SI Course Project Part 2

Sasa Pakvovic

Course project task 2 description

Now in the second portion of the class, 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 hypothesis tests to compare tooth growth by supp and dose. (Use the techniques from class even if there's other approaches worth considering)
4. State your conclusions and the assumptions needed for your conclusions.

Introduction

The ToothGrowth dataset is already available in R so we can load it easily. After loading the dataset a bit more information can be obtained with help(ToothGrowth) command. The data set comes from the study named The Effect of Vitamin C on Tooth Growth in Guinea Pigs and tracks the response in the 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).I believe the basic question we need to ask ourselves during the analysys is: How does tooth growth differ with different dosages and different supplements?

Exploring data

We can load the dataset into memory by calling library(datasets). Also, check the ToothGrowth dataset description page (Reference [1]).

```
library(datasets)
```

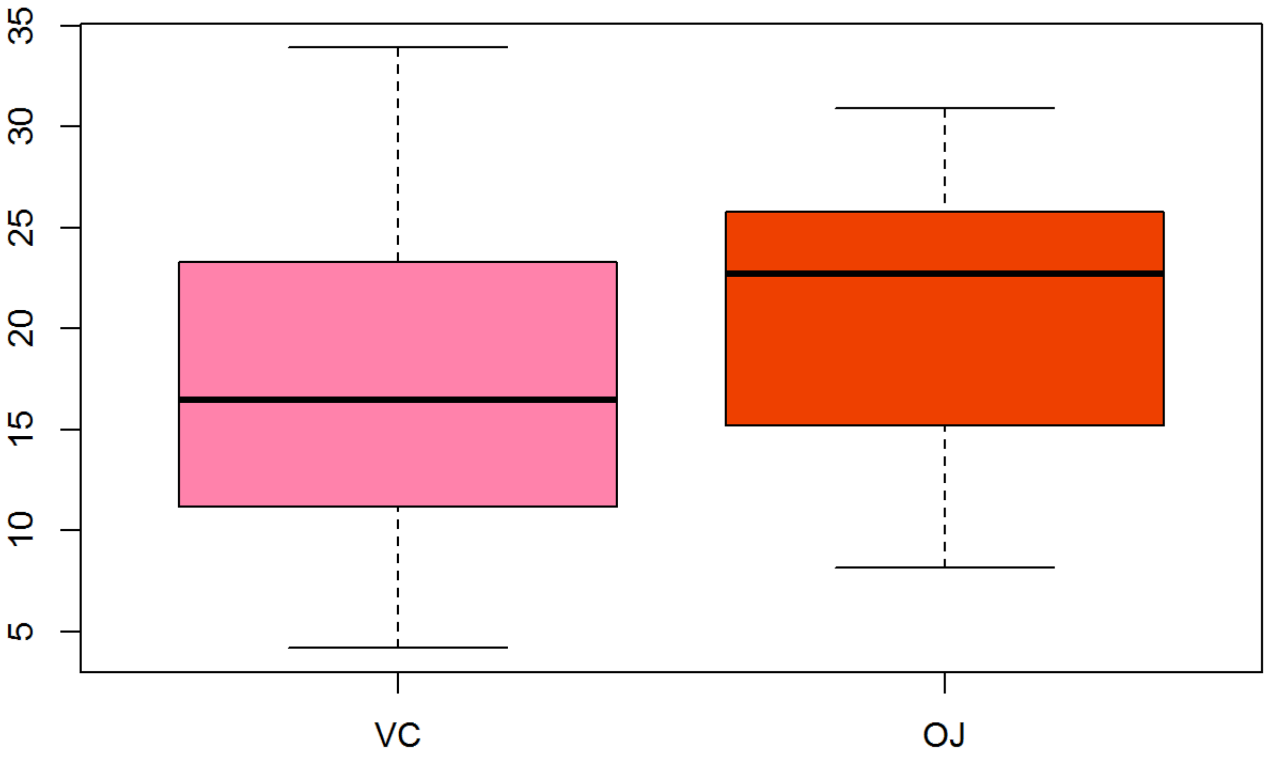
We can see that the dataset consist of only 3 columns and 60 observations. We have a factor variable supp with 2 factors OJ (orange juice) and VC (ascorbic acid).

```
## '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 ...
```

I will add better names for columns so its easier to understand what each of the columns represent.

```
colnames(ToothGrowth) <- c("toothlength","supplement","dosage")
```

From the overview of a single supplement we can see that there are 3 different dosages of the same amount for each of the supplements. We can see the boxplots for VC and OJ respectively here.



Basic summaries

We can show some basic summary of the data by each of the supplements in regards to the tooth length achieved.

First OJ

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     8.2    15.5    22.7    20.7    25.7    30.9
```

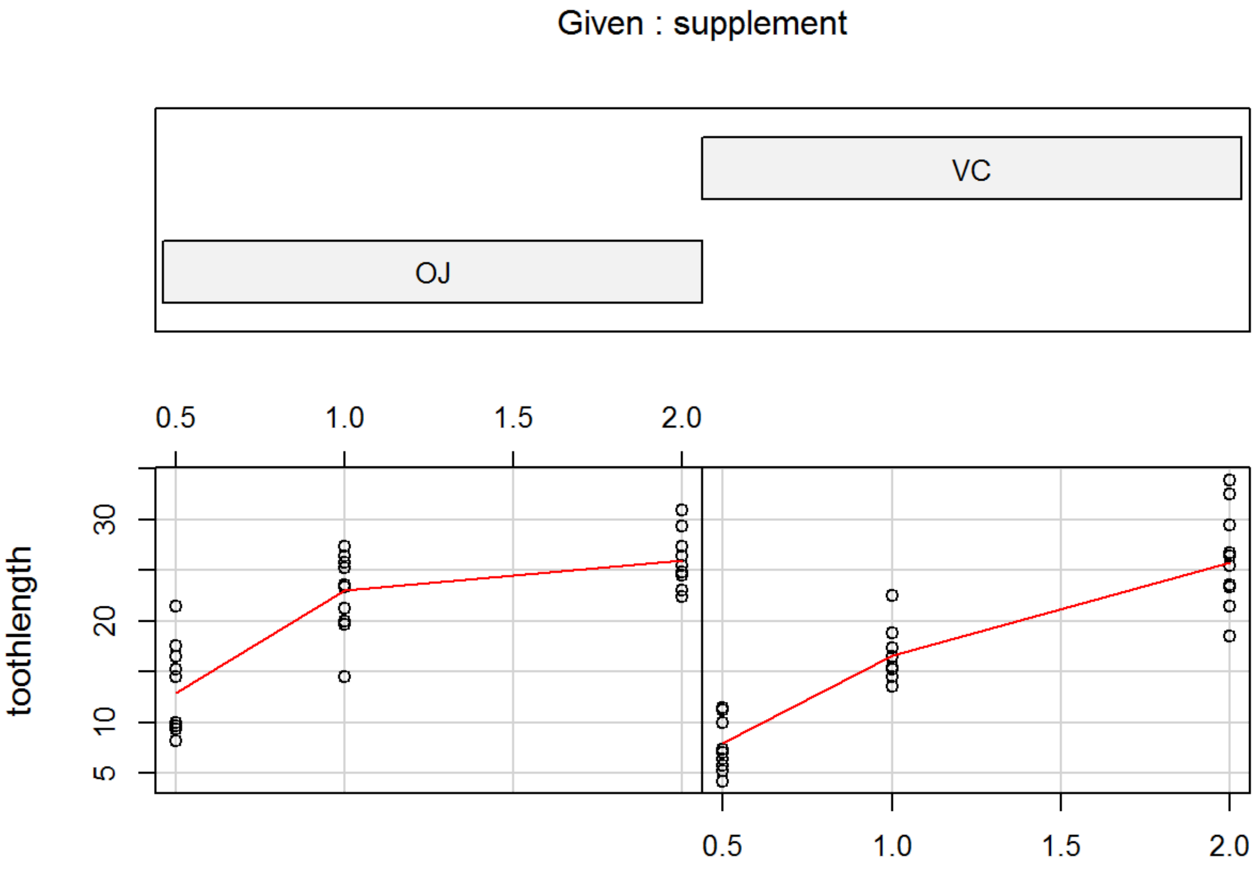
then VC

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     8.2    15.5    22.7    20.7    25.7    30.9
```

We can also notice that we have 6 groups of 10 guinea pigs for each of the combinations of supplement and dosage. with the help of plyr package we can see the summaries for each of the groups a bit clearer.

```
##   supplement dosage  mean stddev variance median
## 1         OJ    0.5 13.23  4.460   19.889  12.25
## 2         OJ    1.0 22.70  3.911   15.296  23.45
## 3         OJ    2.0 26.06  2.655    7.049  25.95
## 4         VC    0.5  7.98  2.747    7.544   7.15
## 5         VC    1.0 16.77  2.515    6.327  16.50
## 6         VC    2.0 26.14  4.798   23.018  25.95
```

Summary from the previous slide we can see in this coniditional plot taken from the description of the ToothGrowth dataste (reference [1]).



ToothGrowth data: length vs dose, given type of supplement

Both the data summary and the coplot suggest, that on average: 1. Vitamin C helps in tooth growth of the subjects 2. Both supplements, orange juice and ascorbic acid, seem to be effective way of delivering vitamin C. 3. The higher the dosage of supplements the higher the tooth lengths are 4. The averages of OJ supplement seems to be producing higher effect than VC in tooth growth for the lower and indermediate dosages, but the effect seems to be the same for the high dosage.

Point 4 we will want to test by forming a hypothesis and testing it.

Analysis

Assumptions

1. No information is available if the the selected subjects are indeed chosen randomly from the population

of subjects. We will assume so.
2. No information is available that confirms or denies if the same 10 subjects were used repeatedly in any or all of the 6 experiments. We will assume that only one combination of supplement and dosage was used per subject.
3. The variances are assumed to be unequal.

Statements

We want to test these statements:

1. Orange juice (OJ) is on average producing a bigger effect in tooth growth than Ascorbic acid (VC) with dosage of 0.5mg.
2. Orange juice (OJ) is on average producing a bigger effect in tooth growth than Ascorbic acid (VC) with dosage of 1mg.
3. Orange juice (OJ) is on average producing a bigger effect in tooth growth than Ascorbic acid (VC) with dosage of 2mg.

As the sample size for each group is relatively small we will use t distribution and t test for testing the hypothesis. But first we need to prepare the data in the format supplement per dosage forming 6 columns of 10 observations.

```
##    OJ05  OJ1  OJ2 VC05  VC1  VC2
## 1  15.2 19.7 25.5  4.2 16.5 23.6
## 2  21.5 23.3 26.4 11.5 16.5 18.5
## 3  17.6 23.6 22.4  7.3 15.2 33.9
## 4   9.7 26.4 24.5  5.8 17.3 25.5
## 5  14.5 20.0 24.8  6.4 22.5 26.4
## 6  10.0 25.2 30.9 10.0 17.3 32.5
## 7   8.2 25.8 26.4 11.2 13.6 26.7
## 8   9.4 21.2 27.3 11.2 14.5 21.5
## 9  16.5 14.5 29.4  5.2 18.8 23.3
## 10  9.7 27.3 23.0  7.0 15.5 29.5
```

Hypothesis

For testing the above mentioned statements we will use two sided t test where population variances are unequal and unknown. We create a null hypothesis (H0) where we say that means of both supplements for the same dosage are equal using 95% interval of the t distribution.

T test for dosage of 0.5mg provides these results:

```
##
##  One Sample t-test
##
## data:  tgperdose["OJ05"] - tgperdose["VC05"]
## t = 2.979, df = 9, p-value = 0.01547
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.263 9.237
## sample estimates:
## mean of x
##      5.25
```

As the p-value is smaller then the 5% we reject the null hypothesis, and say that the statement 1 is true.

T test for dosage of 1mg provides these results:

```
##
##  One Sample t-test
##
## data:  tgperdose["OJ1"] - tgperdose["VC1"]
## t = 3.372, df = 9, p-value = 0.008229
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.952 9.908
## sample estimates:
## mean of x
##      5.93
```

Again, as the p-value is smaller then the 5% we reject the null hypothesis, and say that the statement 2 is true.

T test for dosage of 2mg provides these results:

```
##
##  One Sample t-test
##
## data:  tgperdose["OJ2"] - tgperdose["VC2"]
## t = -0.0426, df = 9, p-value = 0.967
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -4.329  4.169
## sample estimates:
## mean of x
##     -0.08
```

Here we have p-value that is way bigger then the 5% so we have accept the null hypothesis, and say that the statement 3 is false.

Conclusion

With exploratory data analysis we noticed some patterns in data that we wanted to state, but were not completely sure if we should. Watching the grouped data we formed 3 statements that we wanted to check. We wanted to say that orange juice is on average a better supplement for providing vitamin C for tooth growth no matter the dosage. After forming a hypothesis and checking the two data sets with t test confidence interval of 95% we could only confirm that orange juice is on average a supplement which ensures more tooth growth when distributed in dosages of 0.5mg and 1mg. With dosage of 2 mg on average there is no difference in tooth growth when given any of the the two supplements (orange juice or ascorbic acid)

References

[1].The Effect of Vitamin C on Tooth Growth in Guinea Pigs. URL:https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ToothGrowth.html (https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ToothGrowth.html)