Statistical Inference Course Project, Pt 2

K. Divis

Overview

This short report analyzes the ToothGrowth data in the R datasets package. It was completed as part of (and under the guidelines of) the Johns Hopkins Data Science specialization Statistical Inference class on Coursera.

Load data and exploratory analyses

The data can be found in the R datasets package under "ToothGrowth". The following description was provided (using ?ToothGrowth): "The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid)."

Load the data:

```
library(datasets)
data(ToothGrowth)

# Rename variables and treat dose and supplement as factors:
names(ToothGrowth) = c("Length", "Supplement", "Dose")
ToothGrowth$Dose = as.factor(ToothGrowth$Dose)
ToothGrowth$Supplement = as.factor(ToothGrowth$Supplement)
```

Figure 1 shows length of teeth by dose and supplement. The dosage of Vitamin C ranges from 0.5 to 1 to 2. The supplement refers to either orange juice (OJ) in the salmon color or ascorbic acid (VC) in the cyan color. Table 1 shows the mean length of teeth by dose and supplement. Table 2 shows the standard deviation of the length of teeth by dose and supplement.

Based on visual inspection, there appears to be a main effect of increased teeth growth in the OJ compared to VC condition (though this effect weakens or even disappears at the highest dose). There also appears to be a main effect of dose, with higher doses leading to more teeth growth. Based on this plot and inspection, I will look at the main effect of supplement and the main of effect of high vs. low dose.

Confidence Intervals and/or Hypothesis Tests

As noted above, I plan on looking at the overall effect of supplement and the overall effect of high vs. low dose.

1. Overall effect of supplement:
   H0: Difference between mean of OJ and mean of VC is zero
   HA: Difference between mean of OJ and mean of VC is not zero

2. Overall effect of high vs. low dose:
   H0: Difference between mean of low dose and mean of high dose is zero
   HA: Difference between mean of low dose and mean of high dose is not zero

Code for performing the t-test and outputting the confidence intervals and p-value:

```
# Pull appropriate data:
VC = NULL
OJ = NULL
lowDose = NULL
highDose = NULL
countVC = 0
countOJ = 0
countLow = 0
countHigh = 0
for (i in 1:length(ToothGrowth$Length)) {
    if (ToothGrowth$Supplement[i] == "VC") {
        countVC = countVC + 1
        VC[countVC] = ToothGrowth$Length[i]
    } else {
        countOJ = countOJ + 1
        OJ[countOJ] = ToothGrowth$Length[i]
    }
    if (ToothGrowth$Dose[i] == "0.5") {
        countLow = countLow + 1
        lowDose[countLow] = ToothGrowth$Length[i]
    }
    if (ToothGrowth$Dose[i] == "2") {
        countHigh = countHigh + 1
        highDose[countHigh] = ToothGrowth$Length[i]
    }
}
sdVC = sd(VC)
sdOJ = sd(OJ)
sdLow = sd(lowDose)
sdHigh = sd(highDose)

# t-test comparing VC to OJ
suppConf = t.test(VC, OJ, paired = FALSE, var.equal = FALSE)$conf
suppP = t.test(VC, OJ, paired = FALSE, var.equal = FALSE)$p.value

# t-test comparing low dose (.5 mg) to high dose (2 mg)
doseConf = t.test(lowDose, highDose, paired = FALSE, var.equal = FALSE)$conf
doseP = t.test(lowDose, highDose, paired = FALSE, var.equal = FALSE)$p.value
```

The confidence interval for the t-test of the difference between the means of OJ and VC was **(-7.57, 0.17)**. The p-value was **0.06**.

The confidence interval for the t-test of the difference between the means of low dose and high dose was **(-18.16, -12.83)**. The p-value rounds to **0**.

Conclusions and assumptions

I chose to conservatively treat the data as unpaired with unequal variance. The description of the dataset left it unclear whether the guinea pigs were paired, and no ID markers were given if the data was paired. The standard deviation for OJ was 6.61 and the standard deviation for VC was 8.27. The standard deviation for low dose was 4.5 and the standard deviation for high dose was 3.77. Since there these pairs of standard deviations were different, I chose to take the conservative route of setting the variance as unequal in the test. Further follow-up analyses could investigate the validity of that assumption, but I prefer to keep my tests conservative. I set my alpha level at the traditional .05.

Based on those assumptions in the t-test, I found **no evidence to reject the null hypothesis** that OJ and VC supplements were tied to no different levels of tooth growth, on average (p-value $>$ .05; confidence interval contained 0). Based on inspection of Figure 1, the lack of main effect may be driven by a null effect at the higher dosage level (e.g., interaction between dosage and supplement) but that is outside of the bounds of the tests we have covered in class thus far.

Also based on the assumptions in the t-test explained above, I found **evidence to reject the null hypothesis** that low dose and high dose were tied to no difference in levels of tooth growth, on average (p-value $<$ .05; confidence interval did not contain 0). More tooth growth occurred at high doses than at low doses.

Figure and Tables

Figure 1:

```
library(ggplot2)
g <- ggplot(data = ToothGrowth, aes(Dose, Length))
g <- g + geom_point(size = 4, aes(color = Supplement), alpha = 0.8)
g <- g + ggtitle("Teeth Growth by Dose and Supplement") + ylab("Length of Teeth") +
    xlab("Dose in mg")
g
```

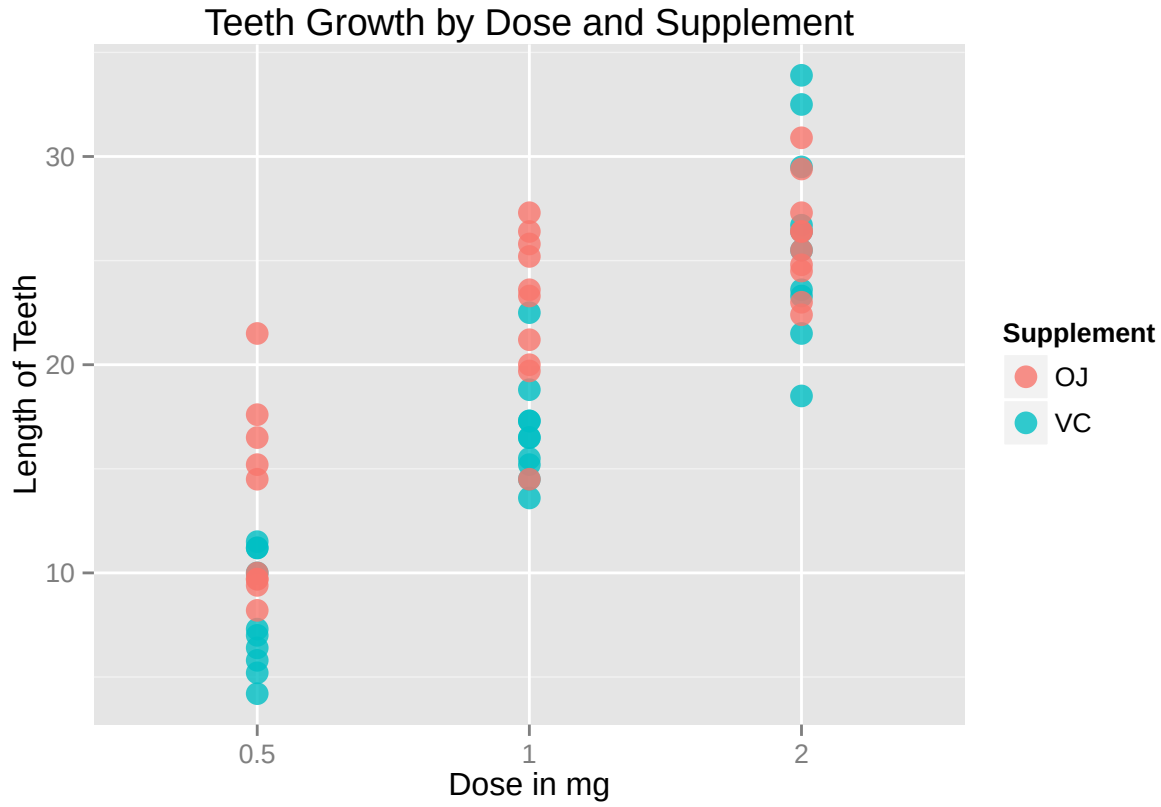

Table 1: Means by Dose and Supplement

```
library(knitr)
allMeans = as.table(tapply(ToothGrowth$Length, list(ToothGrowth$Dose, ToothGrowth$Supplement),
    mean))
kable(allMeans)
```

	OJ	VC
0.5	13.23	7.98
1	22.70	16.77
2	26.06	26.14

Table 2: Standard deviation by Dose and Supplement

```
allSD = tapply(ToothGrowth$Length, list(ToothGrowth$Dose, ToothGrowth$Supplement),
    sd)
kable(allSD, digits = 2)
```

	OJ	VC
0.5	4.46	2.75
1	3.91	2.52
2	2.66	4.80