Part 2: Basic Inferential Data Analysis Instructions

Yan He

2018/8/18

Overview

Now we want to analysis the ToothGrowth datasets in R which response is the length of odontoblasts 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).

First we load the data and summarise the original data:

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
## Using describe() function in Hmisc to get descriptive statistics

data(ToothGrowth)

describe(ToothGrowth)
```

```
## ToothGrowth
##
##  3  Variables      60  Observations
## ---------------------------------------------------------------------------
## len
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       60        0       43    0.999    18.81    8.839     6.37     8.11
##      .25      .50      .75      .90      .95
##    13.07    19.25    25.27    27.30    29.57
##
## lowest :  4.2  5.2  5.8  6.4  7.0, highest: 29.4 29.5 30.9 32.5 33.9
## ---------------------------------------------------------------------------
## supp
##        n  missing distinct
##       60        0        2
##
## Value       OJ  VC
## Frequency   30  30
## Proportion 0.5 0.5
## ---------------------------------------------------------------------------
## dose
##        n  missing distinct     Info     Mean      Gmd
##       60        0        3    0.889    1.167    0.678
##
## Value        0.5   1.0   2.0
## Frequency     20    20    20
## Proportion 0.333 0.333 0.333
## ---------------------------------------------------------------------------
```

It's obvious that the dataseta has three variables: tooth length, supplement type and dose. Because dose variable has only three unique values, we choose to change it to factor variable:

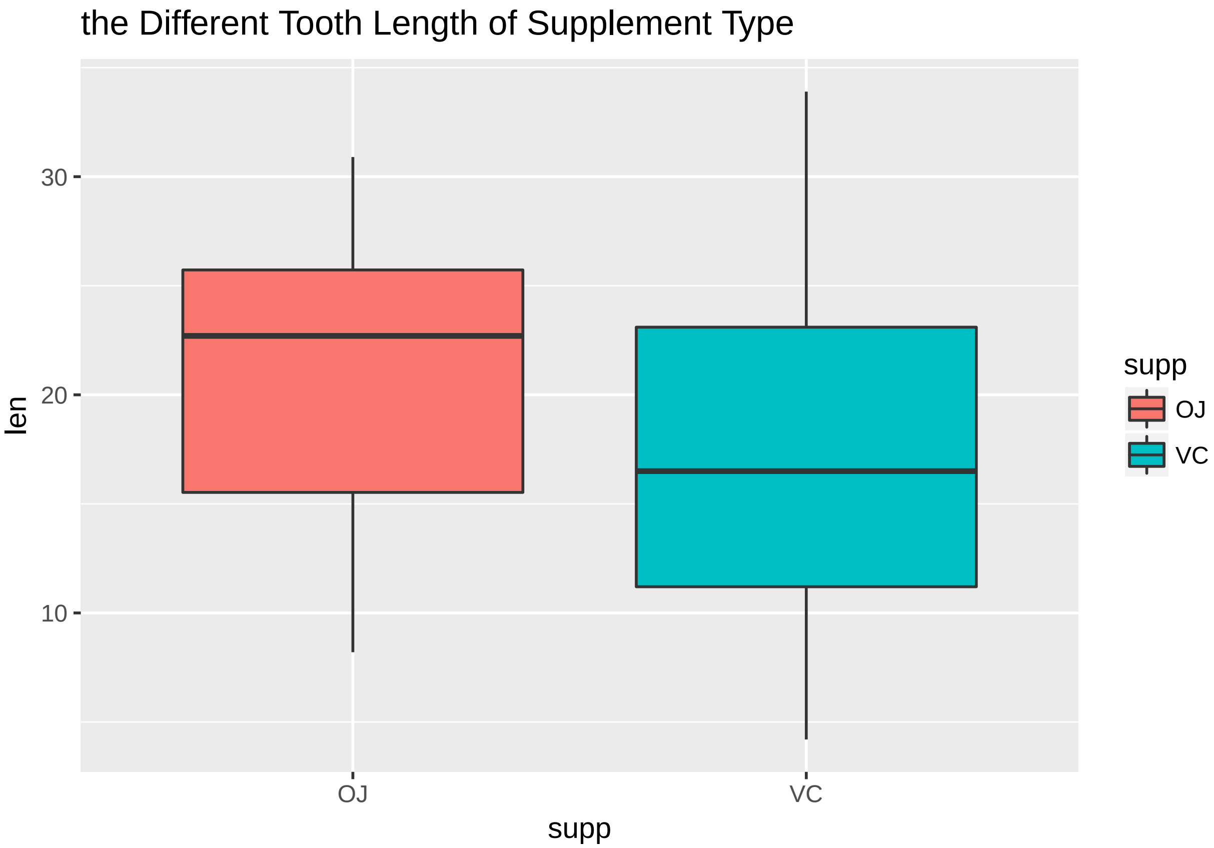
```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
class(ToothGrowth$dose)
```

```
## [1] "factor"
```

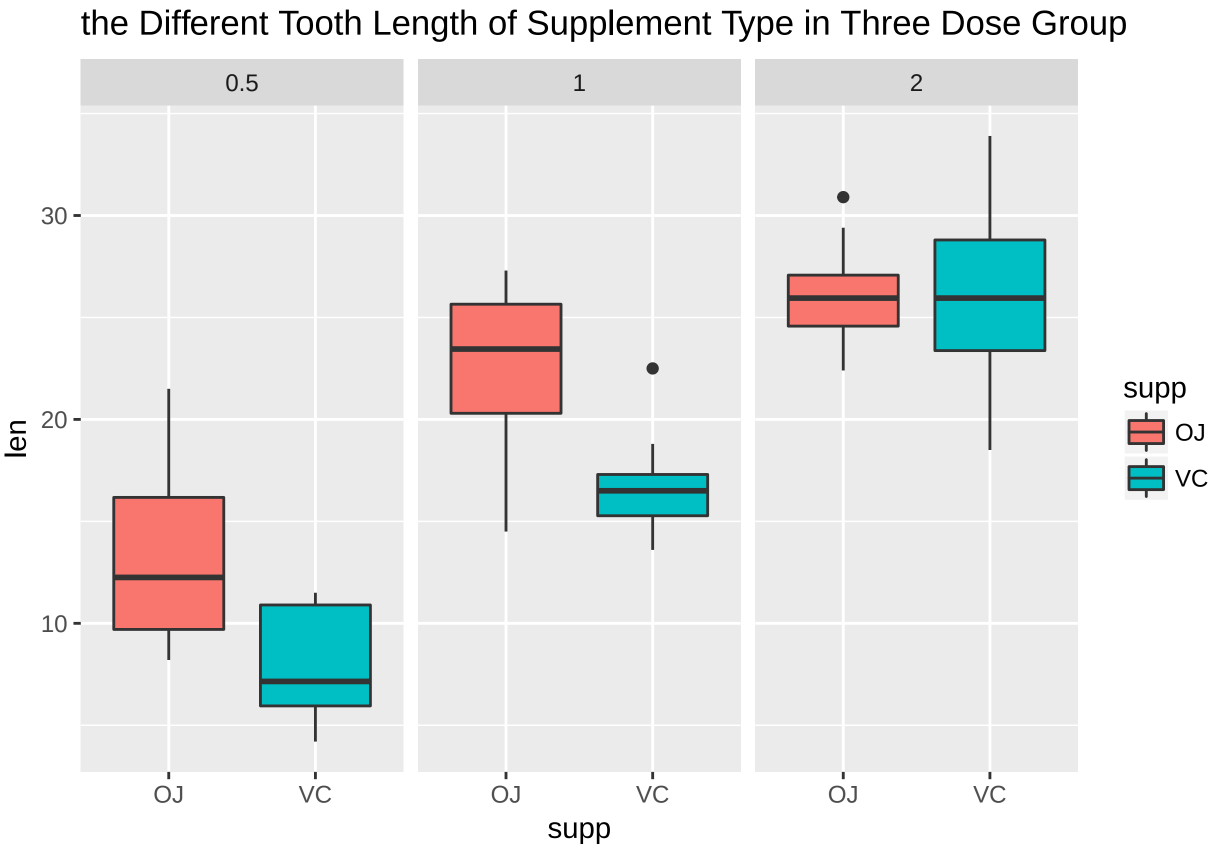
Now we can use ggplot package to get some useful plots:

```
library(ggplot2)

p <- ggplot(data = ToothGrowth, aes(x = supp, y = len, fill = supp)) +
        geom_boxplot() +
        ggtitle('the Different Tooth Length of Supplement Type')
p
```

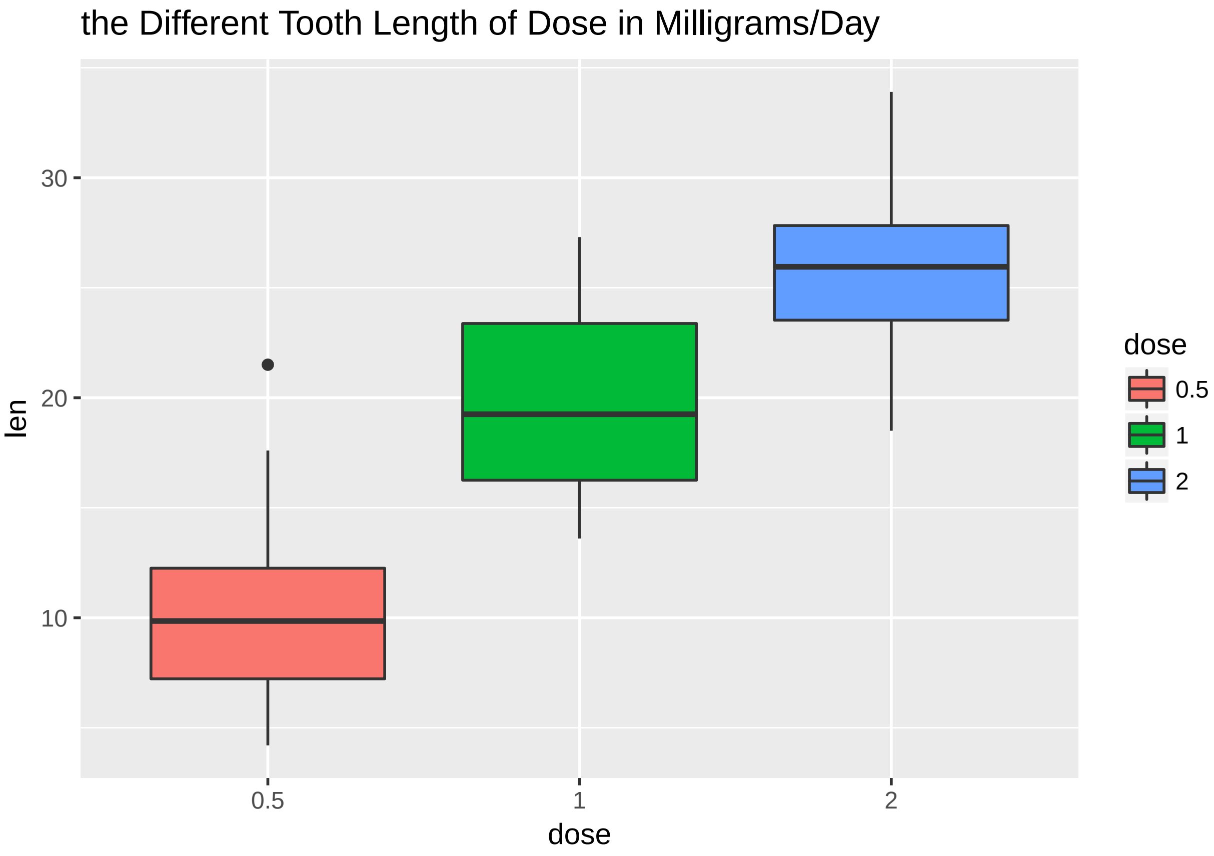


```
q <- ggplot(data = ToothGrowth, aes(x = supp, y = len, fill = supp))+
        geom_boxplot() +
        facet_grid(.~dose) +
        ggtitle('the Different Tooth Length of Supplement Type in Three Dose Group')
q
```



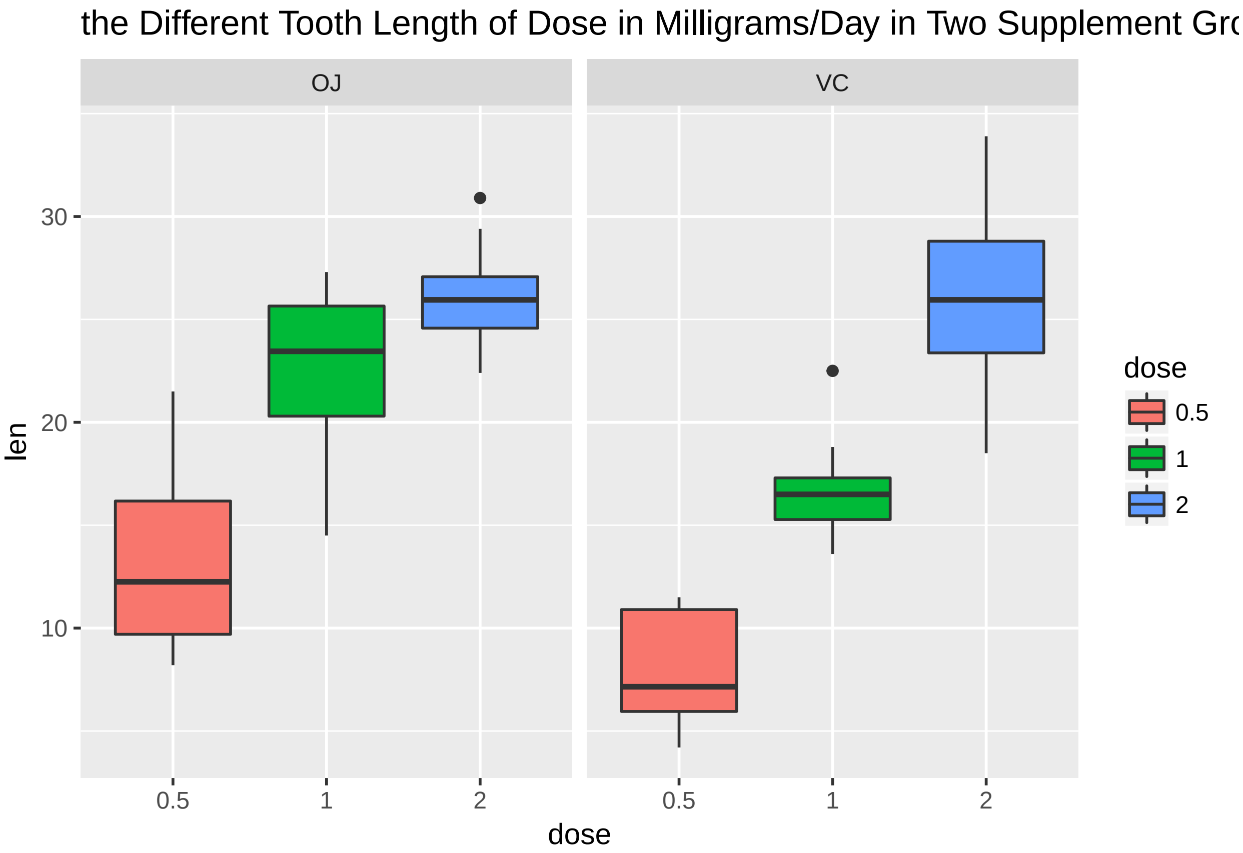
The plot shows that supplement types has different effect on tooth growth. OJ has higher effect than VC.

```
p <- ggplot(data = ToothGrowth, aes(x = dose, y = len, fill = dose)) +
        geom_boxplot() +
        ggtitle('the Different Tooth Length of Dose in Milligrams/Day')
p
```



```r
q <- ggplot(data = ToothGrowth, aes(x = dose, y = len, fill = dose)) +
        geom_boxplot() +
        facet_grid(.~supp) +
        ggtitle('the Different Tooth Length of Dose in Milligrams/Day in Two Supplement Group')

q
```



The plot shows that different doses in milligrams/day have different influnce on tooth growth.

Hypothesis Test

Now we will use hypothesis test for our guesses. The samples of each group is less than 30 which means it's will be a small sample test, so we will use T-test rather than Z-test for our hypothesis tests.

Test for Supplement

```
##
##  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
```

From the T-test we can see that p value is greater than 0.05 (which means no significant) so we can't reject the null hypothesis that the two supp groups have no difference.

Test for dose

The dose variable has three levels so the traditional two sample t-test is not suitable for our job. However, we can develope three methods to solve this problem.

1.using t-test for two of three levels and repeat; 2.using ANOVA method to analysis the paired data; 3.using nonparametric test;

t-test

dose 0.5 and 1

```
sub_data_1 <- subset(ToothGrowth, dose %in% c(0.5, 1))
dose_1 <- sub_data_1$dose
len <- sub_data_1$len
t.test(len~dose_1, alternative = 'two.sided', paired = FALSE)
```

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

The p-value is smaller than 0.05 so we can reject the null hypothesis and make sure that dose 0.5 ans 1 group has significant difference.

dose 1 and 2

```
sub_data_2 <- subset(ToothGrowth, dose %in% c(1,2))
dose_2 <- sub_data_2$dose
len <- sub_data_2$len
t.test(len~dose_2, alternative = 'two.sided', paired = FALSE)
```

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

dose 0.5 and 2

```
sub_data_3 <- subset(ToothGrowth, dose %in% c(0.5, 2))
dose_3 <- sub_data_3$dose
len <- sub_data_3$len
t.test(len~dose_3, alternative = 'two.sided', paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose_3
## 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
```

From above tests we can find that three p-value is smaller than 0.05 so we can reject the null hypothesis and make sure that every dose group has significant difference.

nonparamtetric test

Now we use Kruskal-Wallis test for our propose to analysis the difference of three dose groups.

```
kruskal.test(len~dose, data = ToothGrowth)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  len by dose
## Kruskal-Wallis chi-squared = 40.669, df = 2, p-value = 1.475e-09
```

The p value is so small that we can reject the null hypothesis. The result is same with t-test.

Conclusions

From our analysis, we can make two conclusions of the ToothGrowth dataset. First, there is significant difference of teeth length and dose levels. Second, althrough we offer two supplement types, it's no significant difference of teeth length and supp.