Tooth Growth Data Analysis

Christopher Hair

December 10, 2017

Overview

In this article we will explore the `ToothGrowth` data that is included in the `datasets` package in R. This dataset is used to show the effects of Vitamin C on Tooth Growth in Guinea Pigs. We will show some information about what the data looks like and how the data is structured. We can then look at some basic analysis on comparing the type of supplement and dosage of each supplement. Using this analysis we will attempt to draw a hypothesis to compare which component supplement and dosage has the larger impact on the tooth growth.

Summarizing the Data

First we need to load the `datasets` library and then we can look at some of the basic properties of the `ToothGrowth` sample data and how the information is structured as well as some introduction into the types of values we are are going to evaluate in our testing.

```r
library(datasets)

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 ...
```

```r
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```r
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  
```

Basic Analysis

In this section we will provide some introductory analysis of the `ToothGrowth` dataset to look for indicators as to how to further test the data. We will start by providing some simple analysis information about the dataset as a starting point.

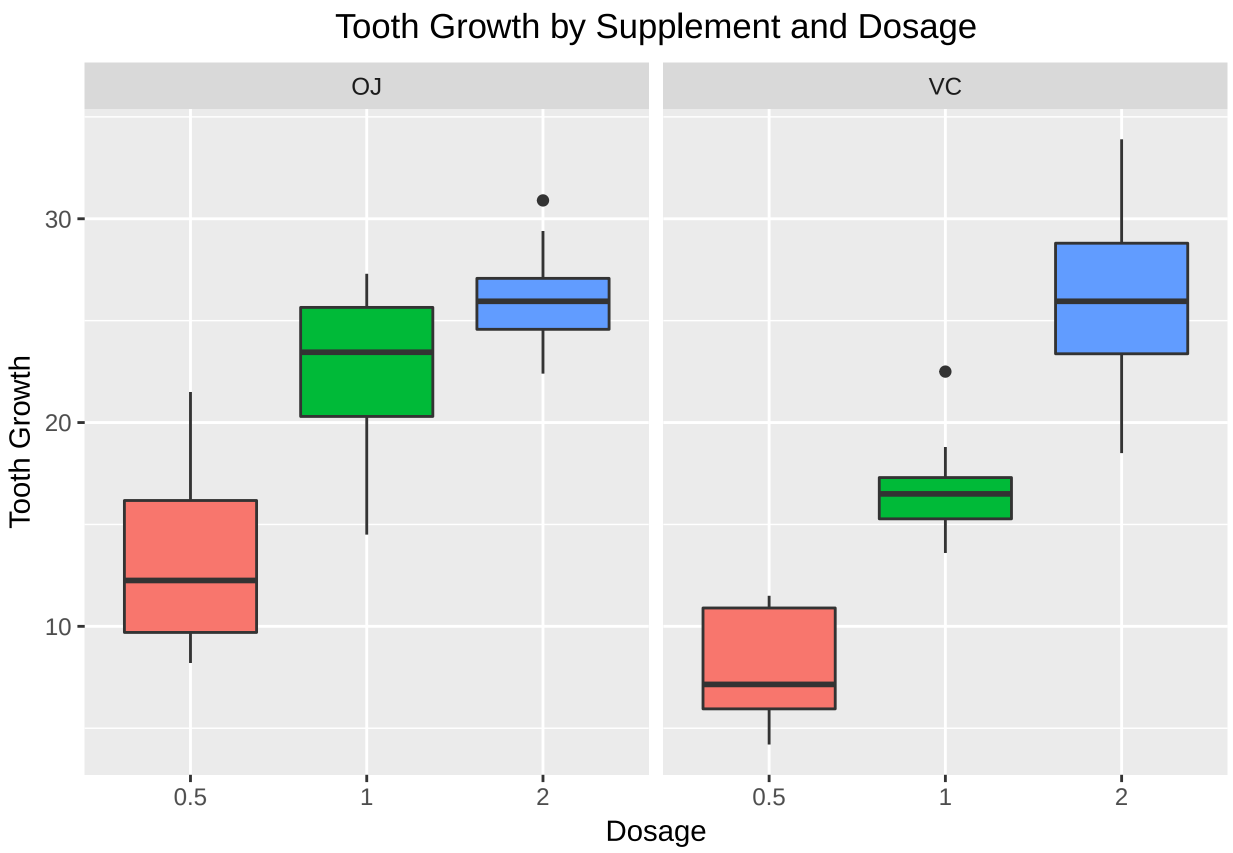
```r
grouped_summay <- group_by(ToothGrowth, supp, dose) %>%
    summarise(group_count=n(),
              group_mean=mean(len),
              group_med=median(len),
              group_sd=sd(len))

as_hux(grouped_summay %>%
           select("Supplement"=supp,
                  "Dosage"=dose,
                  "Count"=group_count,
                  "Mean"=group_mean,
                  "Median"=group_med,
                  "Standard Deviation"=group_sd),
       add_colnames=TRUE) %>%
    set_bold(1, everywhere, TRUE) %>%
    set_all_borders(1) %>%
    set_caption("ToothGrowth Analysis")
```

Table 1: ToothGrowth Analysis

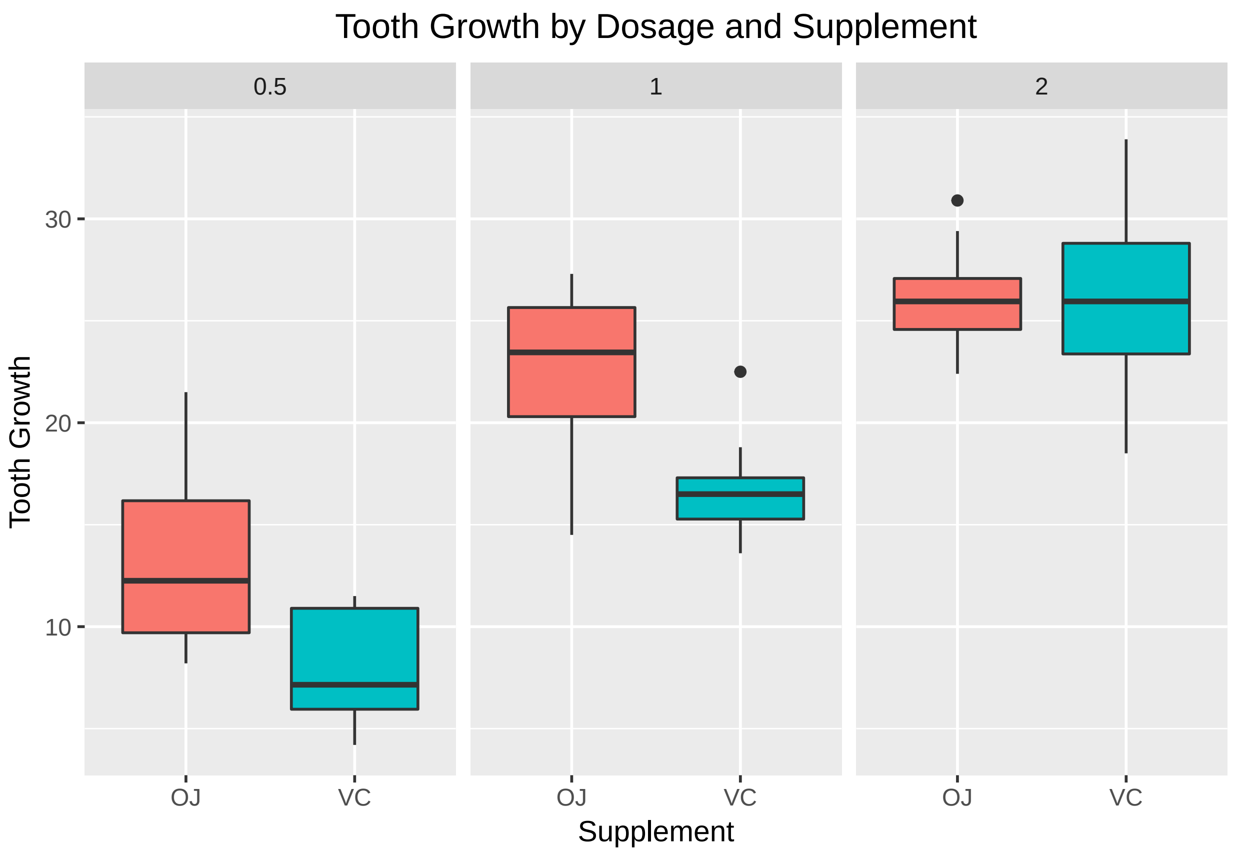
Supplement	**Dosage**	**Count**	**Mean**	**Median**	**Standard Deviation**
OJ	0.50	10.00	13.23	12.25	4.46
OJ	1.00	10.00	22.70	23.45	3.91
OJ	2.00	10.00	26.06	25.95	2.66
VC	0.50	10.00	7.98	7.15	2.75
VC	1.00	10.00	16.77	16.50	2.52
VC	2.00	10.00	26.14	25.95	4.80

This summary shows us that as we increase the dosage of the Orange Juice supplement has a greater chance of success than the Vitamin C supplement. The following plots will show us that whether we look at the data by supplement or by dosage we can see the same results.

```r
ggplot(data = ToothGrowth) +
    geom_boxplot(aes(fill = factor(dose), x = factor(dose), y = len)) +
    facet_grid(.~supp) +
    ggtitle("Tooth Growth by Supplement and Dosage") +
    xlab("Dosage") +
    ylab("Tooth Growth") +
    theme(legend.position="none",
          plot.title = element_text(hjust = 0.5))
```



```r
ggplot(data = ToothGrowth) +
    geom_boxplot(aes(fill = supp, x = supp, y = len)) +
    facet_grid(.~dose) +
    ggtitle("Tooth Growth by Dosage and Supplement") +
    xlab("Supplement") +
    ylab("Tooth Growth") +
    theme(legend.position = "none",
          plot.title = element_text(hjust = 0.5))
```



Hypothesis Testing

From the basic analysis we can now see that we now have the following hypothesis statement to work with.

- $\mathbf{H_0}$ Orange Juice has a greater chance for success on tooth growth in guinea pigs than Vitamin C
- $\mathbf{H_a}$ Orange Juise has a lesser chance for success on tooth growth in guinea pigs than Viatmin C

However, in order to test the hypothesis correctly we need to perform our testing at each of the dosage levels for each supplement. This is done because as the analysis data has also shown there is a difference in the standard deviation at each dosage level. We will use the `t.test` function to compare the supplements at each dosage level to provide us with the information we need to decide if the *null hypothesis* is *TRUE*. In the tests we will assume the standard 95% confidence interval and that these are not paired tests and that the variances are not equal.

Test Summary

Now that we have tested the data we can now summarize the data to see the final results.

```r
small_dose <- subset(ToothGrowth, dose == 0.5)
small_test <- t.test(len ~ supp, data = small_dose, paired = FALSE, var.equal = FALSE)

medium_dose <- subset(ToothGrowth, dose == 1)
medium_test <- t.test(len ~ supp, data = medium_dose, paired = FALSE, var.equal = FALSE)

large_dose <- subset(ToothGrowth, dose == 2)
```

```
large_test <- t.test(len ~ supp, data = large_dose, paired = FALSE, var.equal = FALSE)

dosage <- c("0.5mg/day", "1mg/day", "2mg/day")
p_value <- c(small_test$p.value, medium_test$p.value, large_test$p.value)
ll_value <- c(small_test$conf.int[1], medium_test$conf.int[1], large_test$conf.int[1])
ul_value <- c(small_test$conf.int[2], medium_test$conf.int[2], large_test$conf.int[2])

summary_data <- data.frame(dosage,
                           p_value,
                           ll_value,
                           ul_value,
                           stringsAsFactors = FALSE)

as_huxtable(summary_data %>%
                select("Dosage" = dosage,
                       "P Value" = p_value,
                       "Lower Limit" = ll_value,
                       "Upper Limit" = ul_value),
            add_colnames = TRUE) %>%
    set_number_format(value = 6) %>%
    set_bold(1, everywhere, TRUE) %>%
    set_all_borders(1) %>%
    set_caption("Test Summary Analysis")
```

Table 2: Test Summary Analysis

Dosage	**P Value**	**Lower Limit**	**Upper Limit**
0.5mg/day	0.006359	1.719057	8.780943
1mg/day	0.001038	2.802148	9.057852
2mg/day	0.963852	-3.798070	3.638070

Conclusions

From the final test results we can see that at 0.5mg/day and 1.0mg/day the confience intervals are all above 0 along with a low P value, therefore for these two dosage levels we can say that the *null hypothesis* is *TRUE*. However in the third dosage level of 2.0mg/day the confidence interval contains 0 and we have a very large P value, therefore we can say that the *alternative hypothesis* is *TRUE*.