Course Project for Statistical Inference by Hsin-Hua Lai

Analysis of the ToothGrowth data in the R datasets package

Overview

In this course project we will analyze the ToothGrowth data in the R datasets packages. We use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

Load the ToothGrowth data and perfomr some basic exploratory data analysis

```
## We load the ToothGrowth datasets
data(ToothGrowth)

## Give the summary of the data
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 also give the str information of ToothGrowth
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 ...
```

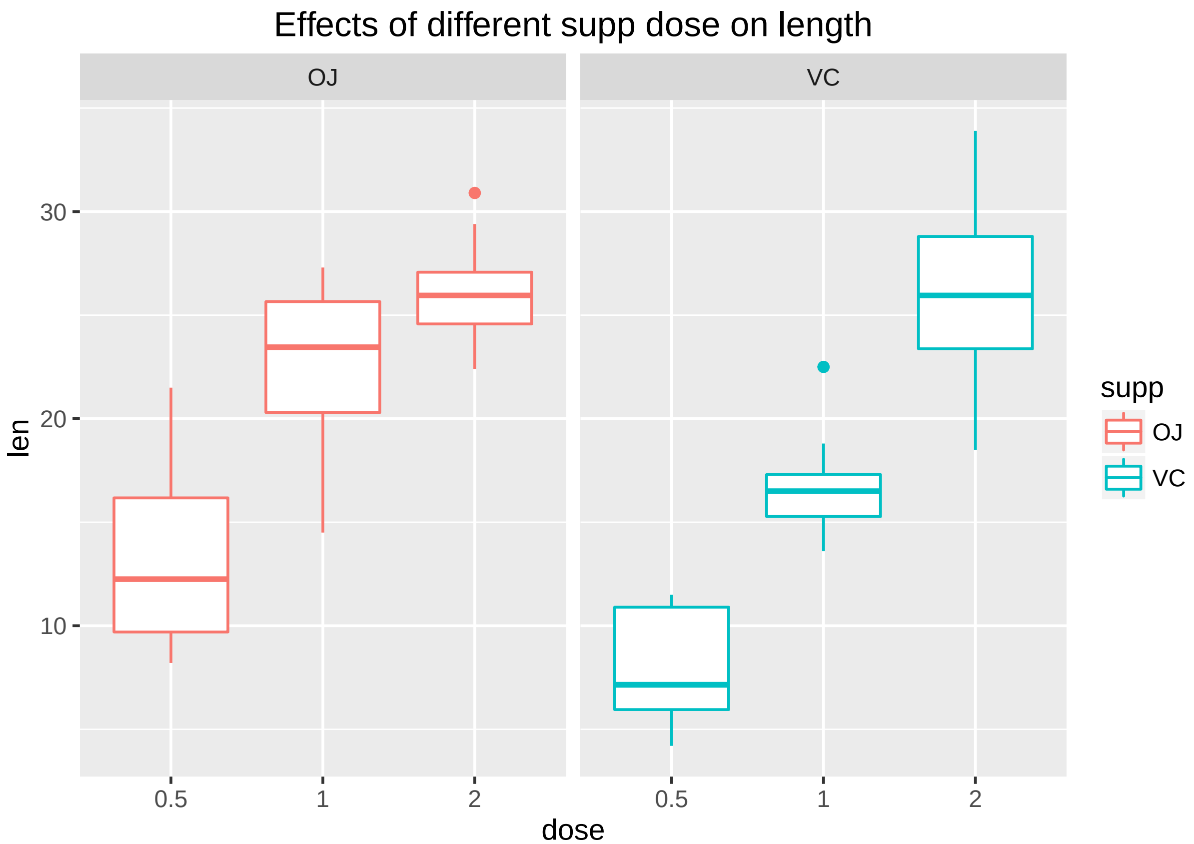
We can see there are three variables—**"len", "supp", and "dose"**, and there are 60 observables. The columns of "len" and "dose" are numeric values in mg and the column of "supp" is a Factor column. In the supp column, 30 observables are OJ and 30 observables are VC, where OJ represents Orange Juice and VC represents Ascorbic Acid (a form of vitamin C).

We first check the effects of dose on the length due to different supps (OJ or VC).

```
## Let's use ggplot2 package
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

```
g <- ggplot(ToothGrowth, aes(x = dose, y=len)) + geom_boxplot(aes(factor(dose),color=supp)) + facet_gri
print(g)
```



From the plot above, we can clearly see that for a lower dose (<2 mg) the OJ supplement (Left) much is more efficient for tooth growth. For a higher dose (2 mg) the VC and OJ are comparable since from the box plot the means of these two data are nearly equal.

Basic summary of the data

In order to have a more qualitative understanding of the data, let's use the dplyr package to reorganize the data to give a qualitative summary of the data

```
## We import dplyr package
library(dplyr)

## Let's group the data by the supp and dose and treat len as a variable
groupdata <- group_by(ToothGrowth, supp, dose)

## Now we can give a summary of the group data
sumdata <- summarize(groupdata, LenMean = mean(len), LenSd = sd(len), number = n())
print(sumdata)
```

```
## Source: local data frame [6 x 5]
## Groups: supp
##
##   supp dose LenMean    LenSd number
## 1   OJ  0.5   13.23 4.459709     10
## 2   OJ  1.0   22.70 3.910953     10
## 3   OJ  2.0   26.06 2.655058     10
```

```
## 4   VC  0.5    7.98 2.746634     10
## 5   VC  1.0   16.77 2.515309     10
## 6   VC  2.0   26.14 4.797731     10
```

We obtain a summary of data giving the means of the len **(LenMean)** for each supp at each dose. We notice that each observable number is 10 and we also obtain the standard deviation of the len **(LenSd)** in the subdata. From this data, we again confirm the conjectures presented above based on the ggplot figures.

Confidence intervals and/or hypothesis tests

Hypothesis: Supplement OJ on average is more efficient on tooth growth than supplement VC is on average. We can use the sumdata we obtained above which contains aveages and standard deviations of data of length of tooth growth for diffent dose and supp, with each sample size being 10. Let's analyze the data for each dose: **(1) 0.5 mg, (2) 1 mg, and (3) 2 mg**

```r
## With Sumdata above, let's first select out dose = 0.5 mg subdata
row0.5 <- which(sumdata$dose == 0.5)
subdata0.5 <- sumdata[row0.5,]

## Let's print out the subdata of dose = 0.5
print(subdata0.5)
```

(1) 0.5 mg

```
## Source: local data frame [2 x 5]
## Groups: supp
##
##   supp dose LenMean    LenSd number
## 1   OJ  0.5   13.23 4.459709     10
## 2   VC  0.5    7.98 2.746634     10
```

```r
## Let's check the t confidence inteval
## First take the LenMean difference
Diff_lenmean0.5 <- subdata0.5$LenMean[1] - subdata0.5$LenMean[2]

## 95% t confidence interval
Diff_lenmean0.5 + c(-1,1)*qt(0.975, 18)*sqrt(subdata0.5$LenSd[1]^2/10 + subdata0.5$LenSd[2]^2/10)
```

```
## [1] 1.770262 8.729738
```

We can see that at dose 0.5 mg, the t confidence interval of difference between the averages of the length of tooth growth of OJ and VC is completely above 0. Therefore, we **ACCEPT** the hypothesis at dose 0.5 mg.

```r
## With Sumdata above, let's first select out dose = 1 mg subdata
row1 <- which(sumdata$dose == 1)
subdata1 <- sumdata[row1,]
```

```r
## Let's print out the subdata of dose = 1
print(subdata1)
```

(2) 1 mg

```
## Source: local data frame [2 x 5]
## Groups: supp
##
##   supp dose LenMean    LenSd number
## 1   OJ    1   22.70 3.910953     10
## 2   VC    1   16.77 2.515309     10
```

```r
## Let's check the t confidence inteval
## First take the LenMean difference
Diff_lenmean1 <- subdata1$LenMean[1] - subdata1$LenMean[2]

## 95% t confidence interval
Diff_lenmean1 + c(-1,1)*qt(0.975, 18)*sqrt(subdata1$LenSd[1]^2/10 + subdata1$LenSd[2]^2/10)
```

```
## [1] 2.840692 9.019308
```

Again we see that at dose 1 mg, the t confidence interval is still completely above 0. We therefore **ACCEPT** the Hypothesis at dose 1 mg.

```r
## With sumdata above, let's first select out dose = 2 mg subdata
row2 <- which(sumdata$dose == 2)
subdata2 <- sumdata[row2,]

## Let's print out the subdata of dose = 2
print(subdata2)
```

(2) 2 mg

```
## Source: local data frame [2 x 5]
## Groups: supp
##
##   supp dose LenMean    LenSd number
## 1   OJ    2   26.06 2.655058     10
## 2   VC    2   26.14 4.797731     10
```

```r
## Let's check the t confidence inteval
## First take the LenMean difference
Diff_lenmean2 <- subdata2$LenMean[1] - subdata2$LenMean[2]

## 95% t confidence interval
Diff_lenmean2 + c(-1,1)*qt(0.975, 18)*sqrt(subdata2$LenSd[1]^2/10 + subdata2$LenSd[2]^2/10)
```

```
## [1] -3.722999  3.562999
```

We can see that at dose 2 mg t confidence interval actually *contains* 0. Therefore, we need to **REJECT** the Hypothesis at dose 2 mg.

Conclusion

We conclude that the Hypothesis that OJ is more efficient than VC is accepted for a lower dose ($<= 1$ mg), while at a higher dose (2 mg) is rejected and the efficiency of OJ and VC are comparable.