Statistical Inference - Course Project (Task 2)

Joscha Frischherz

September 26, 2015

The second task of this project will perform a data analysis over the Tooth Growth dataset in R. We load the ToothGrowth dataset into R by selecting the 'datasets' library and read the data into the 'tg' variable. In addition to that we load the 'sqldf' package for easy interaction with the data via SQL as well as the 'lattice' package for graphical analysis. Refer to Appendix for the relevant code.

```
## Loading required package: gsubfn
## Loading required package: proto
## Loading required package: RSQLite
## Loading required package: DBI
```

Exploratory Analysis

We perform some basic review of the data (Refer to appendix for code and results).

- looking at the top rows of the data to gather a view of how the dataset looks like;
- review the distinct parameters of supp;
- review the distinct parameters of dose;
- run a summary across the dataset to gain additional insights
- plot a histogram to outline the distribution of the len variable; and
- plot a boxplot chart to show how tooth growth is impacted by the supplement and doses.

We learn the following from the above analysis:

- The data set contains tooth growth data depending on two supplements (OJ [Orange Juice] & VC [Vitamin C]), which can be given in three doses (0.5, 1 & 2).
- The dataset has 60 rows.
- The toothgrowth data by itself does not appear normally distributed at first sight.
- Per the box plot, it appears that higher doses of supplements will increase the tooth length.

Data Analysis

We will analyse whether supplement OJ or VC is more effective for tooth growth regardless of the dose provided.

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

```
##
##  Welch Two Sample t-test
##
## data:  tg$len[tg$supp == "OJ"] and tg$len[tg$supp == "VC"]
## 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
```

We conclude from the above that there is no difference in the growth by the supplement given:

- the resulting confidence intervall (95%) [-0.1710156, 7.5710156] includes 0; and
- the resulting p-value of 0.06063 exceeds the required 0.05 (5%)

From the box plot (appendix) it appears that doses of 2 increase the tooth length compared to doses of 0.5 regardless of the supplement. We test this Hypothesis as well:

```
t.test(tg$len[tg$dose == "2"],tg$len[tg$dose == "0.5"], paired = FALSE )
```

```
##
##  Welch Two Sample t-test
##
## data:  tg$len[tg$dose == "2"] and tg$len[tg$dose == "0.5"]
## 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:
##  12.83383 18.15617
## sample estimates:
## mean of x mean of y
##    26.100    10.605
```

We conclude from the above that we accept the Hypothesis that a dose of 2 has a higher impact on tooth growth compared to a dose of 0.5 given:

- the resulting confidence intervall (95%) [12.83383 18.1561] does not include 0; and
- the p value approximates 0 (4.398e-14) which is less than 0.05 (5%).

Next, we would like to also perform this test to compare the dose of 1 to 2.

```
t.test(tg$len[tg$dose == "2"],tg$len[tg$dose == "1"], paired = FALSE )
```

```
##
##  Welch Two Sample t-test
##
## data:  tg$len[tg$dose == "2"] and tg$len[tg$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
```

Again we conclude that the dose of 2 is more effective for tooth growth:

- the resulting confidence intervall (95%) [3.733519 8.996481] does not include 0; and
- the p value approximates 0 (1.906e-05) which is less than 0.05 (5%).

Finally, we want to test whether eigher dose of 2 from OJ or VC is more effective in triggering tooth growt.

```
##Subset the length for OJ and VC with dose of 2:
OJ2<- sqldf("SELECT len FROM tg WHERE dose = '2' AND supp = 'OJ'")
```

```
## Loading required package: tcltk
```

```
VC2<- sqldf("SELECT len FROM tg WHERE dose = '2' AND supp = 'VC'")

##perform statistical test
t.test(OJ2,VC2, paired = FALSE )
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ2 and VC2
## 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 of x mean of y
##     26.06     26.14
```

We conclude that neither OJ nor VC with a dose of 2 is more effective for tooth growth given:

- the resulting confidence intervall (95%) [-3.79807 3.63807] does include 0; and
- the p-value of 0.9639 is larger than 0.05 required by the 95% test.

Conclusion

From the above statistical analysis we conclude:

- Neigher Supplement Orange Juice nor Vitamin C is overall more effective with regards to tooth growth;
- Overall a dose of 2 of either supplement is most effective; and
- For a dose of 2 neither supplement is more effective.

This is under the assumption that all test subjects were independent of each other.

Appendix

R code applied for loading of data and packages

data load

library(datasets)

tg <- ToothGrowth

SQL packages

library(sqldf)

Lattice package

library(lattice)

Code and results from initial data analysis

```
##Top 6 rows
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
```

```
##review of supp
sqldf("SELECT DISTINCT supp FROM tg")
```

```
##   supp
## 1   VC
## 2   OJ
```

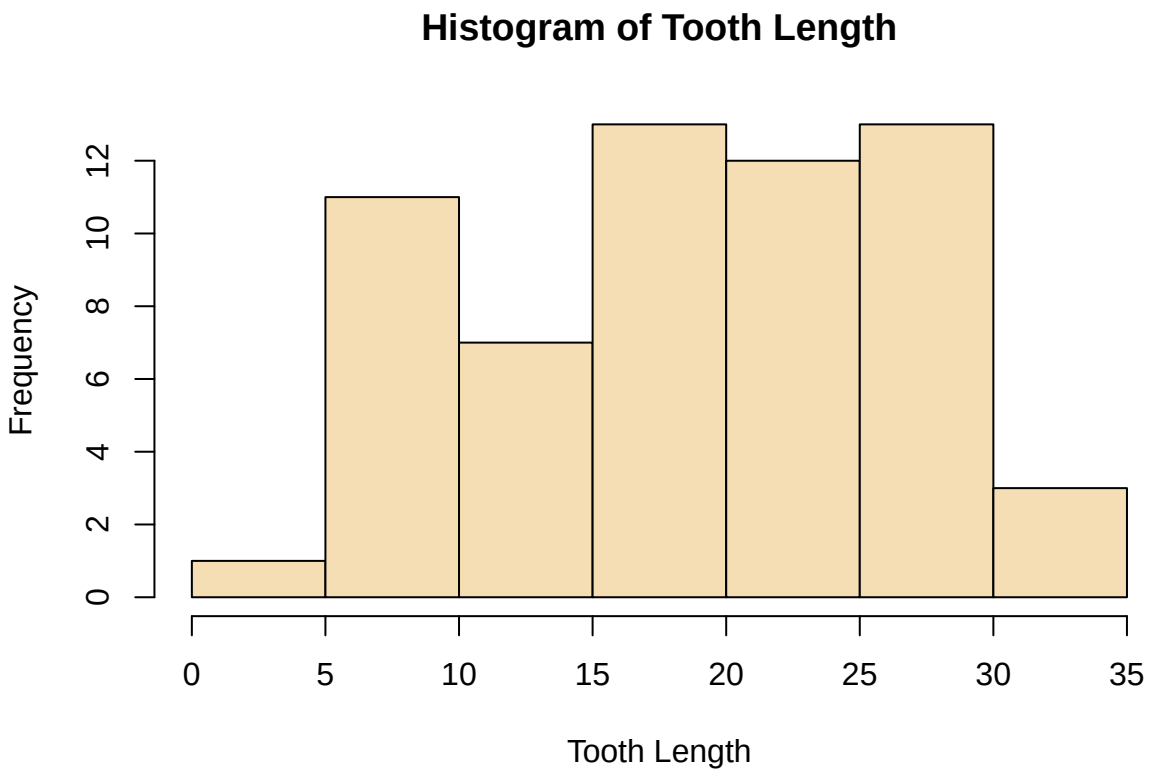
```
##review of dose
sqldf("SELECT DISTINCT dose FROM tg")
```

```
##   dose
## 1  0.5
## 2  1.0
## 3  2.0
```

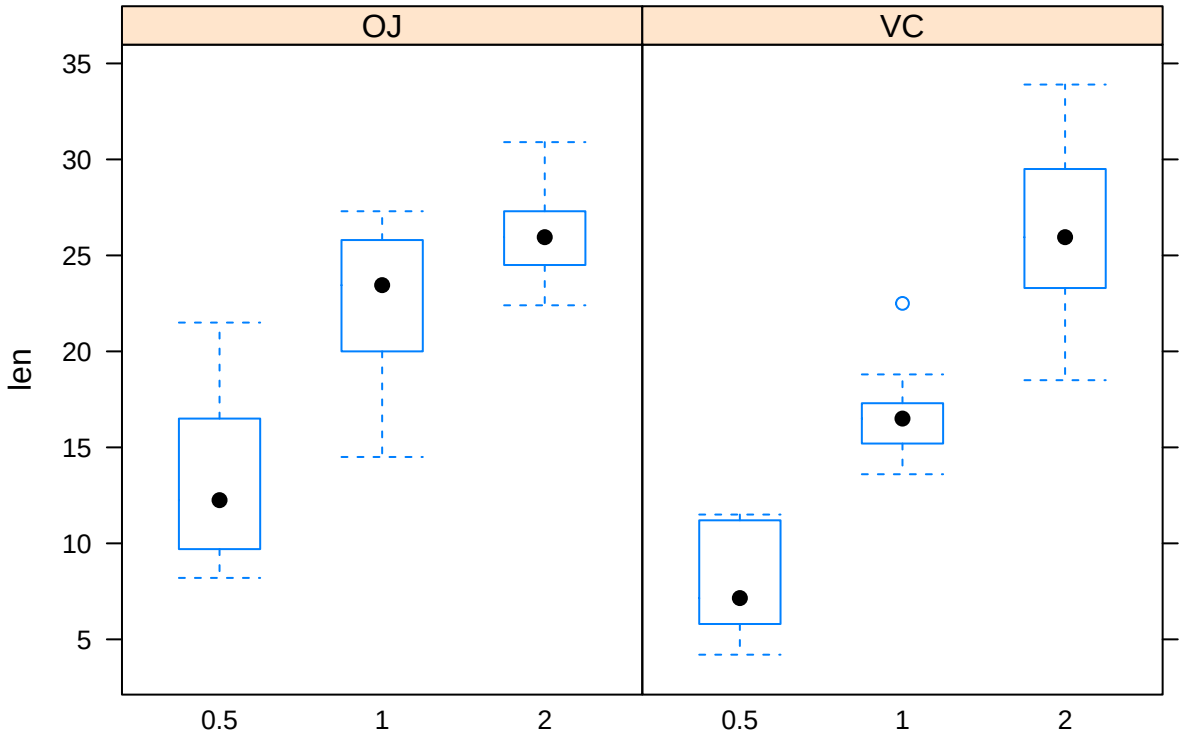
```
##data summary
summary(tg)
```

```
##       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
```

```
##histogram of len
hist(tg$len, col = "wheat", main = "Histogram of Tooth Length", xlab = "Tooth Length")
```




```
##Boxplot of dose & supplements
tg$dose <-as.factor(tg$dose)
bwplot(len ~ dose | supp, data=tg)
```


OJ
VC
len
35
30
25
20
15
10
5
0.5
1
2
0.5
1
2