Project Part 2: A Study on the Effect of Delivery and Dosage on Tooth Growth in Guinea Pigs

Ryan Hammer

Introduction

This second part of the course project in the Coursera Statistical Inference course involves analyzing the ToothGrowth data set that comes built in to R. The data set results from a study on the effect of vitamin C on tooth growth in 60 guinea pigs. Ultimately we would like to answer the question

Do delivery and/or dosage affect the level of tooth growth in guinea pigs?

Data Summary

We begin by loading the data set into a data frame and looking at a summary using R's str() function, the results of which are below.

```
## Classes 'tbl_df', 'tbl' and '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 ...
```

We can see that there are three variables in the data, len for tooth length, supp for supplement type, and dose for the numeric dose of vitamin C given to the test subjects. Supplement type is a factor variable showing whether the vitamin C was administered using orange juice or ascorbic acid, while tooth length and dose are numeric variables measured in length of odontoblasts and milligrams/day, respectively. Using the n_distinct() function from the dplyr R package reveals that the dose variable only has three different values in the data, and using the unique() reveals them to be 0.5, 1.0, and 2.0 mg/day.

```
n_distinct(toothdata$dose)
```

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

```
unique(toothdata$dose)
```

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

With this in mind, we can get an idea of what the data looks like with some general summary calculations and data visualizations. The table below shows the mean and median length for each supplement type for each dose.

Supp_Type	Dose	Mean	Median
OJ	0.5	13.23	12.25
OJ	1.0	22.70	23.45
OJ	2.0	26.06	25.95
VC	0.5	7.98	7.15
VC	1.0	16.77	16.50
VC	2.0	26.14	25.95

The table shows that the means and medians for the length variable increase for both supplements as the dose level rises. For doses of 0.5 and 1.0 milligrams per day, the mean and median length of odontoplasts is larger using orange juice to supplement vitamin C. We can further explore the length data through plotting. Figure 1 below shows boxplots of the length variable for each dose amount separated by supplement type.

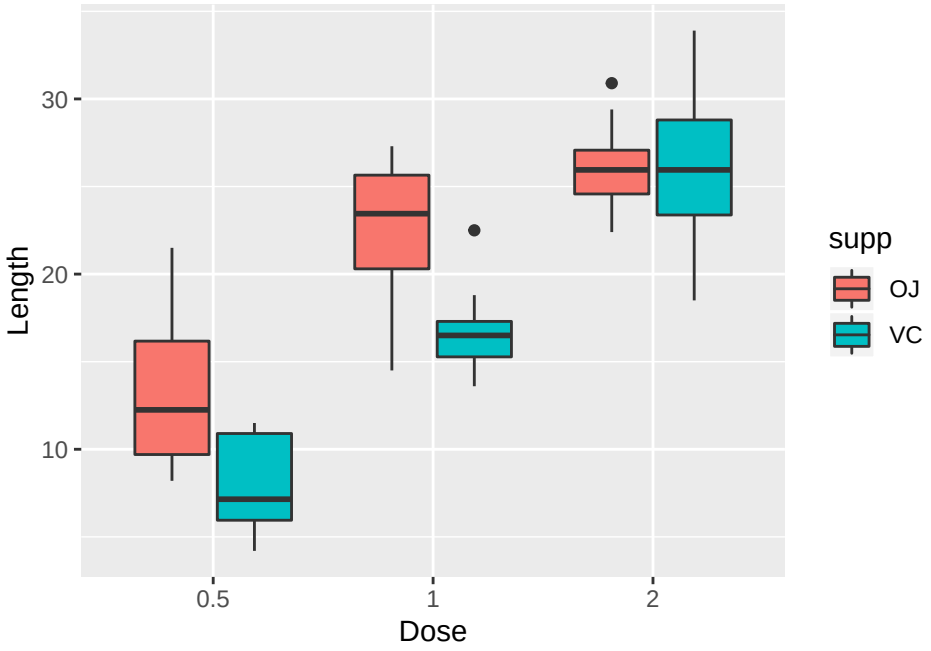

Figure 1: Tooth Length Boxplots

The figure appears reinforce what is seen in the table. There seems to be a trend of larger doses leading to higher lengths for both supplement types. It does also appear that for lower doses there is a more pronounced growth when the vitamin C was administered using orange juice as opposed to using ascorbic acid. We can test the significance of the effects of supplement and dosage using T hypothesis tests. Before starting that process, we should check the normality of the data. A histogram of all 60 observations of the length variable is shown below in Figure 2.

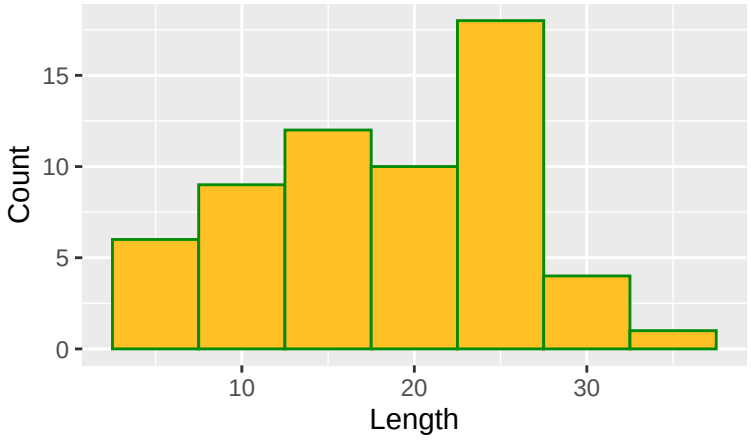

Figure 2: Tooth Lengths

The distribution of lengths shown by the plot is not perfectly normal, but does tend to cluster around its center while thinning out on either end. A further test for normality can be down using a Shapiro-Wilk Test. Using the stat.desc() function in the pastecs package in R, the Shapiro-Wilk probability is p = 0.1091 The alpha value used in test was 0.05. The Shapiro-Wilk Test states that the null hypothesis, that the data in question is normally distributed, should be rejected for probability values lower than alpha. Thus, the

assumption that the tooth length data is normally distributed is reasonable, and we can validly apply T hypothesis tests.

Hypothesis Tests

To attempt to answer the question previously posed in the introduction, it will be necessary to perform hypothesis tests related to the effects of supplement type and dosage amount. First, we will test supplement type. The null hypothesis will be that supplement type does not affect tooth growth length. We will test this by assuming the difference between the mean values for each supplement type is equal to 0. The althernative hypothesis will be that there is a difference between the two mean values. In the data set, 30 obesrvations were given vitamin C using orange juice, and the other 30 using ascorbic acid. The code below will separate the original data by supplement type, then perform a two-sided T test on the two sets of length data.

```
ojdata <- tbl_df(filter(toothdata, supp == "OJ"))
vcdata <- tbl_df(filter(toothdata, supp == "VC"))
t.test(ojdata$len, vcdata$len, conf.level = 0.95)$p.value
```

```
## [1] 0.06063451
```

Using $\alpha = 0.05$ and assuming unequal variances, the test produces a p value of 0.0606. Because $p > \alpha$, we fail to reject the null hypothesis based on the T hypothesis test. Recall the appearance of the boxplot in Figure 1; for the two lower dosage values, there did appear to be a difference in the tooth lengths for each supplement type. To further explore this, we can perform T hypothesis tests separating the length data by both dosage and supplement type. Because this will involve multiple hypothesis tests on the same data, it becomes necessary to use a correction method to control for errors that may arise due to multiple testing. The method used here will be the "BH" method for correcting the False Discovery Error Rate (FDAR) seen in week of the course. The R code below shows creation of data subsets based on dosage and type.

```
oj0.5 <- filter(toothdata, supp == "OJ") %>% filter(dose == 0.5) %>% tbl_df()
vc0.5 <- filter(toothdata, supp == "VC") %>% filter(dose == 0.5) %>% tbl_df()
oj1 <- filter(toothdata, supp == "OJ") %>% filter(dose == 1.0) %>% tbl_df()
vc1 <- filter(toothdata, supp == "VC") %>% filter(dose == 1.0) %>% tbl_df()
oj2 <- filter(toothdata, supp == "OJ") %>% filter(dose == 2.0) %>% tbl_df()
vc2 <- filter(toothdata, supp == "VC") %>% filter(dose == 2.0) %>% tbl_df()
```

Now we can use the subsets to run T tests comparing supplement type on each dosage level and place the results in a vector of p values, including the original p value determined above. Those p values are then run through R's p.adjust() function using the BH correction method to determine how many p values meet a level allowing for rejection of the null hypothesis. The null hypothesis for each test is that the difference between means for each supplement type is zero. The result of those calculations are in the table below.

Supp_overall	Supp_0.5	Supp_1.0	Supp_2.0
0.080846	0.0127172	0.0041535	0.9638516

Conclusions

Using $\alpha = 0.5$ and the "BH" FDAR correction method due to multiple testing, we are able to reject the null hypothesis that mean tooth lengths for each supplement type are the same for dosage levels of 0.5 and 1.0 mg/day. The T tests were two-tailed, and conducted under the assumptions that the distribution of tooth lengths was approximately normal, and that variances between groups were unequal.

Appendix

```
suppressWarnings(library(ggplot2))
suppressWarnings(library(tidyverse))
suppressWarnings(library(knitr))
suppressWarnings(library(pastecs))
toothdata <- tbl_df(ToothGrowth)
str(toothdata)
n_distinct(toothdata$dose)
unique(toothdata$dose)
lengthsummary <- toothdata %>% group_by(supp, dose) %>%
    summarize(Mean = mean(len), Median = median(len))
lengthsummary <- rename(lengthsummary, Supp_Type = supp, Dose = dose)
kable(lengthsummary)
qplot(x=as.factor(dose), y=len, data = toothdata, geom = "boxplot",
      fill = supp, xlab = "Dose", ylab = "Length")
qplot(len, data = toothdata, binwidth = 5, xlab = "Length",
      ylab = "Count", fill = I("goldenrod1"), col = I("green4"))
a <- round((stat.desc(toothdata$len, norm = TRUE)[20]), 4)
cat(a)
ojdata <- tbl_df(filter(toothdata, supp == "OJ"))
vcdata <- tbl_df(filter(toothdata, supp == "VC"))
t.test(ojdata$len, vcdata$len, conf.level = 0.95)$p.value
oj0.5 <- filter(toothdata, supp == "OJ") %>% filter(dose == 0.5) %>% tbl_df()
vc0.5 <- filter(toothdata, supp == "VC") %>% filter(dose == 0.5) %>% tbl_df()
oj1 <- filter(toothdata, supp == "OJ") %>% filter(dose == 1.0) %>% tbl_df()
vc1 <- filter(toothdata, supp == "VC") %>% filter(dose == 1.0) %>% tbl_df()
oj2 <- filter(toothdata, supp == "OJ") %>% filter(dose == 2.0) %>% tbl_df()
vc2 <- filter(toothdata, supp == "VC") %>% filter(dose == 2.0) %>% tbl_df()
pvector <- t.test(ojdata$len, vcdata$len, conf.level = 0.95)$p.value
pvector[2] <- t.test(oj0.5$len, vc0.5$len, conf.level = 0.95)$p.value
pvector[3] <- t.test(oj1$len, vc1$len, conf.level = 0.95)$p.value
pvector[4] <- t.test(oj2$len, vc2$len, conf.level = 0.95)$p.value
pvalsdf <- data.frame(NA)
pvalsdf[1, 1:4] <- p.adjust(pvector, method = "BH")
colnames(pvalsdf) <- c("Supp_overall", "Supp_0.5", "Supp_1.0", "Supp_2.0")
kable(pvalsdf)
```