JHU Coursera: Statistical Inference Assignment1_2:

Dheeraj Agarwal

October 24, 2015

Contents

SCOPE:

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 conclusion

---

Global Settings The below global setting allow the code to be visible and switches the scientific notation of large numbers, if any, off.

```
knitr::opts_chunk$set(echo = TRUE, options(scipen=999))
library(ggplot2)
```

PART 1: DATA LOAD & EXPLORATORY ANALYIS

The below section will load the data as well as perform some function based basic exploratoy analysis.

```
library(datasets)
data("ToothGrowth")
head(ToothGrowth, 5) # Display the first 5 rows and all columns
```

```
##    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
```

```r
dim(ToothGrowth) # No of Rows and Columns in the entire data set
```

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

All the three columns has the following classes: **numeric**, **factor** & **numeric** respectively.

```r
range(ToothGrowth$len)
```

```
## [1]  4.2 33.9
```

```r
unique (ToothGrowth$supp)
```

```
## [1] VC OJ
## Levels: OJ VC
```

```r
unique (ToothGrowth$dose)
```

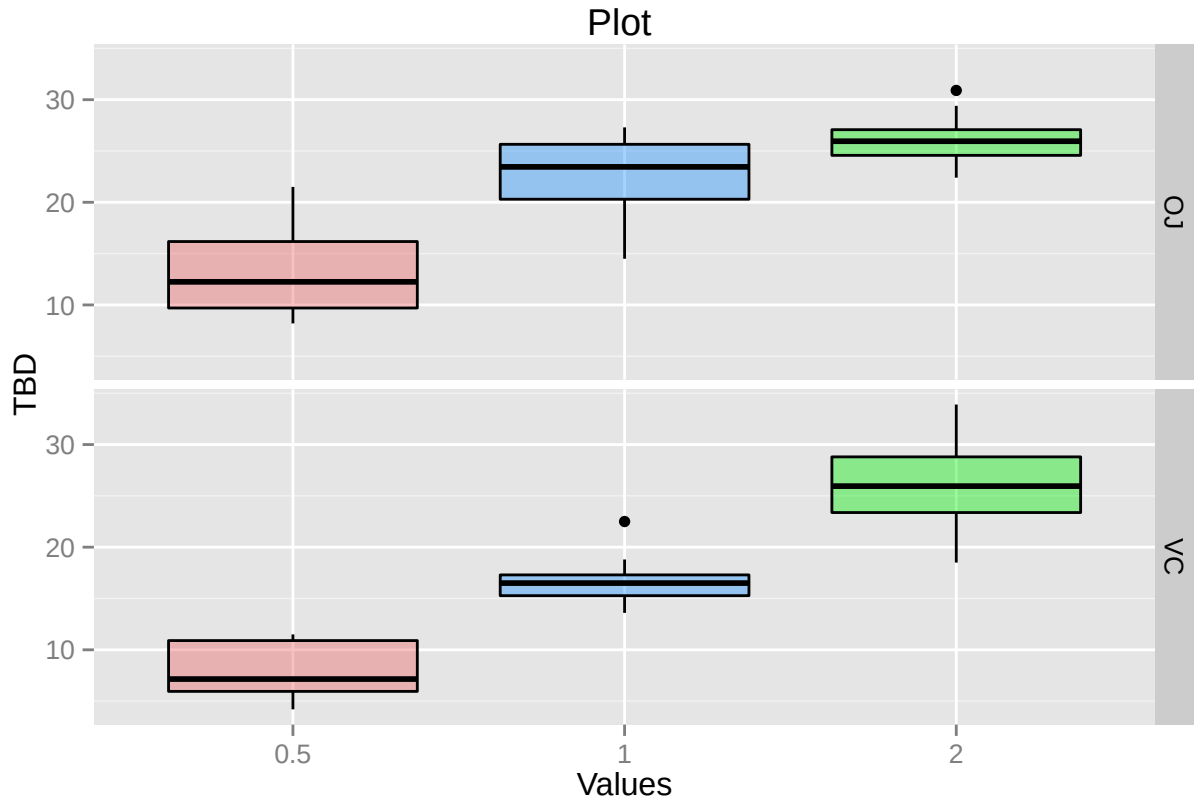
```
## [1] 0.5 1.0 2.0
```

```r
group <- aggregate(ToothGrowth[,1], as.list(ToothGrowth[,2:3]), FUN=median)
colnames(group)[3] <- "MedianLength"
group
```

```
##   supp dose MedianLength
## 1   OJ  0.5        12.25
## 2   VC  0.5         7.15
## 3   OJ  1.0        23.45
## 4   VC  1.0        16.50
## 5   OJ  2.0        25.95
## 6   VC  2.0        25.95
```

The above median calculations are also depicted below with the help of the box plots.

```r
# Plotting the data for each quantity of dose and
 # delivery method against length of tooth in the guinea pigs.
ggplot(data = ToothGrowth, aes(x=as.factor(dose), y=len))+
        geom_boxplot(
                col = "black",
                fill = c("indianred2", "dodgerblue", "green2"),
                alpha = 0.4
        ) +
        labs(title = "Plot ", x = "Values", y = "TBD") +
        facet_grid(supp ~ .)
```



Based on the above boxplot, we can see that there is a marked shift in the tooth length as dosage increase. Based on the median values it also appears that that tooth length vary by the supplement type (OJ vs VC).

PART 2: HYPOTHESIS FORMULATION & INFERENTIAL ANALYSIS

Given the above exploratory analysis, the:
First NULL Hypothesis or H0(1) is: OJ and VC have different effects on tooth length
Second NULL Hypothesis or H0(2) is: Dosage has an increaing effect on tooth length

Data Prep

```
supp_val <- split(ToothGrowth$len, ToothGrowth$supp)
sapply(supp_val,var)
```

```
##       OJ       VC
## 43.63344 68.32723
```

```
dose_val <- split(ToothGrowth$len, ToothGrowth$dose)
sapply(dose_val,var)
```

```
##      0.5        1        2
## 20.24787 19.49608 14.24421
```

These variances are not close to each other hence any subsequent t-tests cannot be assumed to have a constant variance. Also the values cannot be considered paired for t-tests. This is because the tests were performed on different populations. That is the same subjects were not provided both the tests. Now, the reason a t-test is considered and not any other is because the t-test will help support (or contradict) the inferences made by the plots alone. That is test the hypothesis.

Testing NULL Hypothesis 1

```
t.test(
        ToothGrowth$len[ToothGrowth$supp == "VC"],
        ToothGrowth$len[ToothGrowth$supp == "OJ"],
        paired=FALSE, var.equal=FALSE
)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "VC"] and ToothGrowth$len[ToothGrowth$supp == "OJ"]
## 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:
##  -7.5710156  0.1710156
## sample estimates:
## mean of x mean of y
##  16.96333  20.66333
```

From the values above, there are two important observations: 1. the **p value** is **0.06** 2. the 95% confidence interval is **-7.57 to 0.17**

That is the 95% confidence interval contains the value 0 and pvalue is greater than the usual standard of 5%. Hence we ***can reject the NULL***. It cannot be conclusively said if the supplment type makes an impactful difference.

Testing NULL Hypothesis 2

For this test, we will run t-tests for two samples at a time. Since we have 3 different dosages, we will have 3 samples with two dosages selected at a time for tests.

At first, comparing the 0.5 and 1 mg dosages

```
t.test(
        ToothGrowth$len[ToothGrowth$dose==1],
        ToothGrowth$len[ToothGrowth$dose==0.5],
        paired = FALSE, var.equal = FALSE
)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 1] and ToothGrowth$len[ToothGrowth$dose == 0.5]
## t = 6.4766, df = 37.986, p-value = 0.0000001268
## 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
```

Comparing the 0.5 and 2 mg dosage

```
t.test(
        ToothGrowth$len[ToothGrowth$dose==2],
        ToothGrowth$len[ToothGrowth$dose==0.5],
        paired = FALSE, var.equal = FALSE
)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 2] and ToothGrowth$len[ToothGrowth$dose == 0.5]
## t = 11.799, df = 36.883, p-value = 0.00000000000004398
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  12.83383 18.15617
## sample estimates:
## mean of x mean of y
##    26.100    10.605
```

And finally comparing the 1 mg and 2 mg dosage

```
t.test(
        ToothGrowth$len[ToothGrowth$dose==2],
        ToothGrowth$len[ToothGrowth$dose==1],
        paired = FALSE, var.equal = FALSE
)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 2] and ToothGrowth$len[ToothGrowth$dose == 1]
## t = 4.9005, df = 37.101, p-value = 0.00001906
## 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
```

Consolidating all the above results in a table. The confidence interval range is rounded to 2 decimals.

Measure	**p-value**	**95% confidence interval**
0.5 & 1.0 mg	0	06.28 - 11.98
0.5 & 2.0 mg	0	12.83 - 18.16
1.0 & 2.0 mg	0	03.77 - 9.00

In all cases, the **p-value** is **0**. Also the 95% confidence interval in all instances, do not contain **0** hence we ***cannot reject the NULL***. There is a marked impact of dosage on the tooth length.

DISCLAIMER: This concludes my report based on my understanding of the problem statements. This report was created as part of Coursera Data Science Specialization::Reproducible Research course starting Oct 5th 2015. Please feel free to use it as a reference, however if you are taking the course yourself, please do not copy it.