Statistical Inference Project - Part 2

Jonathan Brant

Friday, January 23, 2015

Synopsis

The goal of the second portion of this project is to analyze the ToothGrowth dataset, which ships in the standard R datasets package. The dataset represents a study of the effect of Vitamin C on tooth growth in Guinea Pigs. Tooth length was measured with respect to two supplements (vitamin C and orange juice) and three doses for both (0.5 mg, 1 mg, and 2 mg). The intent of this work is to perform some level of exploratory analysis on the data, produce data summaries for easy visualization of overall trends and data set properties, as well as utilize hypothesis tests and accompanying confidence intervals to compare tooth growth by both supplement and dose. The latter will ideally allow us to make inferential statements regarding supplement and dose efficacy with a reasonable level of confidence.

Data Analysis and Summary

The first two tasks involve loading the ToothGrowth dataset, performing some high level analyses, and generating summaries to give the reader a general overview of the data. First, we want to load the dataset and take a look at its structure:

```
## Load Tooth Growth dataset
data(ToothGrowth)

## Determine the structure of the dataset
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 ...
```

As described, the dataset consists of two supplements (orange juice and vitamin C), three dosages (0.5 mg, 1 mg, and 2 mg), and 60 observations of tooth lengths given the aforementioned factors. In order to get a quick idea of the spread of the data, a summary is printed which calculates the mean, median, minimum/maximum values, and the distribution quantiles.

```
## Get an idea of the spread of the 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
```

As a more thorough summary of the data characteristics, table 1 enumerates the mean tooth length, standard deviation of the tooth length, and variance of the tooth length for each combination of supplement and dosage. Notice what appears to be a linear relationship between dosage and tooth length for both of the supplements with an inverse relationship in variances between the two supplements for the given dosage (i.e. the variance in tooth length for orange juice decreases as the dosage increases while the opposite occurs for vitamin C).

Supplement	Dosage	Mean Length	S.D.	Variance
OJ	0.5	13.23	4.46	19.89
OJ	1	22.7	3.91	15.3
OJ	2	26.06	2.66	7.05
VC	0.5	7.98	2.75	7.54
VC	1	16.77	2.52	6.33
VC	2	26.14	4.8	23.02

Table 1: Tooth Length Summary by Supplement and Dosage

In order to help visualize the distribution, figure 1 depicts the mean and interquartile range for the tooth length resulting from the application of both supplements at each of the three dosages.

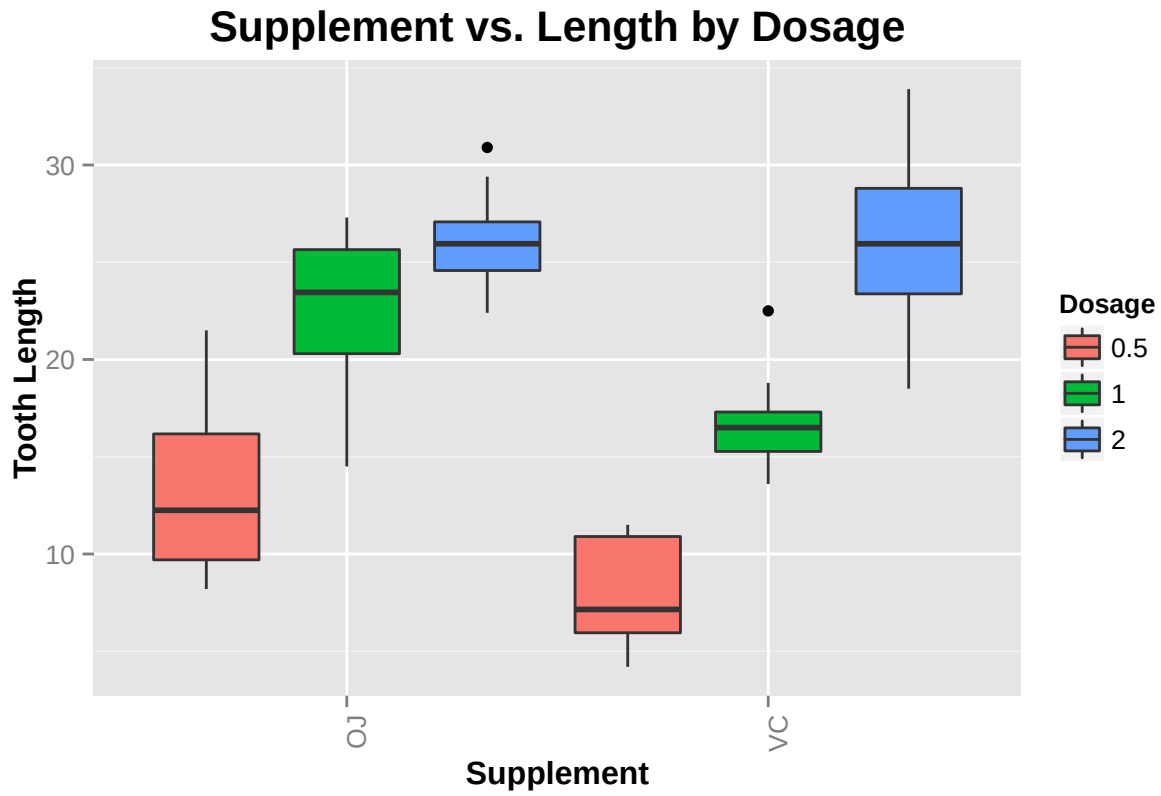

Figure 1: Effect of Supplement and Dosage on Tooth Length

Tooth Growth and Dosage

After analyzing the data above, the natural question is how do both the supplement and dosage compare with regard to their effect on tooth growth. In particular, is there a statistically significant indication that one supplement given at a particular dosage is superior to the other? In order to answer this question, two hypothesis tests are setup per dosage (three test cases in total). The null hypothesis (H_0) is that there is no difference in tooth growth between orange juice and vitamin C at the given dosage. The alternative hypothesis (H_a) posits that there is, in fact, a statistically significant difference in tooth growth between the two supplements at the given dosage. As is customary, the null hypothesis is the default position unless proven otherwise. The test cases are summarized below.

Test case 1

H_0 : There is no difference in tooth growth between orange juice and vitamin C at a dosage of 0.5 mg.
H_a : There is a difference in tooth growth between orange juice and vitamin C at a dosage of 0.5 mg.

Test case 2

H_0 : There is no difference in tooth growth between orange juice and vitamin C at a dosage of 1 mg.
H_a : There is a difference in tooth growth between orange juice and vitamin C at a dosage of 1 mg.

Test case 3

H_0 : There is no difference in tooth growth between orange juice and vitamin C at a dosage of 2 mg.
H_a : There is a difference in tooth growth between orange juice and vitamin C at a dosage of 2 mg.

For the purposes of this study, statistical significance is derived using a two-tailed Student's t-Test for a 95% confidence interval. A positive t-Statistic and a p-Value below the 2.5% cutoff indicates that orange juice has a greater impact on tooth growth while a negative t-Statistic and a p-Value above 97.5% indicates that vitamin C has a greater impact on tooth growth. Table 2 depicts the results.

	t-Statistic	Degrees of Freedom	p-Value
Hypothesis 1	2.98	9	0.02
Hypothesis 2	3.37	9	0.01
Hypothesis 3	-0.04	9	0.97

Table 2: Hypothesis Test Results

The first two hypothesis tests have a postivie t-Statistic and fall below the bottom 2.5% cutoff, necessitating a rejection of the null hypothesis and indicating statistical significance in favor of orange juice as a more effective driver for tooth growth at dosages of 0.5 mg and 1 mg. While the third test (2 mg) is very close, it has a negative t-Statistic and fails to exceed the 97.5% upper tail cutoff, so the null hypothesis is not rejected, meaning that we cannot make the argument that vitamin C results in greater tooth growth than orange juice for doses of 2 mg.

Appendix

Code for generating summary statistics

```
## Calculate the mean of each combination of supplement and dosage
mean.analysis <- aggregate(ToothGrowth$len, list(ToothGrowth$supp, ToothGrowth$dose), mean)

## Calculate the standard deviation of each combination of supplement and dosage (rounding to 2)
sd.analysis <- aggregate(ToothGrowth$len, list(ToothGrowth$supp, ToothGrowth$dose), sd)

## Calculate the variance of each combination of supplement and dosage (rounding to 2)
var.analysis <- aggregate(ToothGrowth$len, list(ToothGrowth$supp, ToothGrowth$dose), var)

## Round standard deviation and variance to 2 decimals
sd.analysis[,3] <- round(sd.analysis[,3], 2)
var.analysis[,3] <- round(var.analysis[,3], 2)

## Set the column names on the resulting data frames
colnames(mean.analysis) <- c("Supplement", "Dosage", "Mean Length")
colnames(sd.analysis) <- c("Supplement", "Dosage", "S.D.")
colnames(var.analysis) <- c("Supplement", "Dosage", "Variance")

## Merge the mean and variance data frames and print
## (in the report, this is done using pandoc)
merge(mean.analysis,
      merge(sd.analysis,
            var.analysis,
            by = c("Supplement", "Dosage")),
      by = c("Supplement", "Dosage"))
```

Code for Plotting Figure 1

```
suppressMessages(library(ggplot2))
suppressMessages(require(gridExtra))

## Calculate the mean of each combination of supplement and dosage
mean.analysis <- aggregate(ToothGrowth$len, list(ToothGrowth$supp, ToothGrowth$dose), mean)

## Calculate the standard deviation of each combination of supplement and dosage (rounding to 2)
sd.analysis <- aggregate(ToothGrowth$len, list(ToothGrowth$supp, ToothGrowth$dose), sd)

## Calculate the variance of each combination of supplement and dosage (rounding to 2)
var.analysis <- aggregate(ToothGrowth$len, list(ToothGrowth$supp, ToothGrowth$dose), var)

## Round standard deviation and variance to 2 decimals
sd.analysis[,3] <- round(sd.analysis[,3], 2)
var.analysis[,3] <- round(var.analysis[,3], 2)

## Set the column names on the resulting data frames
colnames(mean.analysis) <- c("Supplement", "Dosage", "Mean Length")
colnames(sd.analysis) <- c("Supplement", "Dosage", "S.D. of Length")
```

```r
colnames(var.analysis) <- c("Supplement", "Dosage", "Variance of Length")

## Merge the mean and variance data frames and print
## (in the report, this is done using pandoc)
merge(mean.analysis,
      merge(sd.analysis,
            var.analysis,
            by = c("Supplement", "Dosage")),
      by = c("Supplement", "Dosage"))

## Construct box plot of differences between dosage and supplements split into separate facets
dosage.vs.supp.plot <- ggplot(ToothGrowth, aes(x = supp, y = len)) +
  geom_boxplot(
    aes(fill = factor(dose)),
    position = position_dodge(1)) +
  labs(
    title = "Supplement vs. Length by Dosage",
    x = "Supplement",
    y = "Tooth Length",
    fill = "Dosage") +
  theme(plot.title = element_text(face = "bold", size = 16, vjust = 1),
        axis.title.x = element_text(face = "bold"),
        axis.title.y = element_text(face = "bold"),
        axis.text.x = element_text(angle = 90, hjust = 1))

## Display the box plot
suppressMessages(grid.arrange(dosage.vs.supp.plot,
                              sub = textGrob(
                                "Figure 1: Effect of Supplement and Dosage on Tooth Length",
                                gp = gpar(fontface = "bold", col = "grey20", fontsize = 12))))
```

Code for running t-Tests

```r
## Split up the dataset into the two separate supplements
vitamin.c <- subset(ToothGrowth, ToothGrowth$supp == "VC")
orange.juice <- subset(ToothGrowth, ToothGrowth$supp == "OJ")

## Hypothesis will need to be defined per dosage

## Hypothesis 1
### H0: There is no difference in tooth growth between orange juice and vitamin C at a dosage of 0.5 mg
### Ha: There is a difference in tooth growth between orange juice and vitamin C at a dosage of 0.5 mg

## Run confidence interval t-test for Hypothesis 1 (reject H0 for OJ)
hypothesis.1 <- t.test(
  subset(orange.juice, dose == 0.5)$len - subset(vitamin.c, dose == 0.5)$len,
  var.equal=TRUE)

## Hypothesis 2
### H0: There is no difference in tooth growth between orange juice and vitamin C at a dosage of 1 mg
### Ha: There is a difference in tooth growth between orange juice and vitamin C at a dosage of 1 mg
```

```r
## Run confidence interval t-test for Hypothesis 2 (reject H0 for OJ)
hypothesis.2 <- t.test(
  subset(orange.juice, dose == 1)$len - subset(vitamin.c, dose == 1)$len,
  var.equal=TRUE)

## Hypothesis 3
### H0: There is no difference in tooth growth between orange juice and vitamin C at a dosage of 2 mg
### Ha: There is a difference in tooth growth between orange juice and vitamin C at a dosage of 2 mg

## Run confidence interval t-test for Hypothesis 3 (reject H0 for VC)
hypothesis.3 <- t.test(
  subset(orange.juice, dose == 2)$len - subset(vitamin.c, dose == 2)$len,
  var.equal=TRUE)

## Concatenate all t-statistics
t.stats <- c(
  round(hypothesis.1$statistic,2),
  round(hypothesis.2$statistic,2),
  round(hypothesis.3$statistic,2))

## Concatenate all degrees of freedom
deg.freedom <- c(hypothesis.1$parameter, hypothesis.2$parameter, hypothesis.3$parameter)

## Concatenate all p-values
p.values <- c(
  round(hypothesis.1$p.value,2),
  round(hypothesis.2$p.value,2),
  round(hypothesis.3$p.value,2))

## Construct and format the hypothesis data frame
hypothesis.df <- data.frame(t.stats, deg.freedom, p.values,
                            row.names = c("Hypothesis 1", "Hypothesis 2", "Hypothesis 3"))
colnames(hypothesis.df) <- c("t-Statistic", "Degrees of Freedom", "p-Value")
```