Statistical Inference - Project - Basic Inferential Data Analysis

Author: Wilson Mok

Overview

In this assignment, we are going to analyze the ToothGrowth data in R datasets package. We will be performing:
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 conclusions.

Libraries

```
library(data.table)
library(dplyr)
library(ggplot2)
```

Loading data and basic analyses

```
library(data.table)
library(dplyr)
library(ggplot2)

data("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 ...
```

```
levels(ToothGrowth$supp)
```

```
## [1] "OJ" "VC"
```

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

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

```
sum(is.na(ToothGrowth$len))
```

```
## [1] 0
```

```
sum(is.na(ToothGrowth$supp))
```

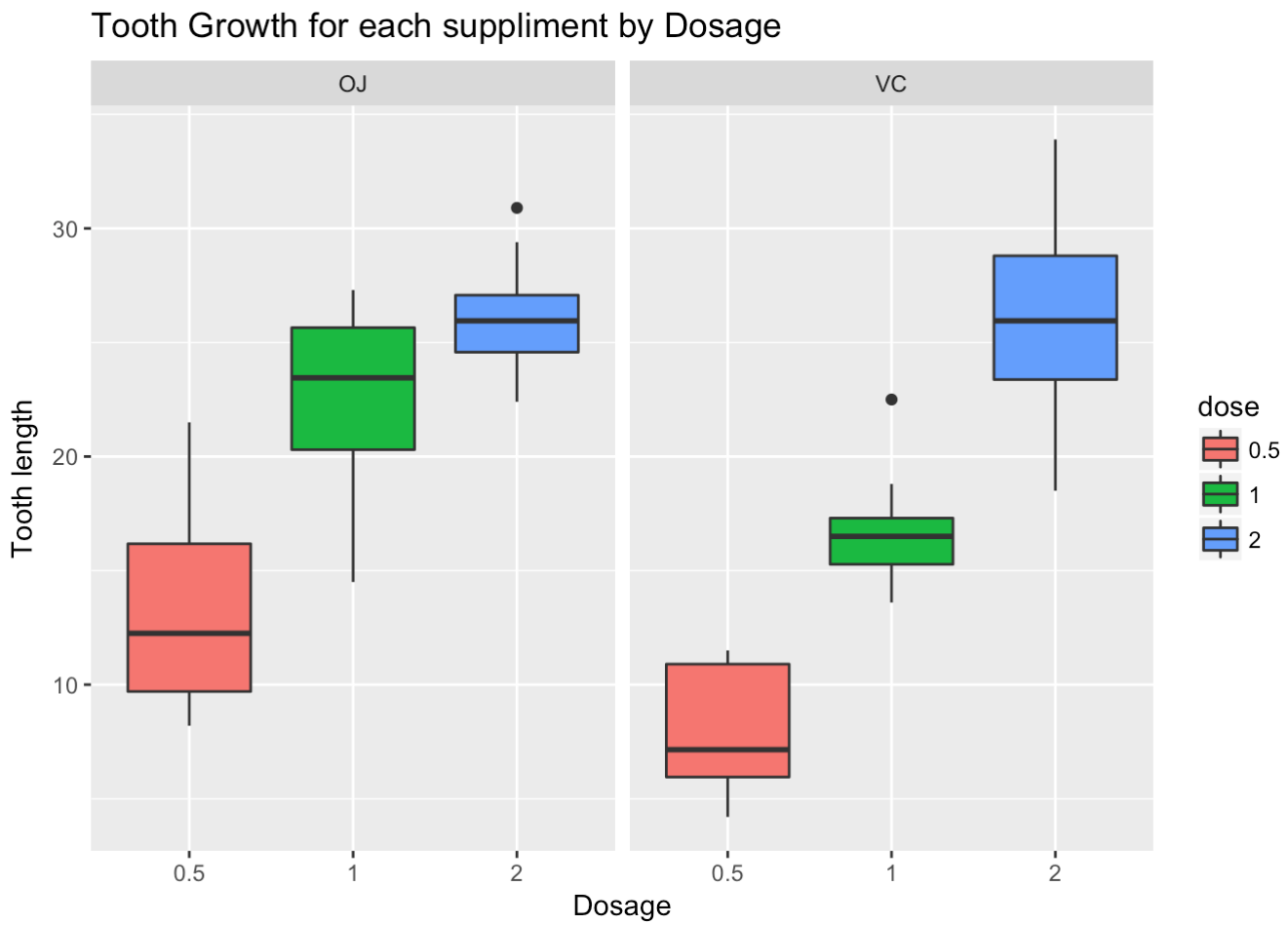
```
## [1] 0
```

```
sum(is.na(ToothGrowth$dose))
```

```
## [1] 0
```

We have 2 variables which impact tooth growth: suppliments and dosage.

```
tg <- ToothGrowth
tg$dose <- factor(tg$dose)
ggplot(data = tg, aes(x=dose, y=len)) +
    geom_boxplot(aes(fill = dose)) +
    facet_wrap(~ supp) +
    labs(title="Tooth Growth for each suppliment by Dosage",
         x="Dosage", y="Tooth length")
```



Interestingly, the mean of tooth growth is similar when the dosage is 2.0 mg/day for both suppliments. Orange Juice appears to be more beneficial toward tooth growth for dosage level of 0.5 mg/day and 1.0 mg/day.

Summary of data

```
dim(ToothGrowth)
```

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

```
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
```

```
sum(ToothGrowth$supp=="OJ")
```

```
## [1] 30
```

```
sum(ToothGrowth$supp=="VC")
```

```
## [1] 30
```

```
sum(ToothGrowth$dose==0.5)
```

```
## [1] 20
```

```
sum(ToothGrowth$dose==1)
```

```
## [1] 20
```

```
sum(ToothGrowth$dose==2)
```

```
## [1] 20
```

Confidence Intervals and/or hypothesis tests to compare tooth growth by supp and dose

We will be running the following t.test to examine the impact to tooth growth:
1. Suppliment 2. Dosage

Before we can begin the testing, we have to define the Null and alternative hypothesis:

- H0 - There is no impact between the test variables and tooth growth.
- H1 - There is impact between the test variables and tooth growth.

Alpha level = 0.05 is used to reject the null hypothesis.

Suppliment

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

```
##
##  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
```

The p-value is 0.061. We cannot exclude the Null hypotheis because it is above the alpha level.

Dosage

Since we have 3 dosages (0.5, 1, and 2), we have to break up the analysis into 3 tests.

Dosage - 0.5 and 1

```
t.test(tg$len[tg$dose==0.5], tg$len[tg$dose==1])
```

```
##
##  Welch Two Sample t-test
##
## data:  tg$len[tg$dose == 0.5] and tg$len[tg$dose == 1]
## 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:
##  -11.983781  -6.276219
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

The p-value is 1.268e-07 and is significantly less than the defined alpha level.

Dosage - 0.5 and 2

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

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

Similarly to the test above, the p-value (4.398e-14) is significantly less than the pre-defined alpha level.

Dosage - 1 and 2

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

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

The p-value is 1.906e-05. This value is significantly less than the pre-defined alpha level as well.

In the dosage tests, we found all the p-values to be below the alpha level.

Conclusions

Based on the analysis, we can conclude the following:
1. It is unclear which suppliment is better for tooth growth.
2. All amount of dosages do impact to the tooth growth. This is align with the finding during the basic analysis phase.
3. From our initial finding, on average both suppliments provide the same tooth growth benefit.