Part 2: Basic Inferential Data Analysis

Siyang Ni

1/6/2021

1. Overview

This is the final project for John Hopkins University's Statistical Inference Course, which consists of two parts:

1. A simulation exercise.
2. A basic inferential data analysis.

This is the second part of the project, in which a basic inferential data analysis will be conducted to explore certain aspects of the `ToothGrowth` dataset.

```
# Setting up the environment
library(tidyverse)
```

2. Exploratory Data Analysis

The dataset `ToothGrowth` contains 3 variables, each with 60 observations. The three variables are:

- `len`: [numeric] Tooth length
- `supp`:[factor] Supplement type (VC or OJ).
- `dose`: [numeric] Dose in milligrams/day

The response variable is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC).

```r
data("ToothGrowth")
df <- as.data.frame(ToothGrowth)

# Peek at the data
dim(df)
```

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

```r
str(df)
```

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

```r
head(df)
```

```
##    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
```

Below is a general summary of the three variables:

```r
summary(df)
```

```
##       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
```

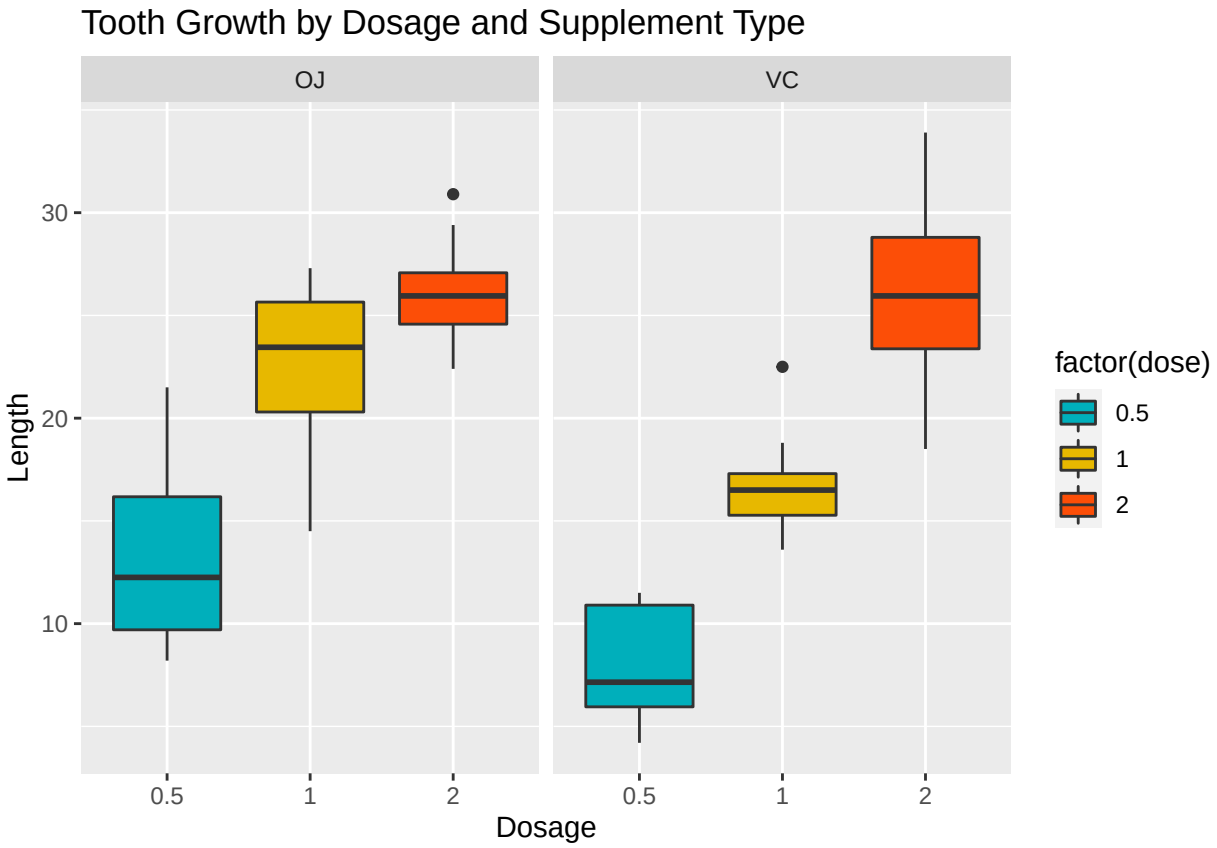
In this small project, **we are interested in the difference between tooth growth given different types of supplement and varied dosage**. A good point of departure would be side-by-side box plots that visualize the difference.

The visualization result illustrates:

1. Dosage is positively correlated to tooth growth for either type of supplement.

2. The two types of supplement have varied effect on tooth growth in terms of dosage. When the dosage is at 0.5mg/day, orange juice seems to be a more effective supplement, though the effect of orange juice displays larger variation than that of ascorbic acid's. When the dosage is at 1.0mg/day, orange juice is apparently the more effective supplement. However, when the dosage is 2.0 mg/day, the effectiveness of the two types of supplement seem to have no obvious distinction, while the variance of the effect of ascorbic acid is larger than that of the orange juice's.

```
g = ggplot(df, aes(factor(dose), len, fill=factor(dose)))
g = g + geom_boxplot() + facet_grid(.~supp)
g = g + labs(x = 'Dosage', y = 'Length', title = "Tooth Growth by Dosage and Supplement
g = g + scale_fill_manual(values = c("#00AFBB", "#E7B800", "#FC4E07"))
g
```



3. Hypothesis

From the above summary, we can draw the following hypotheses:

- Hypothesis 1: Generally speaking, orange juice differs from ascorbic acid in terms of the effect on tooth growth.
- Hypothesis 2: At 0.5 dosage level, orange juice is a more effective supplement for tooth growth.
- Hypothesis 3: At 1.0 dosage level, orange juice is a more effective supplement for tooth growth.
- Hypothesis 4: At 2.0 dosage level, the two types of supplement have no significant difference in effect on tooth growth.

4. Result

For the t-test of hypothesis 1, the p-value is larger than .05. Meanwhile, the confidence interval contains 0. Therefore, we fail to reject the null hypothesis. Consequently, **there is not enough evidence to support our first hypothesis**, which states that the two types of supplement differ in effect.

```
t1 <- t.test(len ~ supp, data = df)
round(t1$p.value, 3)
```

```
## [1] 0.061
```

```
round(t1$conf.int, 3)
```

```
## [1] -0.171  7.571
## attr(,"conf.level")
## [1] 0.95
```

For the t-test of hypothesis 2, the p-value is smaller than .05. Therefore, we can reject the null hypothesis with 95% confidence. Given the confidence interval is completely above zero, **the result of the t-test supports our hypothesis 2**, which states that at 0.5 dosage level, orange juice is a more effective supplement.

```
len2 <- df$len[df$dose==0.5]
supp2 <- df$supp[df$dose==0.5]

t2 <- t.test(len2 ~ supp2)
round(t2$p.value, 3)
```

```
## [1] 0.006
```

```r
round(t2$conf.int, 3)
```

```
## [1] 1.719 8.781
## attr(,"conf.level")
## [1] 0.95
```

For the t-test of hypothesis 3, the p-value is smaller than .01. Therefore, we can reject the null hypothesis with 99% confidence. Given the confidence interval is completely above zero, **the result of the t-test supports our hypothesis 3**, which states that at 1.0 dosage level, orange juice is a more effective supplement.

```r
len3 <- df$len[df$dose==1.0]
supp3 <- df$supp[df$dose==1.0]

t3 <- t.test(len3 ~ supp3)
round(t3$p.value, 3)
```

```
## [1] 0.001
```

```r
round(t3$conf.int, 3)
```

```
## [1] 2.802 9.058
## attr(,"conf.level")
## [1] 0.95
```

For the t-test of hypothesis 4, the p-value is much larger than .05. Meanwhile, the confidence interval contains 0. Therefore, we fail to reject the null hypothesis. Therefore, **the result of our t-test supports hypothesis 4**, which states that at 2.0 dosage level, the two types of supplement have no significant difference in effect.

```r
len4 <- df$len[df$dose==2.0]
supp4 <- df$supp[df$dose==2.0]

t4 <- t.test(len4 ~ supp4)
round(t4$p.value, 3)
```

```
## [1] 0.964
```

```r
round(t4$conf.int, 3)
```

```
## [1] -3.798  3.638
## attr(,"conf.level")
## [1] 0.95
```