Basic Inferential Data Analysis

Devara Izaz Fathan

8/22/2020

1. Overview

In this assignment we're going to analyze the ToothGrowth data in the R datasets package. We will do some hypothesis tests to get the relationship between tooth length, dose, and supplement.

2. Load Libraries and Data

```
##Load the library that will be used
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------------

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v stringr 1.4.0
## v tidyr   1.1.1     v forcats 0.5.0
## v readr   1.3.1

## -- Conflicts ----------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
##LOad the data
data("ToothGrowth")
```

3. Summarizing Data

Look at the structure of the 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 ...
```

The class of the data set is data frame which contain 60 observation of 3 variables. Now we can look at the summary of the ToothGrowth 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
```

This dataset records effect on length of the tooth (len) following treatment with a supplement (supp) and it's dose (dose). The mean tooth length is 18.81 but it varies from 4.2 to a max. of 33.9. Doses vary from 0.5 to 2.0 with a mean of 1.167.

Now, we take a look on summary length that grouped by supp

```
ToothGrowth %>% group_by(supp) %>% summarise(length = mean(len))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 2
##   supp  length
##   <fct>  <dbl>
## 1 OJ      20.7
## 2 VC      17.0
```

Then taka a look on summary length that grouped by dose, but als first grouped by supp.

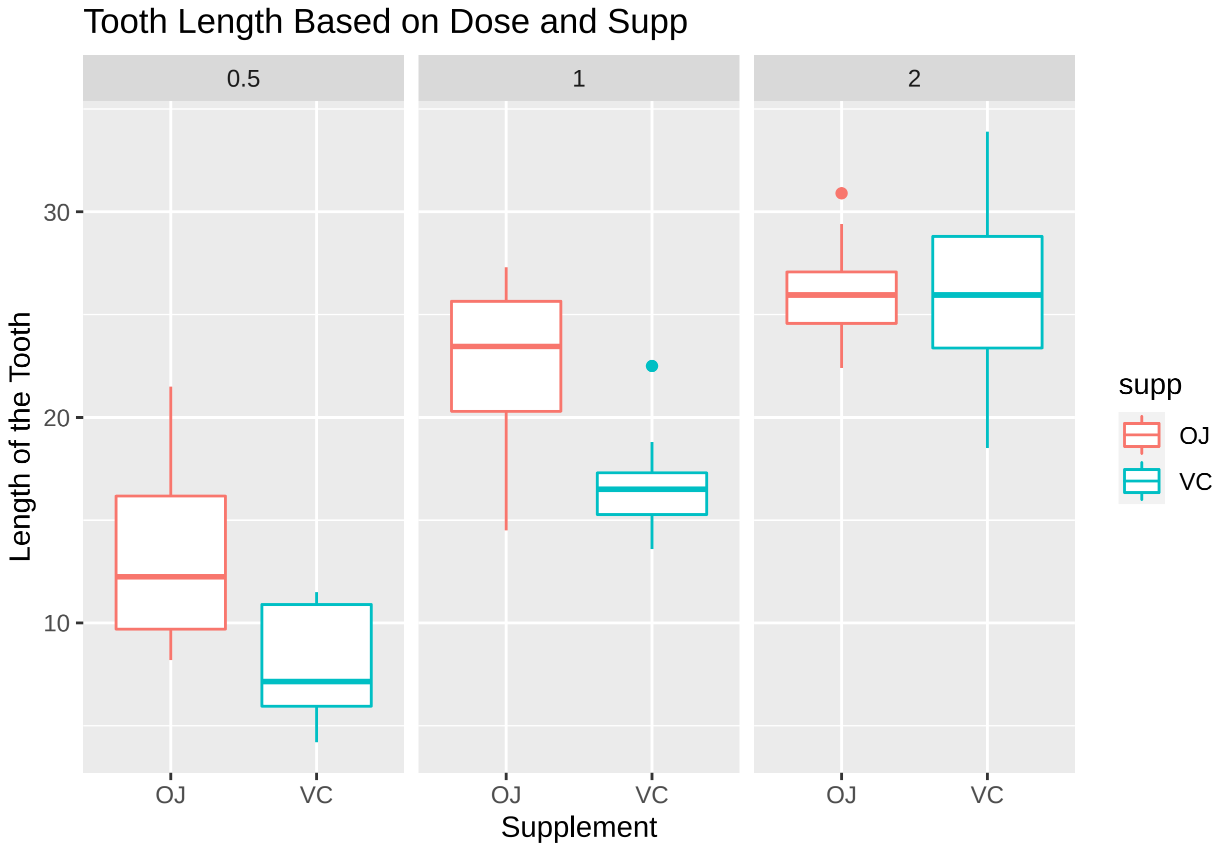
```
ToothGrowth %>% group_by(supp, dose) %>% summarise(length = mean(len))
```

```
## `summarise()` regrouping output by 'supp' (override with `.groups` argument)
```

```
## # A tibble: 6 x 3
## # Groups:   supp [2]
##   supp   dose length
##   <fct> <dbl>  <dbl>
## 1 OJ      0.5  13.2 
## 2 OJ      1    22.7 
## 3 OJ      2    26.1 
## 4 VC      0.5   7.98
## 5 VC      1    16.8 
## 6 VC      2    26.1
```

For making clear the distribution of the data that grouped by dose, we can make a boxplot

```
ToothGrowth %>% ggplot(aes(x=supp, y=len, col=supp)) +
      geom_boxplot() +
      facet_grid(.~dose) +
      ggtitle('Tooth Length Based on Dose and Supp') +
      xlab('Supplement') +
      ylab('Length of the Tooth')
```



From the plot above we can conclude that if we give the dose higher than the tooth are more length. And for the type of the supplement, the OJ give higher effect to the length of the Tooth, but if we look at dose=2, the type of the supplement doesn't give much differences of the tooth length.

4. Confidence Interval and/or Hypothesis Test

We use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

4.1 Hypothesis Test Based on Supplement

We have to construct null hypothesis and alternative hypothesis. Null Hypothesis : There is no relationship between Length of the tooth and The type of the supplement(OJ-VC=0) Alternative Hypothesis : There relationship between Length of the tooth and The type of the supplement(0J-VC $>$ 0) On this hypothesis test we use 0.05 significance level

We should assign variables for length tooth of OJ and VC from the dataset.

```
OJ <- ToothGrowth %>% filter(supp=='OJ') %>% select(len)
VC <- ToothGrowth %>% filter(supp=='VC') %>% select(len)
```

Then we can do the T test for unpaired and unequal variance(our assumption). We take this assumption because there is no direct relationship between OJ and VC.

```
t.test(OJ, VC, alternative = 'greater' ,paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ and VC
## t = 1.9153, df = 55.309, p-value = 0.03032
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4682687       Inf
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

The result of the test show that the p-value is less then 0.05, so that we reject the null hypothesis. Based on the confidence interval, we can say that P(OJ-VC=0) is not in the 95% confidence interval.

However, if we look at the boxplot, we have to give more attention to the dose=2, so we have to test it. We have to assign variables that we are going to use in this test.

```
dose2_OJ <- ToothGrowth %>% filter(dose==2, supp==’OJ’) %>% select(len)
dose2_VC <- ToothGrowth %>% filter(dose==2, supp==’VC’) %>% select(len)
```

Then we can do the T test for unpaired and unequal variance(our assumption). We take this assumption because there is no direct relationship between OJ and VC.

```
t.test(dose2_OJ, dose2_VC, alternative = ’greater’ ,paired = FALSE, var.equal = FALSE, conf.level = 0.95
```

```
##
##  Welch Two Sample t-test
##
## data:  dose2_OJ and dose2_VC
## t = -0.046136, df = 14.04, p-value = 0.5181
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -3.1335     Inf
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

The result of the test show that the p-value=0.5181 is far more than 0.05, so that we can not reject the null hypothesis. Based on the confidence interval, we can say that P(OJ-VC=0) is in the 95% confidence interval

4.2 Hypothesis Test Based on Dose

In the ToothGrowth data we have 3 kinds of dose, so we have to make three test: 1. Test for dose=0.5 and dose=1 2. Test for dose=0.5 and dose=2 3. Test for dose=1 and dose=2

First we have to assign variables for each dose.

```
dose_half <- ToothGrowth %>% filter(dose==0.5) %>% select(len)
dose_one <- ToothGrowth %>% filter(dose==1) %>% select(len)
dose_two <- ToothGrowth %>% filter(dose==2) %>% select(len)
```

Then we can do the T test for unpaired and unequal variance(our assumption). We take this assumption because there is no direct relationship between OJ and VC based on the boxplot that we have been created.

4.2.1 Test for Dose=0.5 and Dose=1

```
t.test(dose_half, dose_one, alternative = ’less’, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  dose_half and dose_one
## t = -6.4766, df = 37.986, p-value = 6.342e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.753323
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

The result of the test show that the p-value=6.342e-08 is far less than 0.05, so that we can reject the null hypothesis. Based on the confidence interval, we can say that P(OJ-VC=0) is not in the 95% confidence interval.

4.2.2 Test for Dose=0.5 and Dose=2

```
t.test(dose_half, dose_two, alternative = ’less’, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  dose_half and dose_two
## t = -11.799, df = 36.883, p-value = 2.199e-14
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -13.27926
## sample estimates:
## mean of x mean of y
##    10.605    26.100
```

The result of the test show that the p-value=2.199e-14 is far less than 0.05, so that we can reject the null hypothesis. Based on the confidence interval, we can say that P(OJ-VC=0) is not in the 95% confidence interval.

4.2.3 Test for Dose=1 and Dose=2

```
t.test(dose_one, dose_two, alternative = ’less’, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  dose_one and dose_two
## t = -4.9005, df = 37.101, p-value = 9.532e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -4.17387
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

The result of the test show that the p-value= 9.532e-06 is far less than 0.05, so that we can reject the null hypothesis. Based on the confidence interval, we can say that P(OJ-VC=0) is not in the 95% confidence interval.

5. Conclusion

1. Based on the test that we have done, in general p-value are less than 0.05 so that we can reject both null hypothesises. However, for test in dose=2, the p-value is far more than 0.05.
2. Tooth length have relationship to the type of the supplement, generally supplement OJ have larger association to Tooth length than VC supplement. However, for dose=2, both VC and OJ approximately having same effect to the tooth length
3. Tooth length have relationship to the amount of dose, the higher the dose, then the longer the tooth.
4. We assume that the sample population is representative of the population, that treatments were randomly assigned and the variances are different.