

```
t.test(ToothGrowth$len[which(ToothGrowth$dose == 1)], ToothGrowth$len[which(ToothGrowth$dose == .5)], pa
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[which(ToothGrowth$dose == 1)] and ToothGrowth$len[which(ToothGrowth$dose == 0
## 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:
##   6.276219 11.983781
## sample estimates:
## mean of x mean of y
```

```
##    19.735    10.605
```

The test shows the following:

- The p-value < 0.05 which indicates signifacnt result and we reject the null hypothesis.
- The confidence interval does not contain 0 therefore we can infer that a dose of 1 mg is better than 0.5 mg.

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

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[which(ToothGrowth$dose == 2)] and ToothGrowth$len[which(ToothGrowth$dose == 1
## 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:
##  3.733519 8.996481
## sample estimates:
## mean of x mean of y
##    26.100    19.735
```

The results of the 2 mg dose vs the 1 mg dose shows similar results that the 2 mg is better than the 1 mg.

The method of intake

```
t.test(ToothGrowth$len[which(ToothGrowth$supp == "OJ")], ToothGrowth$len[which(ToothGrowth$supp == "VC"
```

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

The test shows the following:

- The p-value > 0.05 which indicates non-signifacnt result and we fail to reject the null hypothesis, though it is relatively close to our $\alpha = 0.5$.
- The confidence interval does contain 0 therefore the results is insignificant.