Inferential Data Analysis

Charles

August 7, 2017

Overview

In this report we'll analyze the ToothGrowth data in the R datasets package and have a basic Inferential Data Analysis.

Basic exploratory data analysis

Firstly we set up the environment:

```
setwd("C:/Study/Coursera/1 Data-Science/2 RStudio/6 Class 6/Coursera_DataScience_Class6_FinalProject")
set.seed(135246987)
library(ggplot2)
library(gridExtra)
library(grid)
library(datasets)
```

Then we load the data and summarize the data:

```
data("ToothGrowth")
dataTG <- ToothGrowth
str(dataTG)
```

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

```
head(dataTG)
```

```
##    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
```

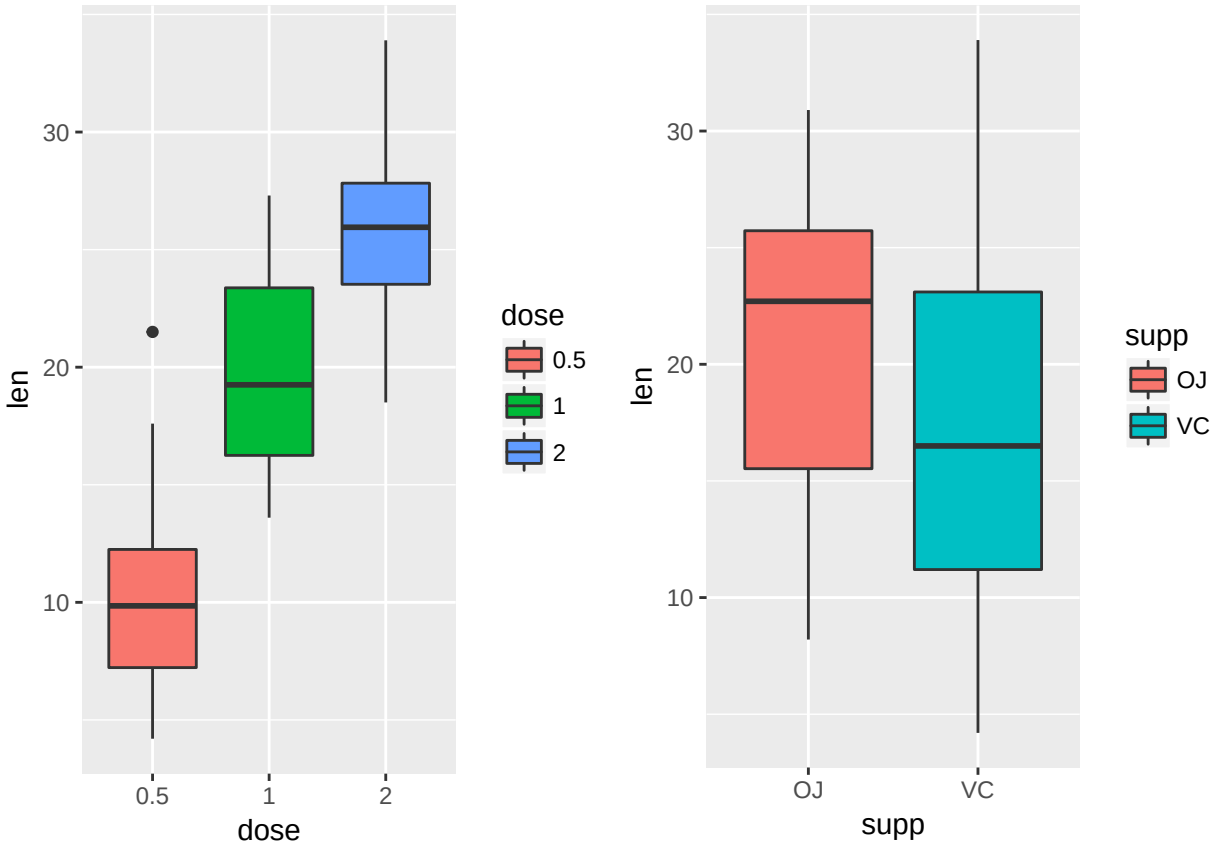
```
summary(dataTG)
```

```
##       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
```

```
dataTG$dose <- as.factor(dataTG$dose)
```

We can get a general idea of the data according to explore above. Then we can do some exploratory data analysis by drawing relationships between length and supp & dose.

```
g1 <- ggplot(data = dataTG, aes(x=dose, y=len)) +
        geom_boxplot(aes(fill = dose))
g2 <- ggplot(data = dataTG, aes(x=supp ,y=len)) +
        geom_boxplot(aes(fill = supp))
grid.arrange(g1,g2,nrow=1)
```



As we can see above, it seems the length grows as dose increases. While the replationship between length and supp is not quite obvious.

T test for length and supp

We conduct t-test for length and supp:

```
t.test(len ~ supp, data = dataTG, alternative ="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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 in group OJ mean in group VC
##         20.66333         16.96333
```

As we can see above, 95% confidence interval contains zero and p-value is about 0.06. So we wouldn't reject the null hypothesis, which means true difference in means is equal to 0.

T test for length and dose.

We divide the data into three groups: dose in 0.5 and 1; dose in 0.5 and 2; dose in 1 and 2.

```
dataTG_0.5_1.0 <- subset(dataTG, dose %in% c("0.5","1"))
dataTG_0.5_2.0 <- subset(dataTG, dose %in% c("0.5","2"))
dataTG_1.0_2.0 <- subset(dataTG, dose %in% c("1","2"))
```

Then do the t-test:

```
t.test(len ~ dose, data = dataTG_0.5_1.0)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## 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 in group 0.5   mean in group 1
##            10.605            19.735
```

```
t.test(len ~ dose, data = dataTG_0.5_2.0)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## 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 in group 0.5   mean in group 2
##            10.605            26.100
```

```
t.test(len ~ dose, data = dataTG_1.0_2.0)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## 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 in group 1 mean in group 2
##          19.735          26.100
```

As we can see above, all three test the p-value is smaller than 5% and the 95% confidence interval dose not contain zero. Besides, they are all negative, which means there's an increase as dose goes up.

Conclusions and assumptions

According to the analysis above, we could see that under 95% confidence interval:

- Supplement has no effect on the length of tooth growth.
- Tooth grows longer as the dose increases.

Assumptions:

- The experiment was properly conducted.
- The sample data is good enough to represent the entire population.