Statistical Inference Project Part II

R3M79

27 de Dezembro de 2017

Synopsis

This document pertains to Cousera's Statistical Inference model Project. The project is divided in two parts

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

In this document we'll address part 2 of the project

Part 2: Basic Inferential Data Analysis

Overview

Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package.

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.

Preparation

First we'll load the required libraries and data

Data Detail and Summary

```
#Preview of data
head(ToothGrowth,15)
```

```
##     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
## 7  11.2   VC  0.5
## 8  11.2   VC  0.5
## 9   5.2   VC  0.5
## 10  7.0   VC  0.5
```

```
## 11 16.5   VC  1.0
## 12 16.5   VC  1.0
## 13 15.2   VC  1.0
## 14 17.3   VC  1.0
## 15 22.5   VC  1.0
```

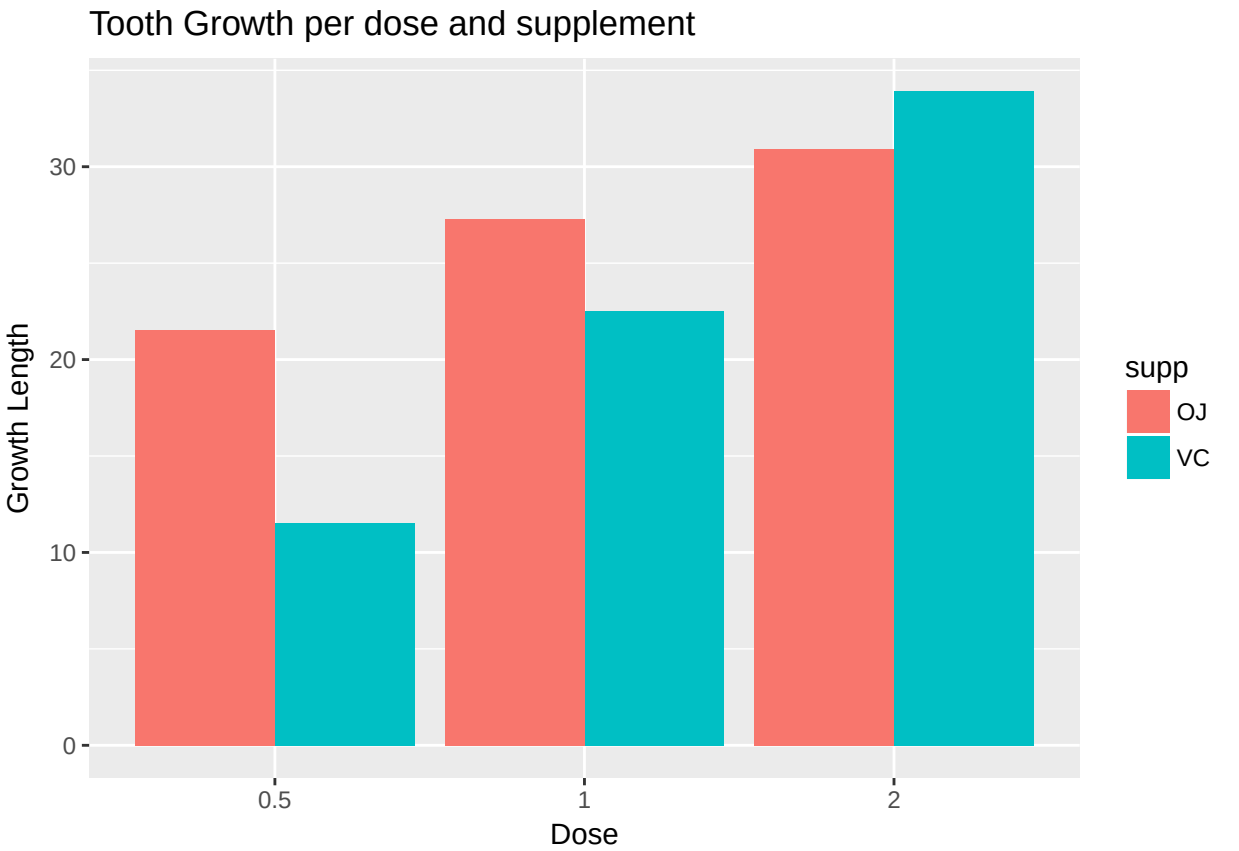
```
#Detail of data
str(ToothGrowth)
```

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

```
#Data Summary
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
```

```
#plot data
plotdata
```



From the above output and plot we can now see how the data is organized. We can see that there's no value

for a dose of 1.5

Hypotesis testing

Growth by Supplement

Let's perform a t test to compare Growth by Supplement

```
cat("Conf Int:",t_gr_sup$conf.int,"p-value",t_gr_sup$p.value)
```

```
## Conf Int: -0.1710156 7.571016 p-value 0.06063451
```

The p-value of this test was 0.06, which is greater than 0.05, and the confidence interval of the test contains zero. With these results we can't reject the null hypothesis that the different supplement types don't have effect on tooth growth length.

Growth by Doses

Now we'll compare tooth growth by dose, testing the different pairs of dose values. (The data doesn't posess any value for doses 1.5)

```
#"Doses 1.0 and 2.0"
cat("Conf Int:",t_gr_dose_1_2$conf.int,"p-value",t_gr_dose_1_2$p.value)
```

```
## Conf Int: -8.996481 -3.733519 p-value 1.90643e-05
```

```
#"Doses 0.5 and 1.0"
cat("Conf Int:",t_gr_dose_5_1$conf.int,
    "p-value",t_gr_dose_5_1$p.value)
```

```
## Conf Int: -11.98378 -6.276219 p-value 1.268301e-07
```

```
#"Doses 0.5 and 2.0"
cat("Conf Int:",t_gr_dose_5_2$conf.int,
    "p-value",t_gr_dose_5_2$p.value)
```

```
## Conf Int: -18.15617 -12.83383 p-value 4.397525e-14
```

From the above results we see that the p-value is always very close to zero. The confidence intervals for each test don't cross the value zero.

Based on these results we can assume that the average tooth length increases with an inceasing dose, and therefore the null hypothesis can be rejected.

Conclusion

In our analysis we considered the following assumptions:

1. The sample is representative of the population
2. The distribution of the sample means follows the Central Limit Theorem

Based on the presented results on the previous chapters we can conclude that the supplement has no effect on tooth growth length. However, as the dosage is increased a growth of the tooth will occur.

Appendix

Code

Below follows all the code necessary for the displayed information and plots.

```
#load libraries
library(dplyr)
library(ggplot2)

#Define Variables for simulation
set.seed(100) # set the seed value for reproducibility

#Load Data
data("ToothGrowth")

#plot data preparation
plotdata <- ggplot(ToothGrowth, aes(as.factor(dose),len,fill = supp))
plotdata <- plotdata + geom_bar(stat = "identity", position = "dodge") +
    labs(title = "Tooth Growth per dose and supplement",
         x = "Dose", y = "Growth Length")


#Preview of data
head(ToothGrowth,15)

#Detail of data
str(ToothGrowth)

#Data Summary
summary(ToothGrowth)

#plot data
plotdata

#Compare tooth growth by supplement using a t-test.
t_gr_sup <- t.test(len~supp,data=ToothGrowth)
cat("Conf Int:",t_gr_sup$conf.int,
    "p-value",t_gr_sup$p.value)

#Compare tooth growth by supplement using a t-test.
# doses 1 and 2
t_gr_dose_1_2 <- t.test(len~dose,data=subset(ToothGrowth,
                                             dose %in% c(1, 2)))
cat("Conf Int:",t_gr_dose_1_2$conf.int,
    "p-value",t_gr_dose_1_2$p.value)

#Compare tooth growth by supplement using a t-test.
# doses .5 and 1
t_gr_dose_5_1 <- t.test(len~dose,data=subset(ToothGrowth,
                                             dose %in% c(.5, 1)))
cat("Conf Int:",t_gr_dose_5_1$conf.int,
    "p-value",t_gr_dose_5_1$p.value)
```

```r
#Compare tooth growth by supplement using a t-test.
# doses .5 and 2
t_gr_dose_5_2 <- t.test(len~dose,data=subset(ToothGrowth,
                                             dose %in% c(0.5, 2)))
cat("Conf Int:",t_gr_dose_5_2$conf.int,
    "p-value",t_gr_dose_5_2$p.value)
```