Analyzing the ToothGrowth Data*

Aliakbar Safilian†

December 13, 2018

1 Overview

We analyze the **ToothGrowth** data in the R **datasets** package. The data is about the effect of vitamin C on tooth growth in guinea pigs. In this data, "the response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of *three dose levels* of vitamin C (**0.5**, **1**, and **2** mg/day) by one of two *delivery methods*, **orange juice** (coded as **OJ**) or **ascorbic acid** (a form of vitamin C and coded as **VC**)."[1] We will further explore the data in the next section.

The structure of the rest of the report is as follows: In Sec. 2, we load the data, and perform some basic summary statistics. Moreover, we provide a basic summary of the data. Sec. 3, we study the affects of dose levels and delivery methods in tooth growth, based on the sample data. In Sec. 4, we study the denisties of the data and some of its subsets. Sec. 5, we compare tooth growth with other elements in the data using by hypothesis tests. Sec. 6 concludes the report with our conclusions and the assumptions needed for them.

2 The Data: Loading and Basic Summary Analysis

Let's first load the data, and take a look at its structure:

```
library(datasets)
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 ...
```

As we see, there are 60 observations with 3 variables in this data. Here is a brief explanation of the variables: **len** denotes the *length* of the growth, **supp** represents the *delivery* (*supplement*) type (either **VC** or **OJ**), and **dose** denotes the *dose* in milligrams/day. We change the names of the variables to **Length**, **Supplement**, and **Dose**, respectively, in the following script:

```
names(ToothGrowth) <- c("Length", "Supplement", "Dose")
```

The following script shows that there are only three unique values for **Dose**: $0.5, 1$, and 2. To be able to make some elegant plots, we transform this variable into an equivalent factor one.

```
unique(ToothGrowth$Dose)
```

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

```
ToothGrowth$Dose <- as.factor(ToothGrowth$Dose)
```

Let us have a summary of the data:

*A part of this analysis report was submitted for the final project-Part2 of the Statistical Inference (Coursera) course at Johns Hopkins University

†Email: a.a.safilian@gmail.com

[1] https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ToothGrowth.html

```
summary(ToothGrowth)
```

```
##       Length      Supplement  Dose
##  Min.   : 4.20   OJ:30      0.5:20
##  1st Qu.:13.07   VC:30      1  :20
##  Median :19.25              2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

3 Growth Length by Dose Levels & Delivery Methods

In the following, we extract the mean, standard deviation, and sum of growth length for each dose level:

```
library(dplyr)
sum_dose <- as.data.frame(ToothGrowth %>%
        group_by(Dose) %>%
        summarize(Mean = mean(Length), SD = sd(Length), Sum = sum(Length)))
sum_dose
```

```
##   Dose   Mean       SD   Sum
## 1  0.5 10.605 4.499763 212.1
## 2    1 19.735 4.415436 394.7
## 3    2 26.100 3.774150 522.0
```

The following script extracts the mean, the standard deviation, and sum of growth length for each delivery method:

```
sum_supp <- as.data.frame(ToothGrowth %>%
        group_by(Supplement) %>%
        summarize(Mean = mean(Length), SD = sd(Length), Sum = sum(Length)))
sum_supp
```

```
##   Supplement     Mean       SD   Sum
## 1         OJ 20.66333 6.605561 619.9
## 2         VC 16.96333 8.266029 508.9
```

The following script extracts the mean, the standard deviation, and sum of growth length for each pair of dose level and delivery method:

```
sum_supp_dose <- as.data.frame(ToothGrowth %>%
        group_by(Supplement, Dose) %>%
        summarize(Mean = mean(Length), SD = sd(Length), Sum = sum(Length)))
sum_supp_dose
```

```
##   Supplement Dose  Mean       SD   Sum
## 1         OJ  0.5 13.23 4.459709 132.3
## 2         OJ    1 22.70 3.910953 227.0
## 3         OJ    2 26.06 2.655058 260.6
## 4         VC  0.5  7.98 2.746634  79.8
## 5         VC    1 16.77 2.515309 167.7
## 6         VC    2 26.14 4.797731 261.4
```

In the rest of this seciton, we do some relevant exploratory analyses.

Fig. 1 represents the relationship between growth length and dose levels categorized by delivery methods.

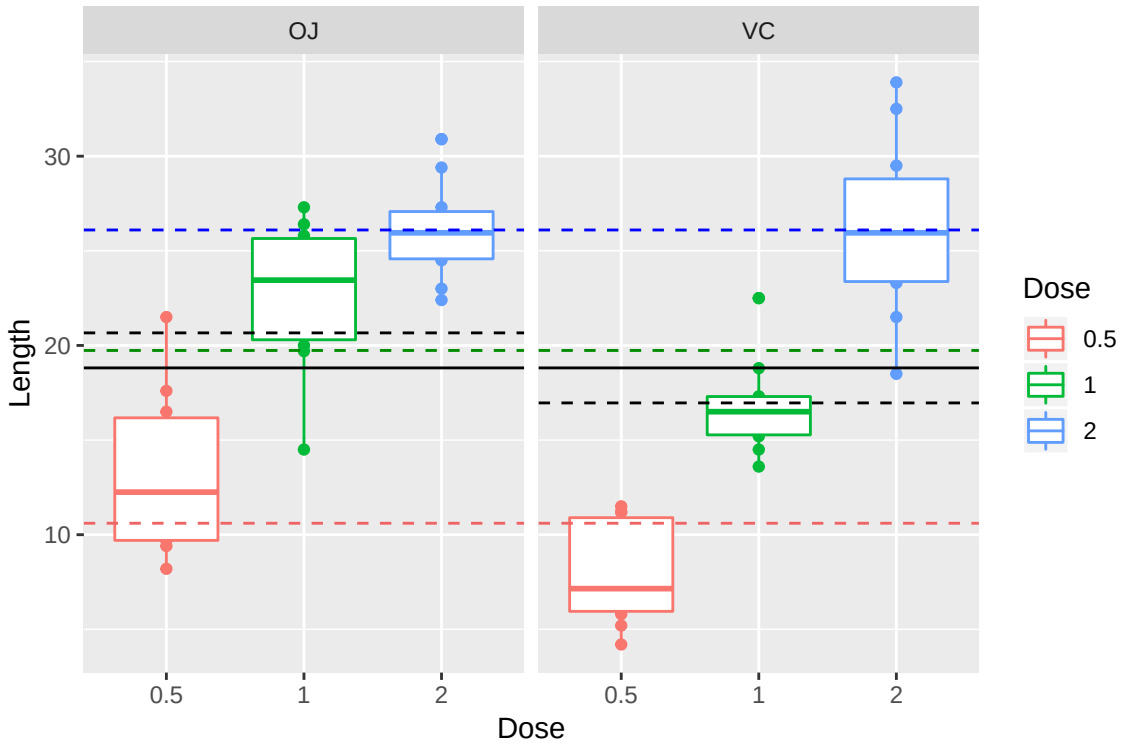

Figure 1: Growth by Dose Levels and Delivery Metods

The *black solid line* represents the mean of growth (i.e., ~**18.81**) in the whole data. A *dashed black* line represents the mean of growth for the associated delivery method (i.e., ~**20.66** and ~**16.96** for methods OJ and VC, respectively). The *red*, *green*, and *blue dashed* lines denote the mean of growth for dose levels 0.5, 1, and 2, respectively (i.e., ~**10.61**, ~**19.73**, and ~**26.1**, respectevily).

```
library(ggplot2)
x <- sum_supp[, 1:2]
m <- mean(ToothGrowth$Length)
qplot(Dose, Length, data = ToothGrowth, color = Dose, facets = .~Supplement) + geom_boxplot() +
geom_hline(yintercept = m, linetype = "solid") +
geom_hline(aes(yintercept = Mean), x, linetype = "dashed") +
geom_hline(yintercept = sum_dose[1, 2], color = "indianred2", linetype = "dashed") +
geom_hline(yintercept = sum_dose[3, 2], color = "blue", linetype = "dashed") +
geom_hline(yintercept = sum_dose[2, 2], color = "green4", linetype = "dashed")
```

Some (interesting) observations from the above figure are as follow:

- As expected, more dose levels results in more growth.
- The delivery method OJ has more positive impact on the growth than VC.
- Almost all growth length for dose level 0.5 (2, respectively) are under (above, respectively) the average growth in the data.
- Almost all growth length for dose level 1 with delivery method OJ are above the average growth, while this is just the other way round for dose 1 with supplement VC.
- The average growth for dose level 1 is very close to the overall average growth, while the average growth for dose level 2 (0.5, respectively) is above (under, respectively) of the overall average growth.

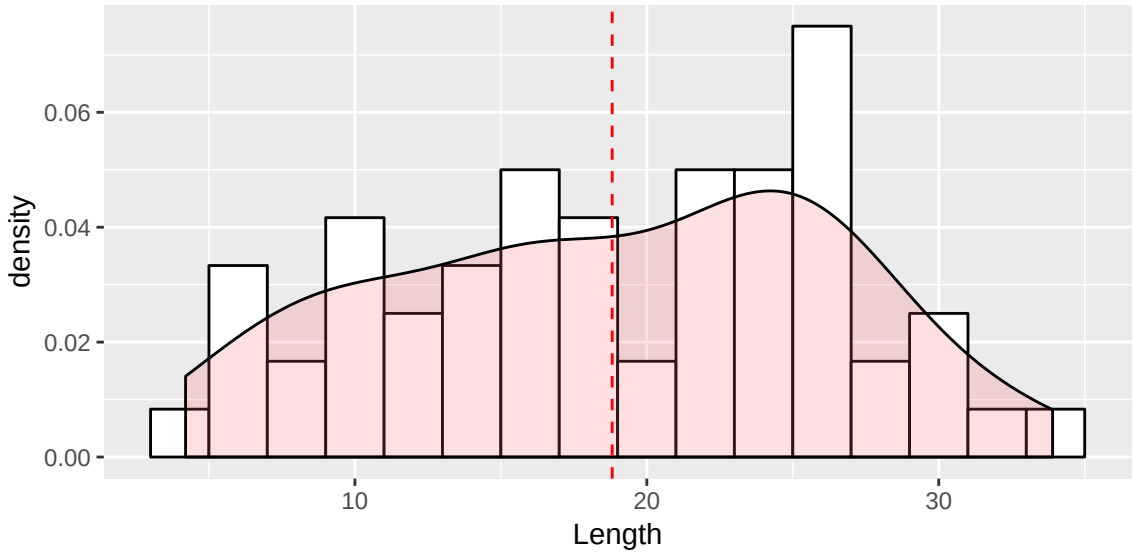

Figure 2: The Density Distribution of Growth Length

4 Density Distribution

In this section, we study the denisty distribution of the data and some of its subsets. The density distribution of growth length in the whole data looks like the curve in Fig. 2, where its mean and standard deviation are approximately **18.81** and **7.65**, respecvtivly. The red dashed line represents the mean of the distribution.

```
ggplot(ToothGrowth, aes(x=Length)) +
        geom_histogram(aes(y=..density..), binwidth=2, colour="black", fill="white") +
        geom_density(alpha=.2, fill="#FF6666") +
        geom_vline(xintercept = m, color = "red", linetype = "dashed")
```

In Fig. 3, we see the density of growth length for each dose level. The mean and standard deviation of growth for dose level 0.5 are approximately **10.61** and **4.5**, respecvtivly. They are approximately **19.73** and **4.42** for dose level 1, and **26.1** and **3.77** for dose level 2. In Fig. 3, the mean of the densities are represented with red dashed lines.

```
y <- sum_dose[, 1:2]
ggplot(ToothGrowth, aes(x=Length)) +
        geom_histogram(aes(y=..density..), binwidth=2, colour="black", fill="white") +
        geom_density(alpha=.2, fill="#FF6666") +
        facet_grid(facets = .~Dose) +
        geom_vline(aes(xintercept = Mean), y, color = "red", linetype = "dashed")
```

Fig. 4 represents the density distribution of growth length for each delivery method. The mean and the standard deviation of growth for the delivery method OJ (VC, respectively) are approximately **20.66** and **6.61** (**16.96** and **8.27**, respectively), respecvtivly. The mean of the densities are represented with red dashed lines.

```
z <- sum_supp[, 1:2]
ggplot(ToothGrowth, aes(x=Length)) +
        geom_histogram(aes(y=..density..), binwidth=2, colour="black", fill="white") +
        geom_density(alpha=.2, fill="#FF6666") +
        facet_grid(facets = .~Supplement) +
        geom_vline(aes(xintercept = Mean), z, color = "red", linetype = "dashed")
```

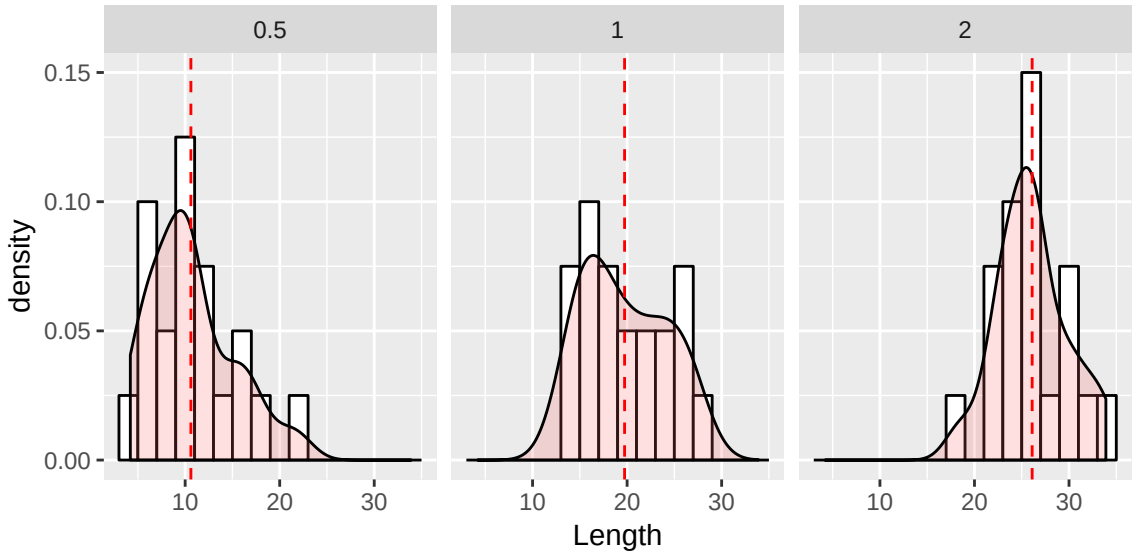

Figure 3: The Density Distribution of Growth Length by Dose Levels

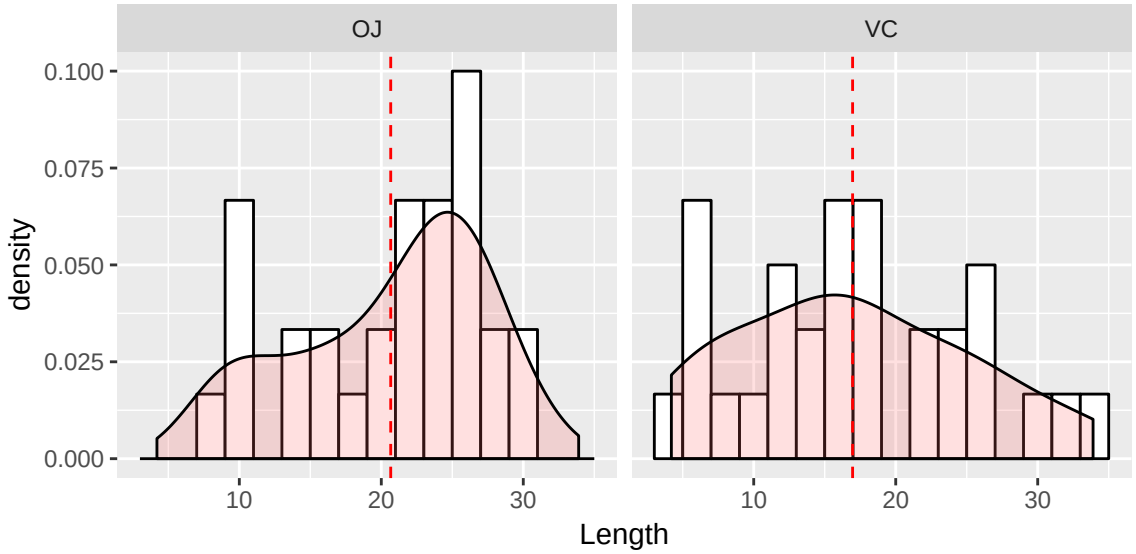

Figure 4: The Density Distribution of Growth Length by Delivery Methods

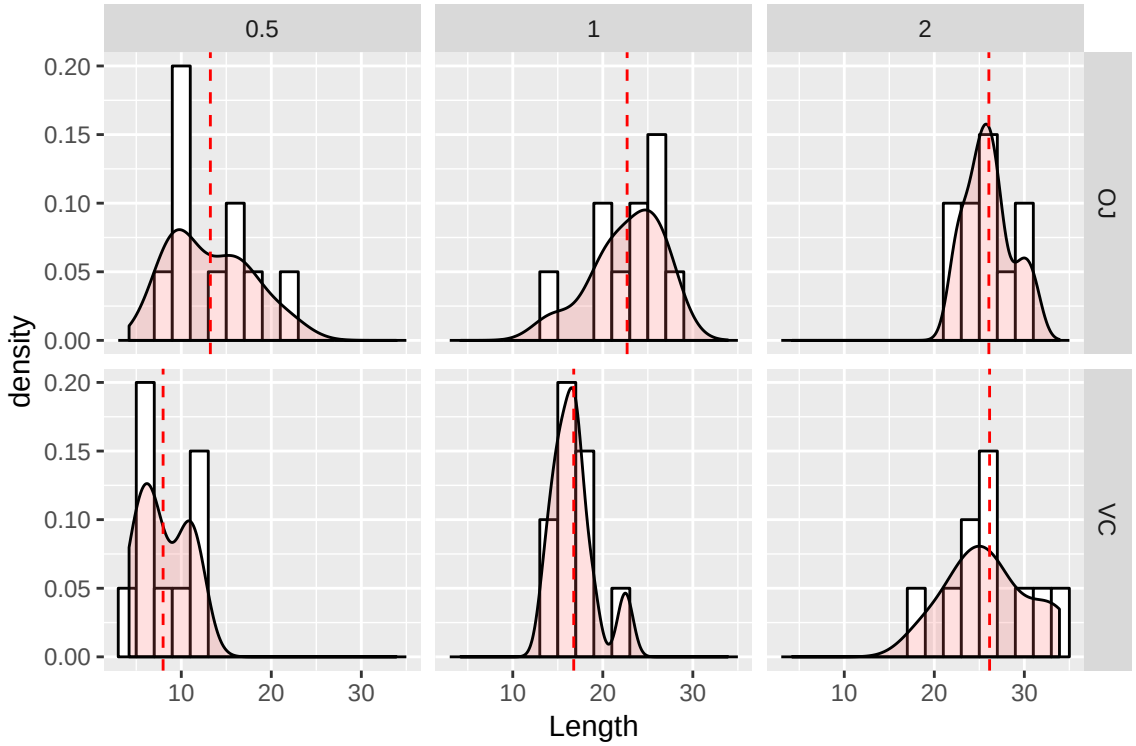

Figure 5: The Density Distribution of Growth Length by Dose Levels and Delivery Methods

Finally, we see, in the Fig. 5, the density distribution of growth length for each pair of delivery methods and dose levels. The mean and the standard deviation of each of these pairs have been already represented in the data frame **sum_supp_dose** (see Sec. 3). The mean of the densities are represented with red dashed lines.

```
v <- sum_supp_dose[, 1:3]
ggplot(ToothGrowth, aes(x=Length)) +
        geom_histogram(aes(y=..density..), binwidth=2, colour="black", fill="white") +
        geom_density(alpha=.2, fill="#FF6666") +
        facet_grid(facets = Supplement~Dose) +
        geom_vline(aes(xintercept = Mean), v, color = "red", linetype = "dashed")
```

5 Hypothesis Tests

In our data, there are two delivery methods (OJ and VC) and three different doses (0.5, 1, and 2). In the follwing, we see the mean of growth length for each pair of (delivery method, dose level):

```
xtabs(Length/10 ~ Supplement + Dose, data = ToothGrowth)
```

```
##            Dose
## Supplement   0.5     1     2
##         OJ 13.23 22.70 26.06
##         VC  7.98 16.77 26.14
```

Therefore, there are 15 potential comparisons of means (hece 15 potential hypothesis tests). In the following, we get the correspodning sub datasets.

```
data_OJ_5 <- ToothGrowth %>% filter(Dose == "0.5" & Supplement == "OJ") %>% select(Length)
data_OJ_1 <- ToothGrowth %>% filter(Dose == "1" & Supplement == "OJ") %>% select(Length)
data_OJ_2 <- ToothGrowth %>% filter(Dose == "2" & Supplement == "OJ") %>% select(Length)
data_VC_5 <- ToothGrowth %>% filter(Dose == "0.5" & Supplement == "VC") %>% select(Length)
data_VC_1 <- ToothGrowth %>% filter(Dose == "1" & Supplement == "VC") %>% select(Length)
data_VC_2 <- ToothGrowth %>% filter(Dose == "2" & Supplement == "VC") %>% select(Length)
```

We test 15 possible null hypotheses by getting the corresponding p-values. The p-values are saved into a vector variable named **pvalues**.

```
pvalues <- NULL
```

We use the notation μ_d^s, where $s \in \{OJ, VC\}, d \in \{0.5, 1, 2\}$, denotes the population mean of growth length for the given deleviry method s and dose d.

Hypothesis Test 1:

$$H_0 : \mu_{0.5}^{OJ} = \mu_{1.0}^{OJ}$$
$$H_a : \mu_{0.5}^{OJ} < \mu_{1.0}^{OJ}$$

```
test1 <- t.test(x = data_OJ_5, y = data_OJ_1, alternative = "less")
pvalues <- c(pvalues, test1$p.value)
```

Hypothesis Test 2:

$$H_0 : \mu_{0.5}^{OJ} = \mu_{2.0}^{OJ}$$
$$H_a : \mu_{0.5}^{OJ} < \mu_{2.0}^{OJ}$$

```
test2 <- t.test(x = data_OJ_5, y = data_OJ_2, alternative = "less")
pvalues <- c(pvalues, test2$p.value)
```

Hypothesis Test 3:

$$H_0 : \mu_{1.0}^{OJ} = \mu_{2.0}^{OJ}$$
$$H_a : \mu_{1.0}^{OJ} \neq \mu_{2.0}^{OJ}$$

```
test3 <- t.test(x = data_OJ_1, y = data_OJ_2)
pvalues <- c(pvalues, test3$p.value)
```

Hypothesis Test 4:

$$H_0 : \mu_{0.5}^{VC} = \mu_{1.0}^{VC}$$
$$H_a : \mu_{0.5}^{VC} < \mu_{1.0}^{VC}$$

```
test4 <- t.test(x = data_VC_5, y = data_VC_1, alternative = "less")
pvalues <- c(pvalues, test4$p.value)
```

Hypothesis Test 5:

$$H_0 : \mu_{0.5}^{VC} = \mu_{2.0}^{VC}$$
$$H_a : \mu_{0.5}^{VC} < \mu_{2.0}^{VC}$$

```
test5 <- t.test(x = data_VC_5, y = data_VC_2, alternative = "less")
pvalues <- c(pvalues, test5$p.value)
```

Hypothesis Test 6:

$$H_0 : \mu_{1.0}^{VC} = \mu_{2.0}^{VC}$$
$$H_a : \mu_{1.0}^{VC} < \mu_{2.0}^{VC}$$

```
test6 <- t.test(x = data_VC_1, y = data_VC_2, alternative = "less")
pvalues <- c(pvalues, test6$p.value)
```

Hypothesis Test 7:

$$H_0 : \mu_{0.5}^{OJ} = \mu_{0.5}^{VC}$$
$$H_a : \mu_{0.5}^{OJ} < \mu_{0.5}^{VC}$$

```
test7 <- t.test(x = data_VC_1, y = data_VC_2, alternative = "less")
pvalues <- c(pvalues, test7$p.value)
```

Hypothesis Test 8:

$$H_0 : \mu_{0.5}^{OJ} = \mu_{1.0}^{VC}$$
$$H_a : \mu_{0.5}^{OJ} \neq \mu_{1.0}^{VC}$$

```
test8 <- t.test(x = data_OJ_5, y = data_VC_1)
pvalues <- c(pvalues, test8$p.value)
```

Hypothesis Test 9:

$$H_0 : \mu_{0.5}^{OJ} = \mu_{2.0}^{VC}$$
$$H_a : \mu_{0.5}^{OJ} < \mu_{2.0}^{VC}$$

```
test9 <- t.test(x = data_OJ_5, y = data_VC_2, alternative = "less")
pvalues <- c(pvalues, test9$p.value)
```

Hypothesis Test 10:

$$H_0 : \mu_{1.0}^{OJ} = \mu_{0.5}^{VC}$$
$$H_a : \mu_{1.0}^{OJ} > \mu_{0.5}^{VC}$$

```
test10 <- t.test(x = data_OJ_1, y = data_VC_5, alternative = "greater")
pvalues <- c(pvalues, test10$p.value)
```

Hypothesis Test 11:

$$H_0 : \mu_{1.0}^{OJ} = \mu_{1.0}^{VC}$$
$$H_a : \mu_{1.0}^{OJ} > \mu_{1.0}^{VC}$$

```
test11 <- t.test(x = data_OJ_1, y = data_VC_1, alternative = "greater")
pvalues <- c(pvalues, test11$p.value)
```

Hypothesis Test 12:

$$H_0 : \mu_{1.0}^{OJ} = \mu_{2.0}^{VC}$$
$$H_a : \mu_{1.0}^{OJ} \neq \mu_{2.0}^{VC}$$

```
test12 <- t.test(x = data_OJ_1, y = data_VC_2)
pvalues <- c(pvalues, test12$p.value)
```

Hypothesis Test 13:

$$H_0 : \mu_{2.0}^{OJ} = \mu_{0.5}^{VC}$$
$$H_a : \mu_{2.0}^{OJ} > \mu_{0.5}^{VC}$$

```
test13 <- t.test(x = data_OJ_2, y = data_VC_5, alternative = "greater")
pvalues <- c(pvalues, test13$p.value)
```

Hypothesis Test 14:

$$H_0 : \mu_{2.0}^{OJ} = \mu_{1.0}^{VC}$$
$$H_a : \mu_{2.0}^{OJ} > \mu_{1.0}^{VC}$$

```
test14 <- t.test(x = data_OJ_2, y = data_VC_1, alternative = "greater")
pvalues <- c(pvalues, test14$p.value)
```

Hypothesis Test 15:

$$H_0 : \mu_{2.0}^{OJ} = \mu_{2.0}^{VC}$$
$$H_a : \mu_{2.0}^{OJ} \neq \mu_{2.0}^{VC}$$

```
test15 <- t.test(x = data_OJ_2, y = data_VC_2)
pvalues <- c(pvalues, test15$p.value)
```

Now, let us take a look at our p-values. They have been rounded with 2 digits:

```
round(pvalues, 2)
```

```
##  [1] 0.00 0.00 0.04 0.00 0.00 0.00 0.00 0.05 0.00 0.00 0.00 0.10 0.00 0.00
## [15] 0.96
```

Since we have done multiple hypothesis tests, we need to adjust our p-values. In the following, we do so using by the Bonferronni method:

```
pvalues <- p.adjust(pvalues, method = "bonferroni")
round(pvalues, 2)
```

```
##  [1] 0.00 0.00 0.59 0.00 0.00 0.00 0.00 0.69 0.00 0.00 0.01 1.00 0.00 0.00
## [15] 1.00
```

Therefore, we reject the null hypothesis for the following tests:

```
which(pvalues < 0.05)
```

```
##  [1]  1  2  4  5  6  7  9 10 11 13 14
```

And we failed to reject the null hypothesis of the follwoing tests with 95% confidence:

```
which(pvalues >= 0.05)
```

```
## [1]  3  8 12 15
```

That is, with 95% confidence, we choose the following hypotheses:

- $\mu_{0.5}^{OJ} < \mu_{1.0}^{OJ}$ (the alternative hypothesis of Test 1).
- $\mu_{0.5}^{OJ} < \mu_{2.0}^{OJ}$ (the alternative hypothese of Test 2).
- $\mu_{1.0}^{OJ} = \mu_{2.0}^{OJ}$ (the null hypothese of Test 3).
- $\mu_{0.5}^{VC} < \mu_{1.0}^{VC}$ (the alternative hypothese of Test 4).
- $\mu_{0.5}^{VC} < \mu_{2.0}^{VC}$ (the alternative hypothese of Test 5).
- $\mu_{1.0}^{VC} < \mu_{2.0}^{VC}$ (the alternative hypothese of Test 6).
- $\mu_{0.5}^{OJ} < \mu_{0.5}^{VC}$ (the alternative hypothese of Test 7).
- $\mu_{0.5}^{OJ} = \mu_{1.0}^{VC}$ (the null hypothesis of Test 8)

- $\mu_{0.5}^{OJ} < \mu_{2.0}^{VC}$ (the alternative hypothese of Test 9).
- $\mu_{1.0}^{OJ} > \mu_{0.5}^{VC}$ (the alternative hypothesis of Test 10)
- $\mu_{1.0}^{OJ} > \mu_{1.0}^{VC}$ (the alternative hypothesis of Test 11)
- $\mu_{1.0}^{OJ} = \mu_{2.0}^{VC}$ (the null hypothesis of Test 12)
- $\mu_{2.0}^{OJ} > \mu_{0.5}^{VC}$ (the alternative hypothesis of Test 13)
- $\mu_{2.0}^{OJ} > \mu_{1.0}^{VC}$ (the alternative hypothesis of Test 14)
- $\mu_{2.0}^{OJ} = \mu_{2.0}^{VC}$ (the null hypothesis of Test 15)

6 Conclusions

In this report, we have studies the ToothGrowth dataset. We have provided some basic summary analyses of the data. We have also provided some basic explorary analyses. Moreover, we have provided the density distributions of the growth length for all combinations of dose levels and delivery methods. Finally, we have done a mutiple hypothesis tests. A summary of our results is as follows:

- As expected, more dose levels results in more growth.
- The delivery method OJ has more positive impact on the growth than VC.

See our hypothesis tests (and oher sections) for a detailed analysis.