Inference of the Tooth Growth Data Set

Travis Fell

Thursday, June 18, 2015

Overview

This analysis puts into action tools and techniques from the Exploratory Data Analysis class as well as hypothesis testing from the Statistical Inference class using the ToothData dataset in R.

Load and Explore the Data

First, we'll load the datasets package and other packages we will use in this analysis.

```
library(dplyr)
library(datasets)
data(ToothGrowth)
```

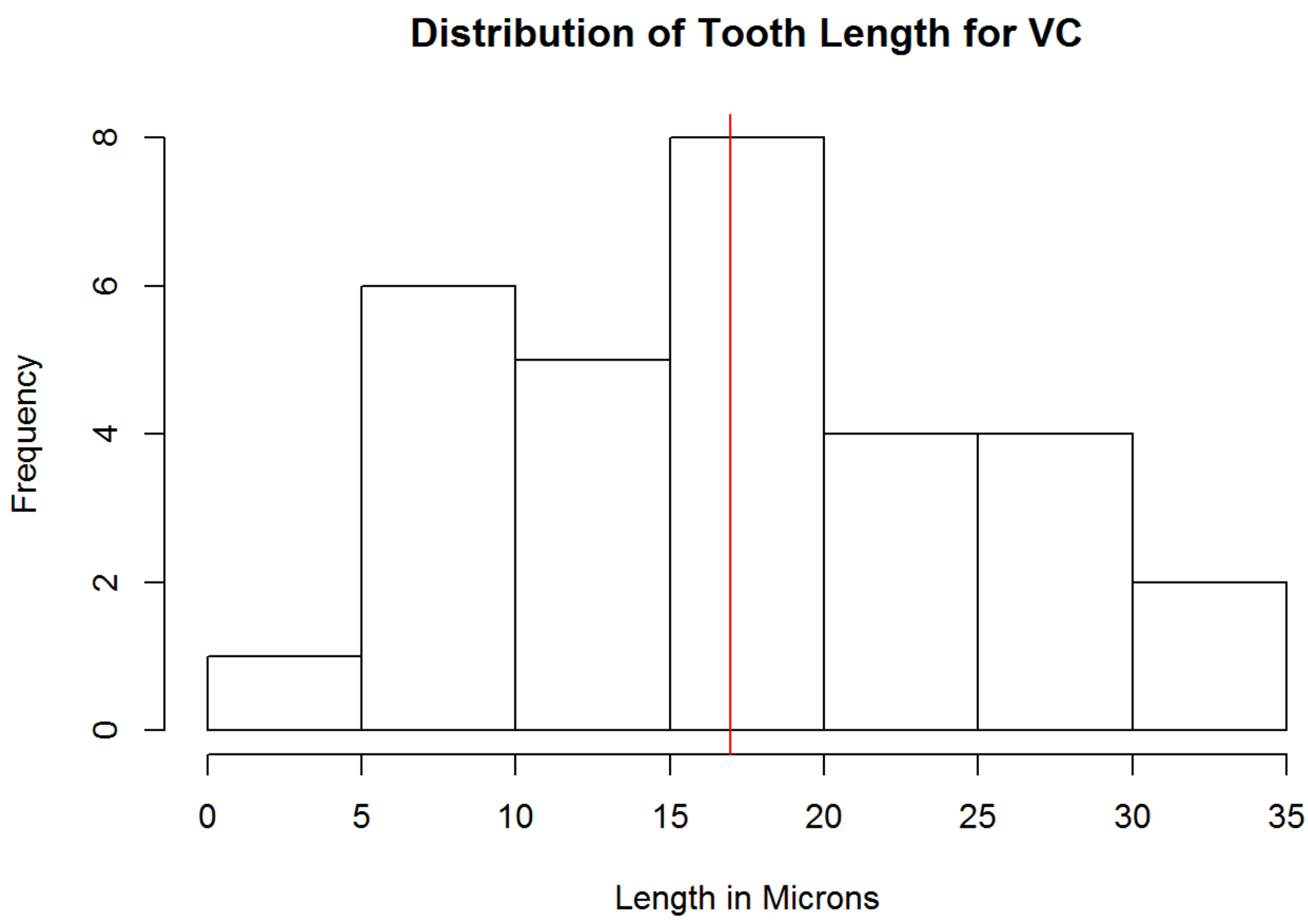
Now, we'll do some basic exploratory analysis just to get a feel for this data set.

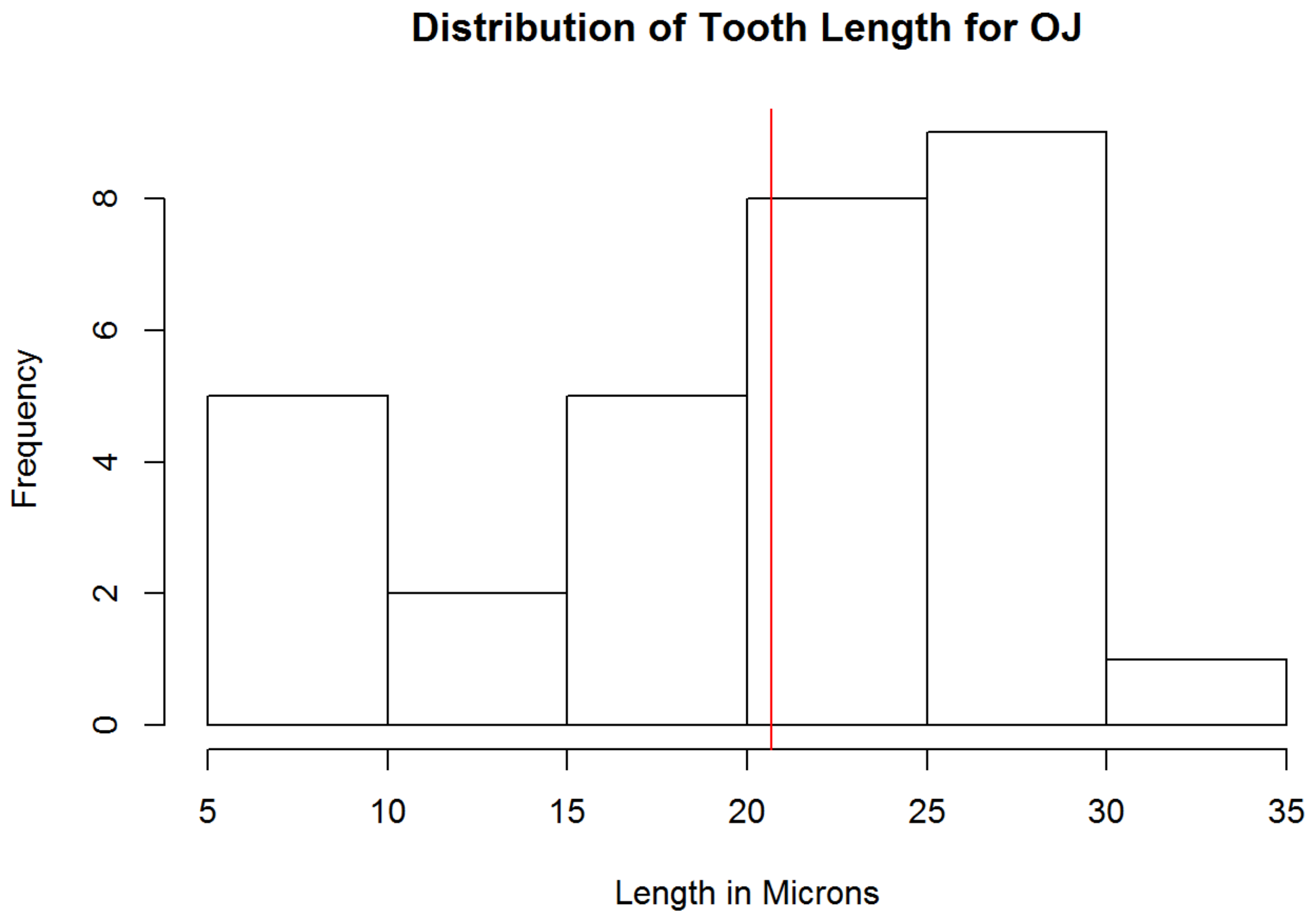
```
ToothGrowth[!complete.cases(ToothGrowth)]
head(ToothGrowth)
lenave <- mean(ToothGrowth$len)
lenrange <- range(ToothGrowth$len)
str(ToothGrowth)
summary(ToothGrowth$len)
nrow(subset(ToothGrowth, supp == "OJ"))
nrow(subset(ToothGrowth, supp == "VC"))
nrow(subset(ToothGrowth, dose == 0.5))
nrow(subset(ToothGrowth, dose == 1.0))
nrow(subset(ToothGrowth, dose == 2.0))
```

So far, the exploration of this data set has revealed the following:

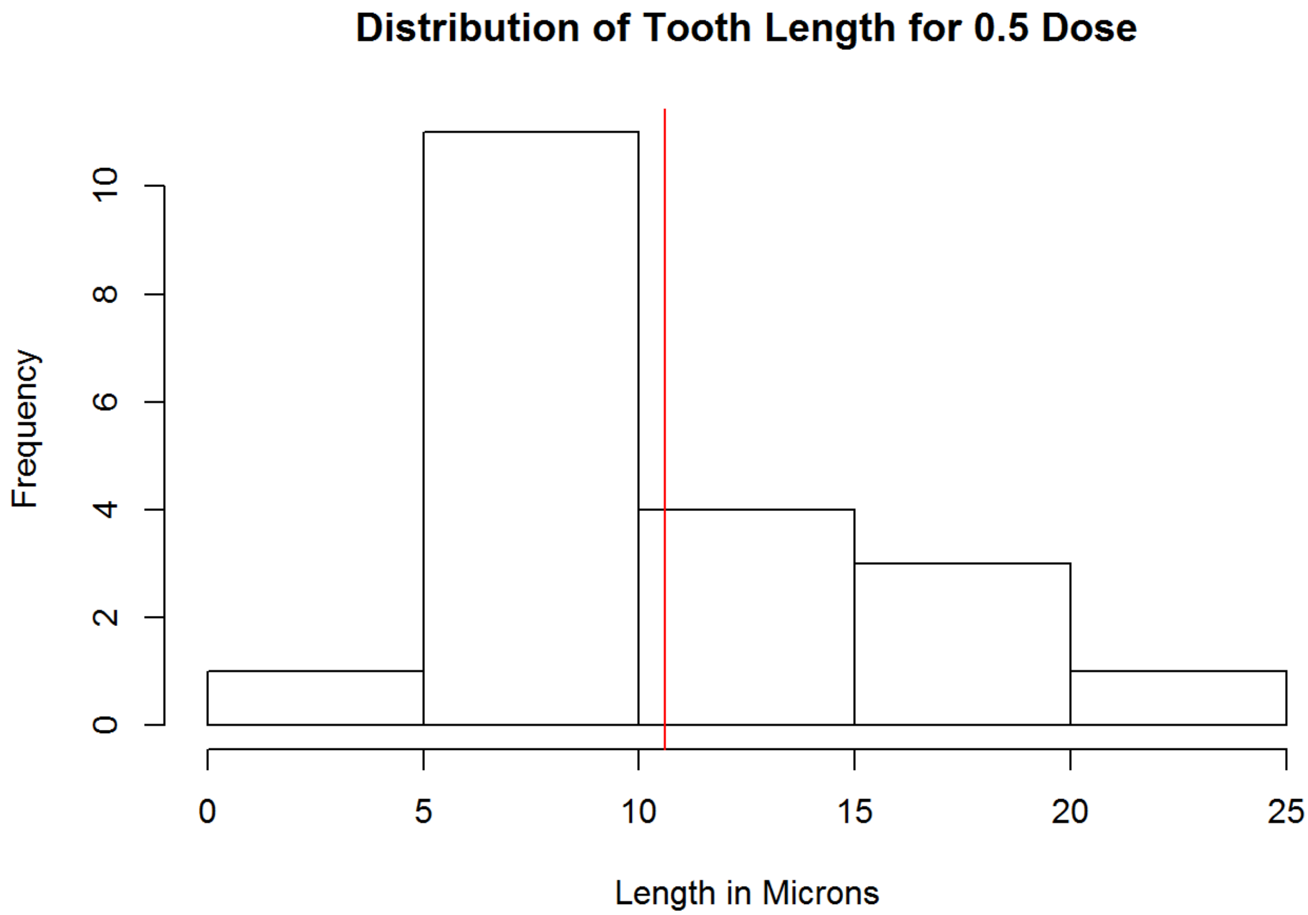
- The data set is 60 rows long with 3 columns, len, supp and dose, with no NA values
- The overall range in tooth lengths is 4.2, 33.9
- The average tooth length is 18.8133333
- The supp column is a factor that contains two Vitamin C supplement delivery methods, OJ and VC
- Half the data set has OJ for the supp value, the other half has VC
- There are three distinct dosage levels: 0.5, 1.0, 2.0, each with 20 records for each dosage level

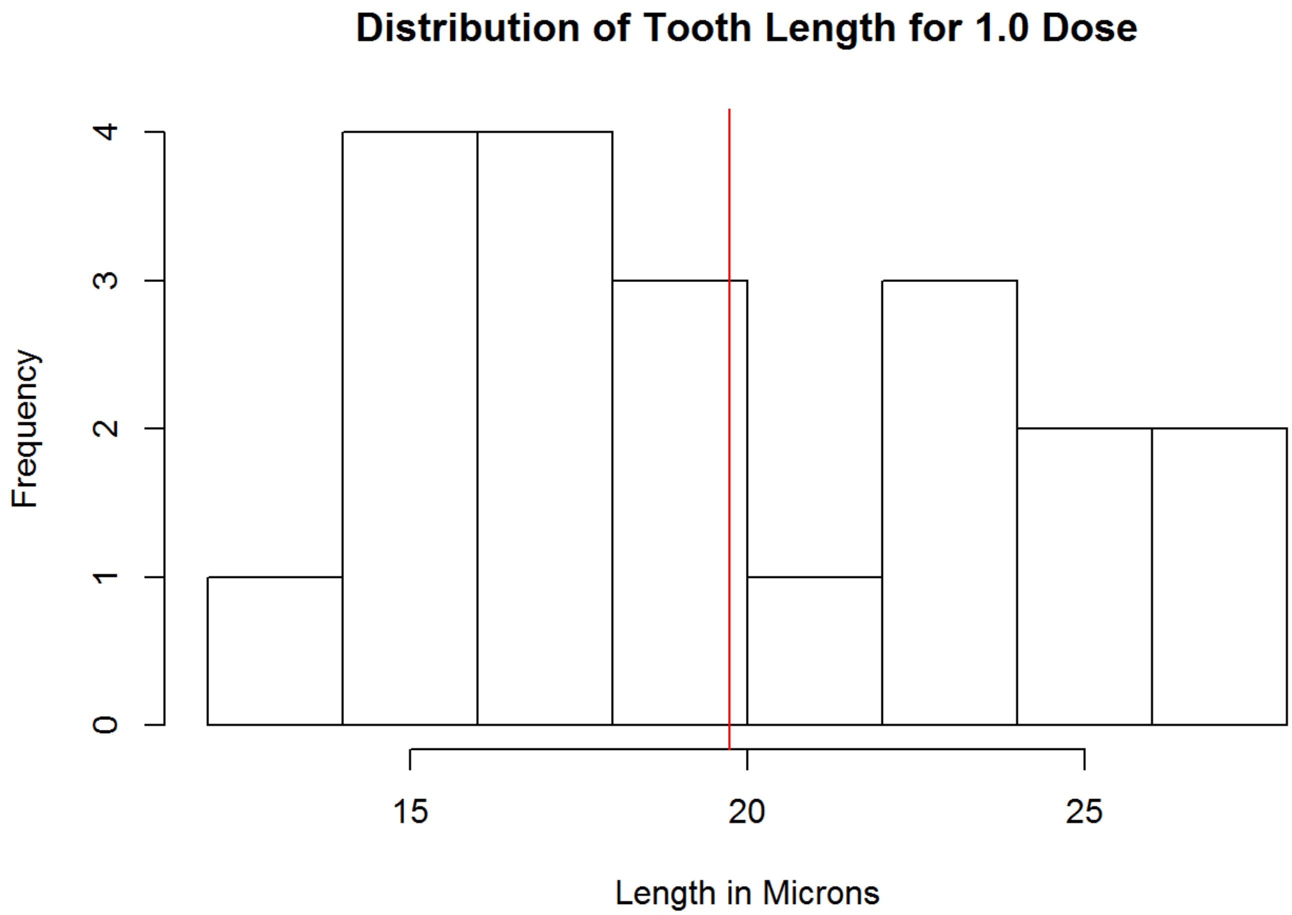
Let's look at the distribution of tooth lengths in microns by the dimensions of supp and dose. The means for the various distributions below are identified with a red vertical bar. Note: the code for these plots are in the appendix.

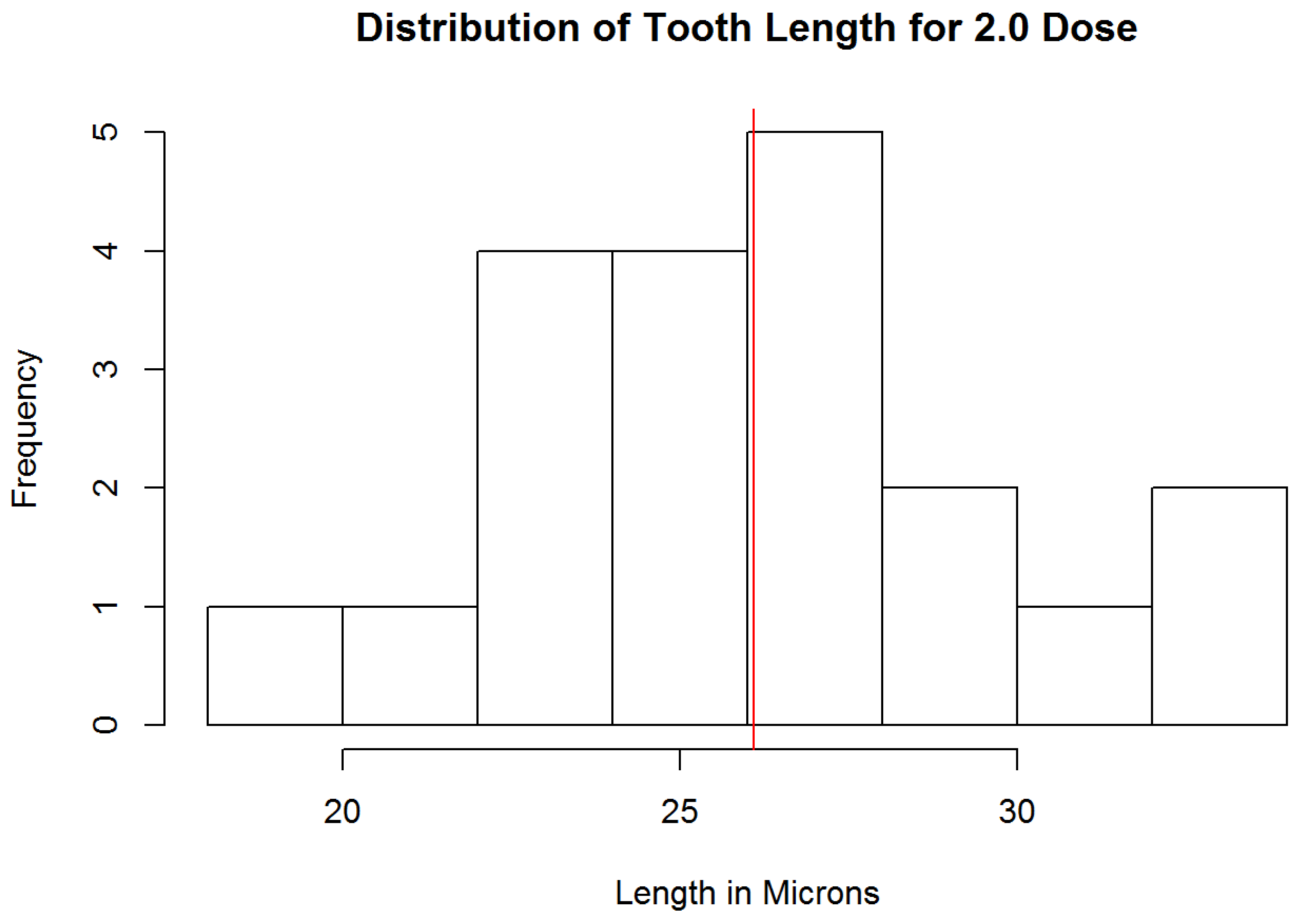




Comparing the tooth lengths between VC and OJ, the VC distribution appears normal-ish, but the OJ distirbution is skewed toward the higher end. Now, let’s look at the distribution of the various dosage levels.


Distribution of Tooth Length for 0.5 Dose
Frequency
0
2
4
6
8
10
0
5
10
15
20
25
Length in Microns


Distribution of Tooth Length for 1.0 Dose
Frequency
0
1
2
3
4
15
20
25
Length in Microns



The .05 distribution is concentrated towards the lower end, with over ten observations between 5 and 10 microns. The 1.0 distribution appears bimodal, with peaks between 15 and 20 microns, then a smaller peak between 20 and 25 microns. The 2.0 distribution is the most normal looking of the three, though does have a little pick up near the top end, with two observations in the 34 micron range. ##Summarize the Data To further understand the tooth length data along by supplement method and dosage, we will find the average tooth length and standard deviation by supplement method, by dosage level and both supplement method and dosage level.

```
growthBySupp <- summarize(group_by(ToothGrowth, supp), mean(len), sd(len))
colnames(growthBySupp) <- c("supp", "meanlen", "sdlen")
growthBySupp
growthByDose <- summarize(group_by(ToothGrowth, dose), mean(len), sd(len))
colnames(growthByDose) <- c("supp", "meanlen", "sdlen")
growthByDose
growthBySuppAndDose <- summarize(group_by(ToothGrowth, supp, dose), mean(len), sd(len))
colnames(growthBySuppAndDose) <- c("supp", "dose", "meanlen", "sdlen")
growthBySuppAndDose
```

Looking at the summary of tooth growth supplement, the mean of of OJ is higher than that of VC and has a tighter standard deviation, which suggests this supplement may be more effective at growing teeth.

Generally, the higher the dose, the greater the tooth growth and the tighter the standard deviation, suggesting that greater doses may be more effective at growing teeth.

Finally, it is interesting to look at the mean tooth growth and standard deviation grouped by both supplement and dose. The higher doses for both supplement types yield greater tooth growth. But what is surprising is that the variability as measured by standard deviation decreases with dose for the OJ supplement but actually increases with dose for the VC supplement. This may mean the VC supplement has more variable (and possibly less reliable) results at higher doses, while the OJ supplement is more variable in smaller doses.

Testing Tooth Growth

The data shows tooth growth by two dimensions: delivery method and doses. Since the sample sizes are relatively small, we'll use the T distribution to compare our hypotheses.

The first hypothesis we'll test is that the supplemental delivery method OJ, the alternative hypothesis, has a greater impact on tooth growth than VC, the null hypothesis. Since the exploratory analysis above suggested that tooth lenght with the OJ supplement tended to have more observations with longer tooth lengths than the VC supplement, let's look deeper into this data.

```
growthVC <- subset(ToothGrowth, supp == "VC", select = "len")
growthOJ <- subset(ToothGrowth, supp == "OJ", select = "len")
t.test(growthOJ - growthVC, var.equal = TRUE)
```

The hypothesis test reveals a 95% confidence interval of 1.4086586, 5.9913414. As this interval is entirely above zero, we can reasonably conclude that the OJ supplement does indeed have a greater impact on tooth growth than the and we can accept the alternative hypothesis.

Now, let's have closer look at the dose data. We'll test two hypotheses: 1. The 1.0 dose has a greater impact on tooth growth than the 0.5 dose. 2. The 2.0 does has a greater impact on tooth growth than the 1.0 dose.

```
growth0.5 <- subset(ToothGrowth, dose == "0.5", select = "len")
growth1.0 <- subset(ToothGrowth, dose == "1", select = "len")
t.test(growth1.0 - growth0.5, var.equal = TRUE)
```

This hypothesis test reveals a 95% confidence interval of 6.3871212, 11.8728788. As the previous exploratory analysis suggested, increasing the dosage by .5 leads to significantly more tooth growth. What about the impact of increasing from 1.0 to 2.0?

```
growth2.0 <- subset(ToothGrowth, dose == "2", select = "len")
t.test(growth2.0 - growth1.0, var.equal = TRUE)
```

This hypothesis test reveals a 95% confidence interval of 3.4718143, 9.2581857. Though this is an increase in tooth growth, it is interesting to note that confidence interval is not as high above 0 as the the hypothesis test from 0.5 to 1.0. Also, the increase in tooth growth was an average of 9 microns greater when adding .5 units from .5 to 1.0, but only 7 microns greater when adding 1 unit, from 1.0 to 2.0.

Conclusions and Assumptions

The exploration and testing of the ToothGrowth data above points us to a few conclusions

1. As a general rule, we can conclude from our that the OJ supplement is more leads to greater tooth growth than the VC supplement assuming vitamin C was the only tooth growing nutrient the subjects ingested during the experiment time period.
2. The OJ supplement is more reliable at higher doses and the VC supplement is more reliable at lower doses assuming standard deviation is a valid measure of reliability for the supplements in question.
3. Continuing to increase the dosage will eventually reach the point of diminishing returns as it pertains to tooth growth, assuming large doses follow the pattern observed above of successively large doses yielding successively smaller tooth growth results.

Appendix

This section contains code for the plots referenced earlier in this analysis.

```
VClen <- subset(ToothGrowth, supp == "VC", select = "len")
hist(VClen$len, main = "Distribution of Tooth Length for VC", xlab = "Length in Microns")
abline(v=mean(VClen$len), col = "red")
```

```
OJlen <- subset(ToothGrowth, supp == "OJ", select = "len")
hist(OJlen$len, main = "Distribution of Tooth Length for OJ", xlab = "Length in Microns")
abline(v=mean(OJlen$len), col = "red")
```

```
zerofivelen <- subset(ToothGrowth, dose == "0.5", select = "len")
hist(zerofivelen$len, main = "Distribution of Tooth Length for 0.5 Dose", xlab = "Length
in Microns")
abline(v=mean(zerofivelen$len), col = "red")
```

```
onezerolen <- subset(ToothGrowth, dose == "1", select = "len")
hist(onezerolen$len, main = "Distribution of Tooth Length for 1.0 Dose", xlab = "Length i
n Microns")
abline(v=mean(onezerolen$len), col = "red")
```

```
twozerolen <- subset(ToothGrowth, dose == "2", select = "len")
hist(twozerolen$len, main = "Distribution of Tooth Length for 2.0 Dose", xlab = "Length i
n Microns")
abline(v=mean(twozerolen$len), col = "red")
```