# Part 2: Basic Inferential Data Analysis

*Erik Johnson*

*10/30/2017*

## Overview

In this report, we analyze the ToothGrowth data in the R datasets package. We show that there is a statistically significant difference in tooth growth between the supplements OJ and VC at the 0.5 and 1.0 dose levels, but there is no significant difference between the supplements at the 2.0 dose level or when compared irrespective of dose.

## Load the data

We start by loading the ToothGrowth data from the R datasets package and displaying some basic summary information about the dataset.

```
library(dplyr)
library(ggplot2)
```

```
data("ToothGrowth")

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 ...
```

We see that the dataset has 3 variables: len (numeric), supp (factor), and dose (numeric).

## Exploratory data analysis

We start our analysis by showing the unique supplement and dose values.

```
unique(ToothGrowth$supp)
```

```
## [1] VC OJ
## Levels: OJ VC
```

```
unique(ToothGrowth$dose)
```

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

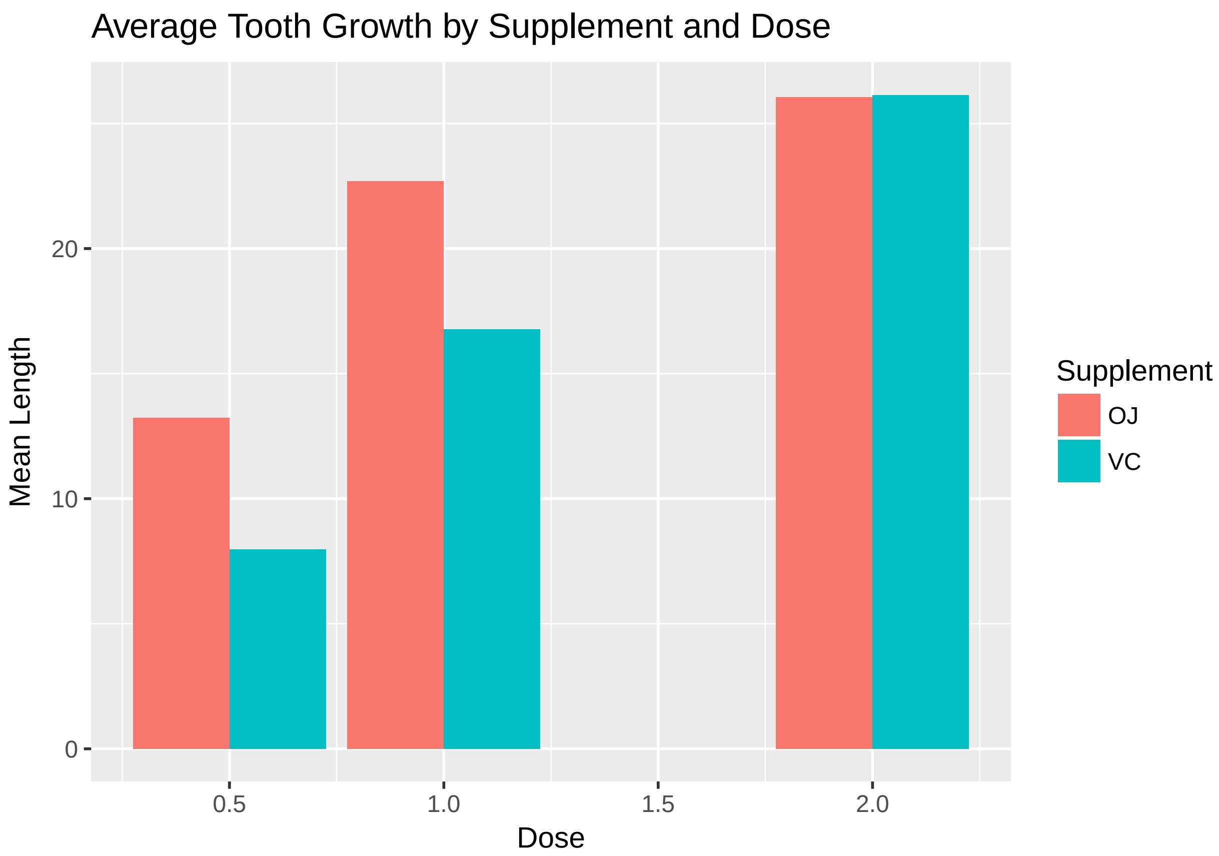
To compare tooth growth by supp and dose, we first group the data by supp and then by dose, and then take the mean tooth growth for each grouping.

```
summarizedToothGrowth <-
    ToothGrowth %>%
    group_by(supp, dose) %>%
    summarize(len.mean = mean(len))
head(summarizedToothGrowth)
```

```
## # A tibble: 6 x 3
## # Groups:   supp [2]
##     supp  dose len.mean
##   <fctr> <dbl>    <dbl>
## 1     OJ   0.5    13.23
## 2     OJ   1.0    22.70
## 3     OJ   2.0    26.06
## 4     VC   0.5     7.98
## 5     VC   1.0    16.77
## 6     VC   2.0    26.14
```

Next, we plot the tooth growth on the y-axis versus dose level, with the bar colors representing the supplement.

```
g <- ggplot(summarizedToothGrowth, aes(x = dose, y = len.mean, fill = supp)) +
    geom_col(position = "dodge") +
    labs(
        x = "Dose",
        y = "Mean Length",
        title = "Average Tooth Growth by Supplement and Dose",
        fill = "Supplement"
    )
print(g)
```



We see that OJ appears to result in higher tooth growth than VC at the 0.5 and 1.0 dose levels. However, at the 2.0 dose level, OJ and VC appear to result in similar tooth growth. In the next section, we explore if these differences are statistically significant.

## Confidence intervals and hypothesis tests

We want to know if the difference in tooth growth that we observed between VC and OJ is significant.

Our null hypothesis is that there is no difference in average tooth growth between VC and OJ at any dose level. Our alternate hypothesis is that average tooth growth is different between OJ and VC. We test this at each dose level, as well as irrespective of dose.

To start, we retrieve the growth length data for each supplement at each dose level, as well as irrespective of dose.

```r
len_OJ_05 <- ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 0.5]
len_OJ_10 <- ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 1.0]
len_OJ_20 <- ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 2.0]
len_VC_05 <- ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose == 0.5]
len_VC_10 <- ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose == 1.0]
len_VC_20 <- ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose == 2.0]
len_OJ <- ToothGrowth$len[ToothGrowth$supp == "OJ"]
len_VC <- ToothGrowth$len[ToothGrowth$supp == "VC"]
```

Next, we perform unpaired two-sided t-tests between the OJ and VC tooth growth data at each dose level as well as irrespective of dose.

```r
t.test(len_OJ_05, len_VC_05)
```

```
##
##  Welch Two Sample t-test
##
## data:  len_OJ_05 and len_VC_05
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean of x mean of y
##     13.23      7.98
```

```r
t.test(len_OJ_10, len_VC_10)
```

```
##
##  Welch Two Sample t-test
##
## data:  len_OJ_10 and len_VC_10
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean of x mean of y
##     22.70     16.77
```

```r
t.test(len_OJ_20, len_VC_20)
```

```
##
##  Welch Two Sample t-test
##
## data:  len_OJ_20 and len_VC_20
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

```
t.test(len_OJ, len_VC)
```

```
##
##  Welch Two Sample t-test
##
## data:  len_OJ and len_VC
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

## Conclusion

We find that average tooth growth for OJ is different than that for VC at the 0.5 and 1.0 dose levels (p = 0.006, 0.001, respectively). This means that the observed difference is likely significant. However, at the 2.0 level, we can't reject the null hypothesis as p=0.96. We also can't reject the null when we compare OJ and VC irrespective of dose, as p = 0.06.

Our alternate hypothesis assumes that OJ and VC result in different tooth growth, but we don't make an assumption as to which results in more or less (therefore we use a two-sided t-test). We assume that the tooth growth means are normally distributed, and we set alpha = 0.05 so that we only reject the null hypothesis for p <= 0.05. We also assume that the variances are unequal since we're comparing different supplements, and that the test subjects across groups are unrelated so we use unpaired t-tests.