Part 2: Analysis of ToothGrowth dataset

Shayonendra Nath Tagore

25 May, 2020

1 Introduction

The Tooth Growth dataset records the effects of vitamin C on Guinea Pig tooth growth. The Vitamin C is delivered through either Orange Juice (OJ) or Vitamin C (VC). The dataset states the final tooth length (len), delivery method/supplement used (supp), and dosage (dose).

2 Basic Exploration

The code below gives a surface-level analysis of the dataset:

```
## Returns surface-level analysis
dim(ToothGrowth)
```

```
## [1] 60  3
```

```
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 ...
```

```
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
```

Dose needs further analysis:

```
## [1] 0.5 1.0 2.0
```

We can see that:

- len (length) - Numeric, over a range of values.
- supp (supplement) - Factor. This is the supplement given; either OJ (Orange Juice) or VC (Vitamin C)
- dose (dosage) - Numeric. Covers 3 distinct dosages: 0.5, 1.0, or 2.0.

Dose handled as numeric will cause problems downstream. To improve statistical handling, this column will be converted to a factor.

```
## '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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

3 Basic Summary

I will look at the length at each dose, separated by supplement.

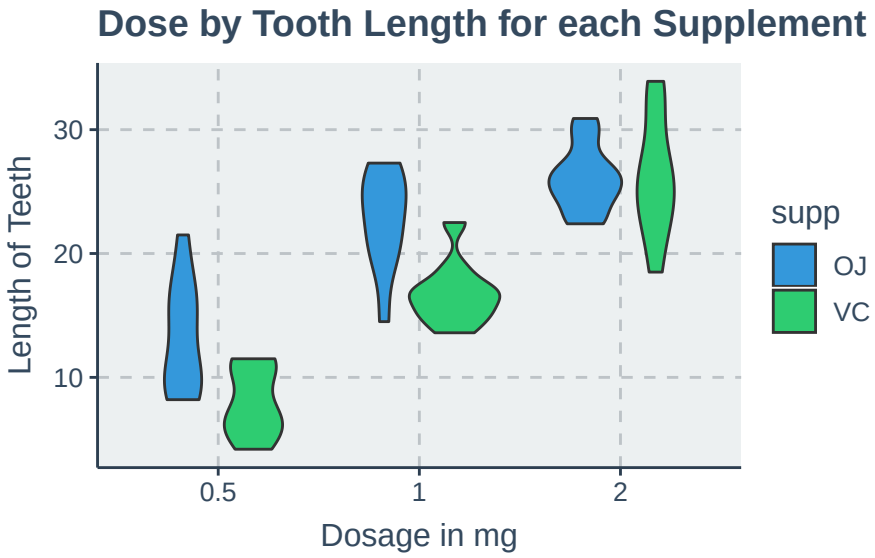

Figure 1: Violin plot of tooth growth at each dose, separated by supplement.

Between OJ and VC at each dose, OJ appears to led to result in greater lengths except at dose 2. Further, OJ at dose 0.5mg appears as competitive as VC at 1mg. These observations would be better confirmed through statistical analysis.

4 Confidence Interval and Hypothesis Testing

The length of the tooth growth may be influenced by Vitamin C's delivery method. To explore this, the differences in the mean will be explored using the two-tailed T-test.

Hypothesis (H_a) : Different delivery methods results in a difference in the mean lengths.

Null Hypothesis (H_0) : There is no difference in the means of lengths due to delivery method.

Dosage is another influencing factor in this study. To isolate it, the T-test will look at length due to delivery method for each distinct dose.

With $\alpha = 0.05$, only two of the tests are non-significant (P-value $> \alpha$): OJ1~VC2 and OJ2~VC2.

5 Conclusion

The following assumptions have been made:

- The observed variables are representative of the population.
- The variables follow a normal distribution.
- The samples are independent.
- Guinea Pigs were randomly assigned a supplement and dosage.

Table 1: t.tests for length and delivery method at varying doses

variations	p.value	conf.int.lower	conf.int.upper
OJ0.5~OJ1	0.0000878	-13.4156344	-5.5243656
OJ0.5~OJ2	0.0000013	-16.3352406	-9.3247594
OJ0.5~VC0.5	0.0063586	1.7190573	8.7809427
OJ0.5~VC1	0.0460103	-7.0081090	-0.0718910
OJ0.5~VC2	0.0000072	-17.2635219	-8.5564781
OJ1~OJ0.5	0.0000878	5.5243656	13.4156344
OJ1~OJ2	0.0391951	-6.5314425	-0.1885575
OJ1~VC0.5	0.0000000	11.5185075	17.9214925
OJ1~VC1	0.0010384	2.8021482	9.0578518
OJ1~VC2	0.0965261	-7.5643336	0.6843336
OJ2~OJ0.5	0.0000013	9.3247594	16.3352406
OJ2~OJ1	0.0391951	0.1885575	6.5314425
OJ2~VC0.5	0.0000000	15.5418168	20.6181832
OJ2~VC1	0.0000002	6.8596674	11.7203326
OJ2~VC2	0.9638516	-3.7980705	3.6380705
VC0.5~VC1	0.0000007	-11.2657120	-6.3142880
VC0.5~VC2	0.0000000	-21.9015120	-14.4184880
VC1~VC0.5	0.0000007	6.3142880	11.2657120
VC1~VC2	0.0000916	-13.0542667	-5.6857333
VC2~VC0.5	0.0000000	14.4184880	21.9015120
VC2~VC1	0.0000916	5.6857333	13.0542667

From the analysis above, it is clear that OJ has a greater effect on tooth growth except at dose 2mg, at which point VC and OJ are equivalent. From Figure 1 and Table 1, increasing the dosage for OJ led to improved tooth growth. The only two tests that were non-significant show that OJ at 1mg or 2mg is equivalent with VC at 2mg.

6 Appendix

```
### loading relevant packages and datasets
pacman::p_load(datasets,  # holds the ToothGrowth dataset
               magrittr,  # piping
               ggthemr,   # theming
               knitr,     # table formating
               dplyr      # data handling
               )
ggthemr::ggthemr("flat")  # theming
data(ToothGrowth)         # loads the ToothGrowth dataset

### surface-level analysis of the dataset
## Returns surface-level analysis
dim(ToothGrowth)
str(ToothGrowth)
summary(ToothGrowth)

## Further investigation of ToothGrowth$dose
ToothGrowth$dose %>% unique

## converting ToothGrowth$dose to a factor
ToothGrowth$dose <- factor(ToothGrowth$dose)
str(ToothGrowth)

### Basic Summary
## Generate scatter plot (figure 1)
ggplot(ToothGrowth, aes(x = dose, y = len, fill = supp)) +
    geom_violin(position=position_dodge(0.7)) +
    scale_x_discrete("Dosage in mg") +
    scale_y_continuous("Length of Teeth") +
    ggtitle("Dose by Tooth Length for each Supplement")

## Code to complete T-tests on tooth length and delivery method, at varying doses.
## Returns a formatted table.
save2table <- function(tmpVar,i,k,var1,var2) {
    tmpDF <- data.frame(variations=character(),
                        var1=character(),
                        var1.doses=numeric(),
                        var2=character(),
                        var2.doses=numeric(),
                        p.value=numeric(),
                        conf.int.lower=numeric(),
                        conf.int.upper=numeric(),
                        stringsAsFactors=FALSE)
    tmpDF[1,1] <- paste(var1,doses[i],"~",var2,doses[k], sep = "")
    tmpDF[1,2] <- var1
    tmpDF[1,3] <- doses[i]
    tmpDF[1,4] <- var2
    tmpDF[1,5] <- doses[k]
    tmpDF[1,6] <- tmpVar$p.value
    tmpDF[1,7] <- tmpVar$conf.int[1]
    tmpDF[1,8] <- tmpVar$conf.int[2]
    return(tmpDF)
```

```
}

OJ  <- ToothGrowth[ToothGrowth$supp=="OJ", ]
VC  <- ToothGrowth[ToothGrowth$supp=="VC", ]

doses       <- c(0.5,1,2)

hypotesting <- data.frame(variations=character(),
                          var1=character(),
                          var1.doses=numeric(),
                          var2=character(),
                          var2.doses=numeric(),
                          p.value=numeric(),
                          conf.int.lower=numeric(),
                          conf.int.upper=numeric(),
                          stringsAsFactors=FALSE)

# Handles all variations of possible t.tests in a non-optimized manner.
for (i in 1:length(doses)) {
    for (k in 1:length(doses)) {
        hypotesting <-
            t.test(OJ$len[OJ$dose == doses[i]],
                   VC$len[VC$dose == doses[k]]) %>%
            save2table(i,k,"OJ","VC") %>%
            dplyr::bind_rows(hypotesting)
        hypotesting <-
            t.test(OJ$len[OJ$dose == doses[i]],
                   OJ$len[OJ$dose == doses[k]]) %>%
            save2table(i,k,"OJ","OJ") %>%
            dplyr::bind_rows(hypotesting)
        hypotesting <-
            t.test(VC$len[VC$dose == doses[i]],
                   VC$len[VC$dose == doses[k]]) %>%
            save2table(i,k,"VC","VC") %>%
            dplyr::bind_rows(hypotesting)
    }
}
hypotesting <- hypotesting[order(hypotesting$variations),]

# Removing variations where it's testing against itself. redundent.
for (i in 1:length(hypotesting$p.value)) {
    ifelse((hypotesting$p.value[i] == 1.000000e+00),
           hypotesting <- hypotesting[-i, ],
           NA)
}
rownames(hypotesting) <- 1:nrow(hypotesting)

## Making a table to display the discovered T-test relationships.
hypotesting %>%
    select(variations,
           p.value,
           conf.int.lower,
           conf.int.upper) %>%
```

```
    kable(format = "latex",
          booktabs = TRUE,
          caption = "t.tests for length and delivery method at varying doses",
          row.names = NA)

## checking for T-tests yielding non-significant relationships.
count_nonsig <- data.frame(variations=character(),
                           var1=character(),
                           var1.doses=numeric(),
                           var2=character(),
                           var2.doses=numeric(),
                           p.value=numeric(),
                           conf.int.lower=numeric(),
                           conf.int.upper=numeric(),
                           stringsAsFactors=FALSE)
for (i in 1:length(hypotesting$p.value)) {
    ifelse((hypotesting$p.value[i] > 0.05),
           count_nonsig <- dplyr::bind_rows(count_nonsig, hypotesting[i, ]),
           NA)
}
```