Statistical Inference: Toothgrowth experiment

Sanjay Somraj

March 23, 2017

Statistical Inference Course Project - PART 2

Overview

Load the ToothGrowth data and perform some basic exploratory data analyses
a) Provide a basic summary of the data.
b) 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)
c) State your conclusions and the assumptions needed for your conclusions.

Load libraries

```
# Load libraries for data processing and plotting
library(ggplot2)
library(datasets)
library(gridExtra)
library(grid)

# The Effect of Vitamin C on Tooth Growth in Guinea Pigs
data(ToothGrowth)
```

Basic Summary of the data

```
# Structure of the data frame
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 ToothGrowth data set consists of 60 observations of 3 variables:

- len: Tooth length in millimeters
- supp: Supplement type
- dose: Dose in milligrams

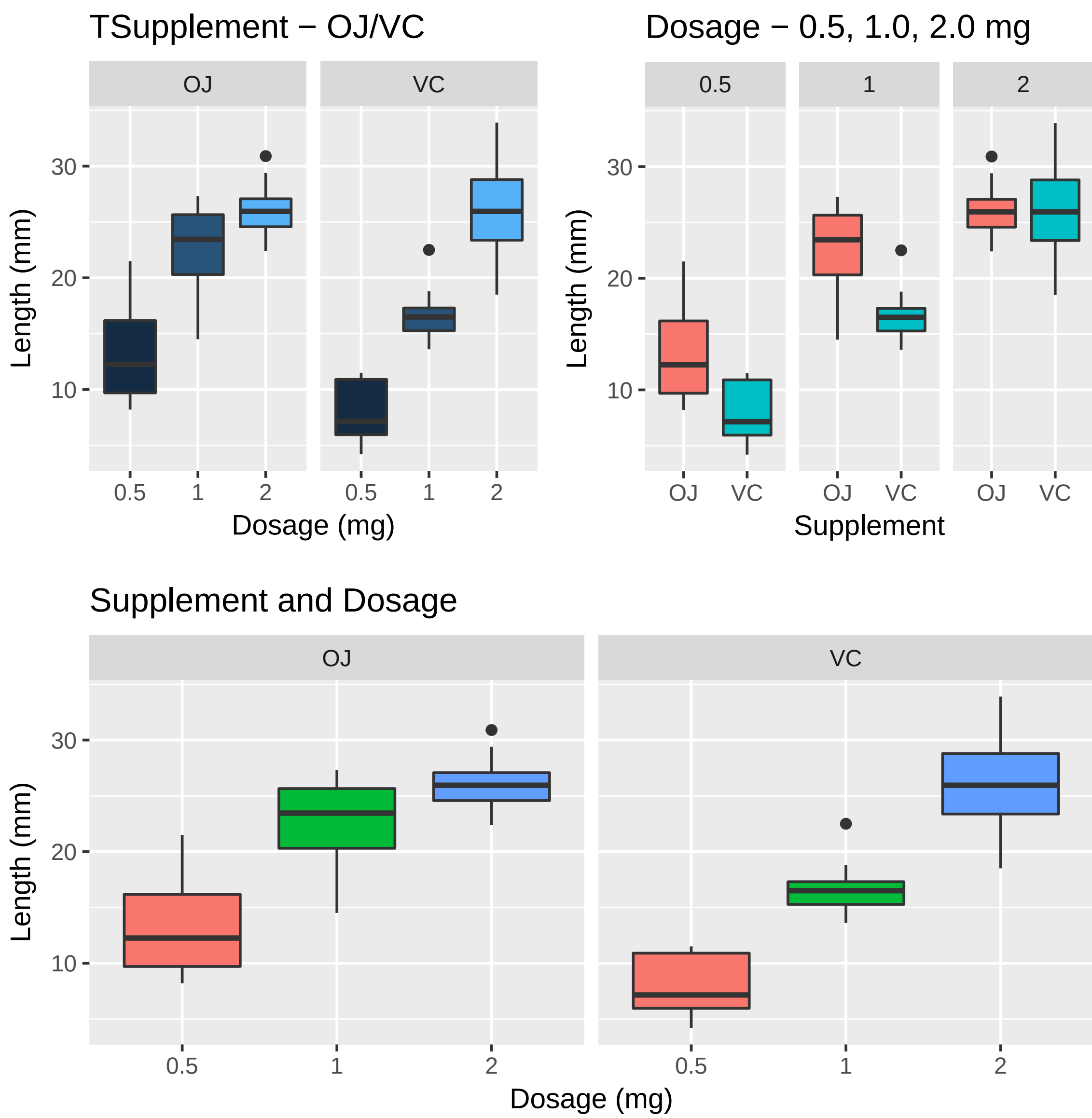
The structure of the Toothgrowth dataframe shows that SUPPLEMENT (supp) has 2 levels -
* OJ - Orange Juice
* VC - Vitamin C

Similary, the DOSAGE (dose) has 3 levels - * 0.5 mg * 1.0 mg * 2.0 mg

The average tooth length is **18.813** with a standard deviation of **7.649**

We focus on finding differences in tooth length across different groups of supplement types, dose levels and their respective combinations.

The below plot shows the tooth growth for each supplement types, Dosage and both (Dosage and Supplement)



Inference from Plot 1: The first box plot (LEFT) reveals that average tooth growth with Orange juice is greater than the average tooth growth than those which got their dose using Ascorbic acid (a form Vitamin C).

Inference from Plot 2: The second box plot (RIGHT) also gives a similar inference that the average tooth growth with Orange juice is greater than the average tooth growth than those which got their dose using Ascorbic acid (a form of Vitamin C). This plot also shows that the diferences in tooth growth greater and the average length.

Inference from Plot 3: from the third plot (BOTTOM), the Orange Juice group distribution is skewed to the left whereas the Ascorbic acid group is fairly symmetric.

Hypothesis Tests

We will use the t distribution for our hypothesis tests and when constructing confidence intervals.

We will have perform the t-tests to check if the p-value > 0.05 and if the confidence interval spans through 0.

First, we will perform the t-test based on **Supplement** Next, we will perform the t-tests for each pair-wise combination of the dosage levels

Assumption: We assume that the subjects were randomly assigned to the groups and that they were sampled from a normal population.

Supplement Type

NULL Hypothesis: We will consider that the average difference in tooth length is 0, implying that the supplement type does not impact the tooth growth.

We will first conduct t-test with unequal variances to check if the difference in average tooth length between subjects who received their dose using Orange juice and those who received their dose using Ascorbic acid is statistically different from 0.

```
ttest_supp <- t.test(len ~ supp, ToothGrowth, var.equal = FALSE)
```

Following the values from the t-test,
- **p Value** = 0.061
- **confidence intervals:** [-0.1710156, 7.5710156]

The **p Value** = **0.061** is greater than the significance level 0f 0.05 AND the confidence intervals contain 0.

We fail to reject the Null hypothesis.

Dosage

We will have to perform t-test for every value of the dosage administered (0.5, 1.0 and 2.0 mg) in three 3 pairwise comparisons (0.5, 1.0), (1.0,2.0) and (0.5,2.0) to determine if the dosage values impact the tooth growth.

NULL Hypothesis: We will consider that the average difference in tooth length is 0, implying that the Dosage does not impact the tooth growth.

```
##             headinglabels         test1         test2         test3
## 1                 p Value  1.266297e-07  1.810829e-05  2.837553e-14
## 2  Confidence Interval-Low -1.198000e+01 -8.990000e+00 -1.815000e+01
## 3 Confidence Interval-High -6.280000e+00 -3.740000e+00 -1.284000e+01
```

In each of the t-tests, we can see that the p value is very small, and less than the significance level 0.5. Also the confidence intervals do not include 0.

We can reject the Null hypothesis.

Conclusion

From the given sample data, we can infer that-
* The Supplement type, Orange juice of Ascorbic acid, does not impact the tooth growth
* The dosage levels does impact the tooth growth

Appendix (Source code)

Please find below the complete source code below.

```r
# Load libraries for data processing and plotting
library(ggplot2)
library(datasets)
library(gridExtra)
library(grid)

# The Effect of Vitamin C on Tooth Growth in Guinea Pigs
data(ToothGrowth)

str(ToothGrowth)

# The below plot shows the tooth growth for
#    1) Each supplement types,
#    2) Dosage
#    3) Both (Dosage and Supplement)

p1 <- ggplot(data=ToothGrowth, aes(x=factor(dose),y=len,fill=dose)) +
     geom_boxplot() +
     theme(legend.position="none") +
     facet_grid(.~supp) +
     labs(title="TSupplement - OJ/VC",
          x = "Dosage (mg)",
          y = "Length (mm)")

p2 <- ggplot(data=ToothGrowth, aes(x=supp,y=len,fill=supp)) +
     geom_boxplot() +
     theme(legend.position="none") +
     facet_grid(.~dose)+
     labs(title="Dosage - 0.5, 1.0, 2.0 mg",
          x = "Supplement",
          y = "Length (mm)")

p3<- ggplot(ToothGrowth, aes(as.factor(dose), len)) +
  geom_boxplot(aes(fill = as.factor(dose))) +
  facet_grid(. ~ supp) +
  xlab('Dosage (mg)') +
  ylab('Length (mm)') +
  ggtitle('Supplement and Dosage') +
  theme(legend.position = "none")

grid.newpage()
pushViewport(viewport(layout = grid.layout(1,2, widths = c(0.5, 0.5))))
print(p1, vp = viewport(layout.pos.row = 1, layout.pos.col = 1))
print(p2, vp = viewport(layout.pos.row = 1, layout.pos.col = 2))

print(p3)

# The t-test based on Supplement
ttest_supp <- t.test(len ~ supp, ToothGrowth, var.equal = FALSE)

# The t-tests for each pair-wise combination of the dosage levels
ttest_dose1<-t.test(ToothGrowth$len[ToothGrowth$dose==0.5],
                    ToothGrowth$len[ToothGrowth$dose==1],
```

```
                    paired = FALSE, var.equal = TRUE)

ttest_dose2<-t.test(ToothGrowth$len[ToothGrowth$dose==1],
                    ToothGrowth$len[ToothGrowth$dose==2],
                    paired = FALSE, var.equal = TRUE)

ttest_dose3<-t.test(ToothGrowth$len[ToothGrowth$dose==0.5],
                    ToothGrowth$len[ToothGrowth$dose==2],
                    paired = FALSE, var.equal = TRUE)

# extract the right data and create a dataframe
test1 <- c(ttest_dose1$p.value,
           round(ttest_dose1$conf.int[1],2),
           round(ttest_dose1$conf.int[2],2))
test2 <- c(ttest_dose2$p.value,
           round(ttest_dose2$conf.int[1],2),
           round(ttest_dose2$conf.int[2],2))
test3 <- c(ttest_dose3$p.value,
           round(ttest_dose3$conf.int[1],2),
           round(ttest_dose3$conf.int[2],2))

headinglabels <- c("p Value",
                   "Confidence Interval-Low",
                   "Confidence Interval-High")

test_df <- data.frame(headinglabels,test1,test2,test3)
test_df
```