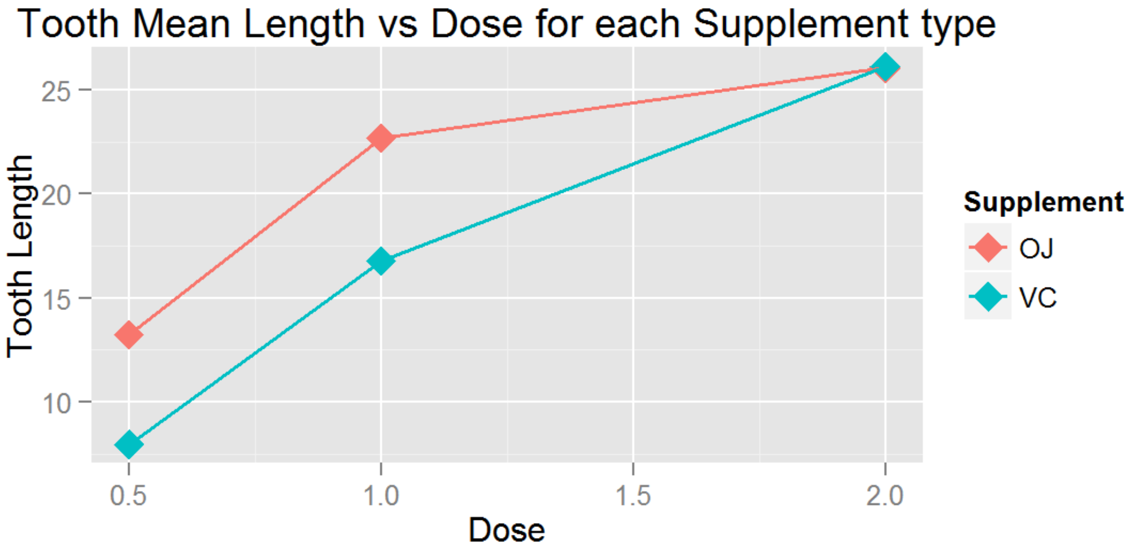
Statistical Inference - Toothgrowth Analysis

M. Nelson

November 17, 2015

Exploratory Data Analysis

This plot is the basis for a brief exploratory data analysis of the ToothGrowth data set. It shows increasing tooth length on the y-axis plotted against the three steps of increasing vitamin C dosage levels on the x-axis given in the data set. Two lines are plotted representing the supplement delivery type: orange juice (OJ) or ascorbic acid (VC). The plot implies that tooth growth may increase within this range of increasing dosage. It also shows that the OJ delivery method may be more effective than the VC method. (Code for the plot is in the Appendix.) These statements will be tested in the more detailed analysis that follows.



Summary of Data Tests

The detailed analysis will test two hypothesis using T-tests with 95% two-sided confidence intervals. These two null hypothesis are:

H0d = There is no difference in tooth length using these different dosages.

H0s = There is no difference in tooth length using these different delivery methods.

Confidence Intervals by Dose

This R code selects all tooth lengths from the lowest dosage (0.5) in one vector (lowdose) and all tooth lengths from the highest dosage (2.0) in another vector (highdose). The result is a confidence interval of 12.8 - 18.2. Because zero is not within this confidence interval, the null hypothesis (H0d) that there is no difference between the means of the two groups receiving different dosages is rejected.

```
lowdose <- ToothGrowth %>% filter(dose == 0.5) %>% select(Length = len)
highdose <- ToothGrowth %>% filter(dose == 2.0) %>% select(Length = len)
t.test(highdose, lowdose, alternative = "two.sided", paired = FALSE, var.equal = FALSE)$c
onf
```

```
## [1] 12.83383 18.15617
## attr(,"conf.level")
## [1] 0.95
```

Confidence Intervals by Supplement Type

This R code selects all tooth lengths from the orange juice delivery in one vector (oj) and all tooth lengths from the ascorbic acid delivery in another vector (vc). The result is a confidence interval of -0.2 - 7.6. Because zero is within this confidence interval, the null hypothesis (H0s) that there is no difference between the means of the two groups receiving different delivery methods cannot be rejected.

```
oj <- ToothGrowth %>% filter(supp == "OJ") %>% select(Length = len)
vc <- ToothGrowth %>% filter(supp == "VC") %>% select(Length = len)
t.test(oj, vc, alternative = "two.sided", paired = FALSE, var.equal = FALSE)$conf
```

```
## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

Conclusion and Assumptions

The first conclusion is that higher dosages of vitamin C within this dosage range does increase the tooth growth of these subjects.

The second conclusion is that the different delivery methods of orange juice or ascorbic acid did not produce statistically different tooth growth.

As the selection method of the test subjects is not known, the variances of each subset of subjects is not assumed to be the same. These are also independent subject groups so this does not represent a paired data test. It was possible for the mean difference to be either above or below the hypothesis that the means are the same. So a two-sided test was used to account for this. Finally the standard confidence interval of 95% was used (Type I error rate of 5%.)

Appendix

Code for EDA Plot:

```
library(dplyr)
library(ggplot2)

data("ToothGrowth")

dosage <- ToothGrowth %>%
    mutate(Supplement = supp) %>%
    group_by(Supplement, dose) %>%
    summarise(tl = mean(len))

p<-ggplot(dosage, aes(x=dose, y=tl, color=Supplement, group=Supplement)) +
    geom_point(shape=18, size=5) +
    geom_line(size=0.6) +
    labs(x="Dose",y="Tooth Length") +
    ggtitle("Tooth Mean Length vs Dose for each Supplement type")

print(p)
```