Analyzing ToothGrowth data in R dataset packages

Araks Stepanyan

7/20/2017

Executive Summary

This report is created for the final project of the Statistical Inference course offered by Johns Hopkins University through Coursera. We are going to explore ToothGrowth data in R dataset packages.

Our objective is to use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose(only using the techniques from class, even if there are other approaches worth considering).

The documentation of the data set says: *"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 or ascorbic acid (a form of vitamin C and coded as VC)."*

4 major parts of the report are
Tooth growth

a. by supp when we ignore the dose,
b. by dose when we ignore the supp,
c. by dose when we account for supp,
d. by supp when we account for dose,

We are correcting α level, whenever there are multiple comparisons in order to avoid increased type 1 error.

We find that

a. when we ignore the dose, we don't have enough evidence to say that the tooth growth in guinea pigs receiving orange juice is different from the tooth growth of the guinea pigs receiving ascorbic acid.
b. when we ignore the type of supplement, we reject the null hypothesis, which says that there are no differences in tooth growth depending on the amount of dose received.
c. when we account for the type of the supplement, we again reject the null hypothesis that there are no differences in tooth growth depending on the amount of dose received. **Except**, now we don't have enough evidence to reject the hypothesis that the tooth growth of the guinea pigs receiving orange juice of dose 1 is different from the tooth growth of guinea pigs receiving orange juice of dose 2.
d. when we account for the dose, we reject the null hypothesis that there is no difference in tooth growth depending on the type of the supplement received. **Except**, we don't have enough evidence to reject the hypothesis that the tooth growth of guinea pigs receiving orange juice of dose 2 is different from the tooth growth of guinea pigs receiving ascorbic acid of dose 2.

ToothGrowth data

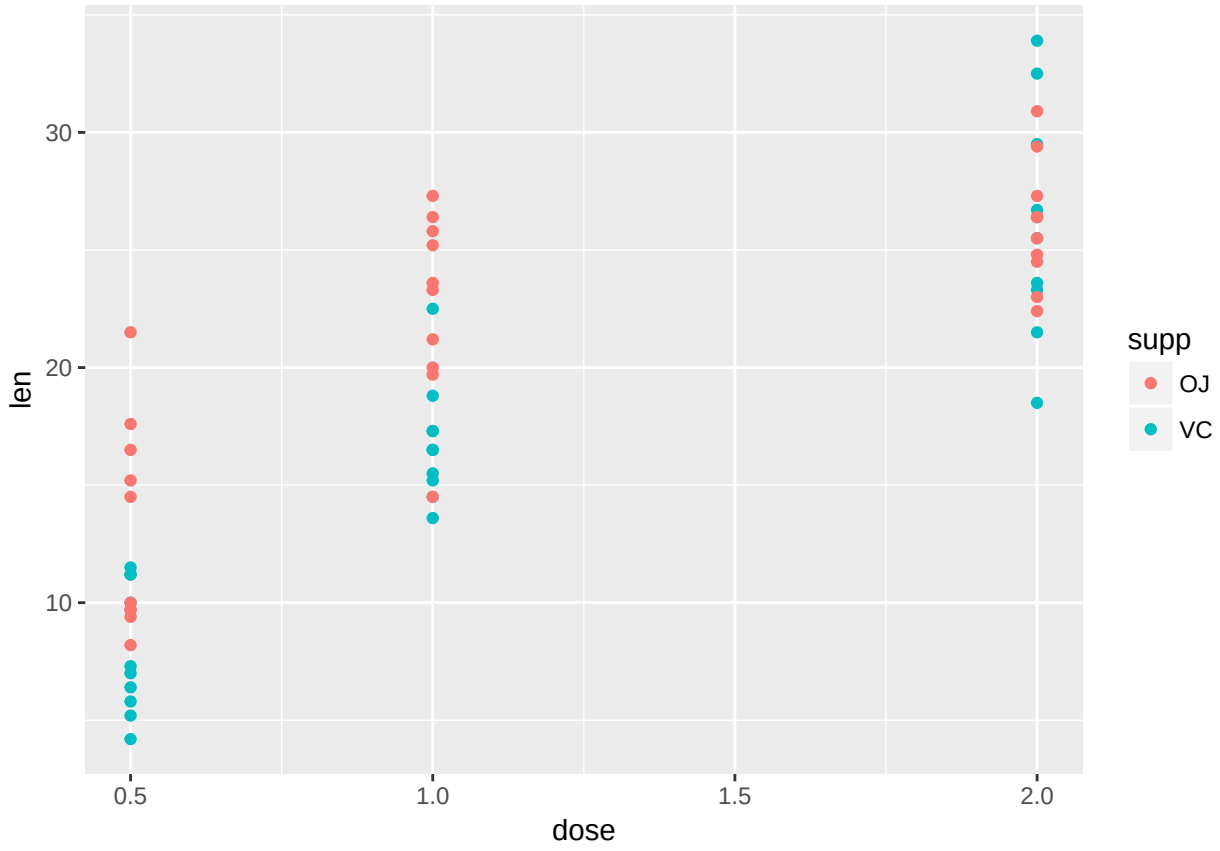
Let's load and look at ToothGrowth data

```
library(ggplot2); library(dplyr); 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 have a data frame with three variables (**len**, **supp**, **dose**). **len** and **dose** are numeric, and **supp** is a factor with two levels. Also, from the plot bellow, we see that there are three values of dose and as the dose increases, we observe higher values of tooth growth (although dose 1 and dose 2 of orange juice (OJ) doesn't look extremely different).

```
qplot(dose, len, color = supp, data = ToothGrowth)
```



Let's start our analysis!

a. Toothe Growth by supp Ignoring the dose (len (OJ) vs len (VC))

From ToothGrowth data let's create two data frames, one for each supp.

```
oj.df <- ToothGrowth %>% filter(supp == "OJ")
vc.df <- ToothGrowth %>% filter(supp == "VC")
dim(oj.df)
```

```
## [1] 30  3
```

```
dim(vc.df)
```

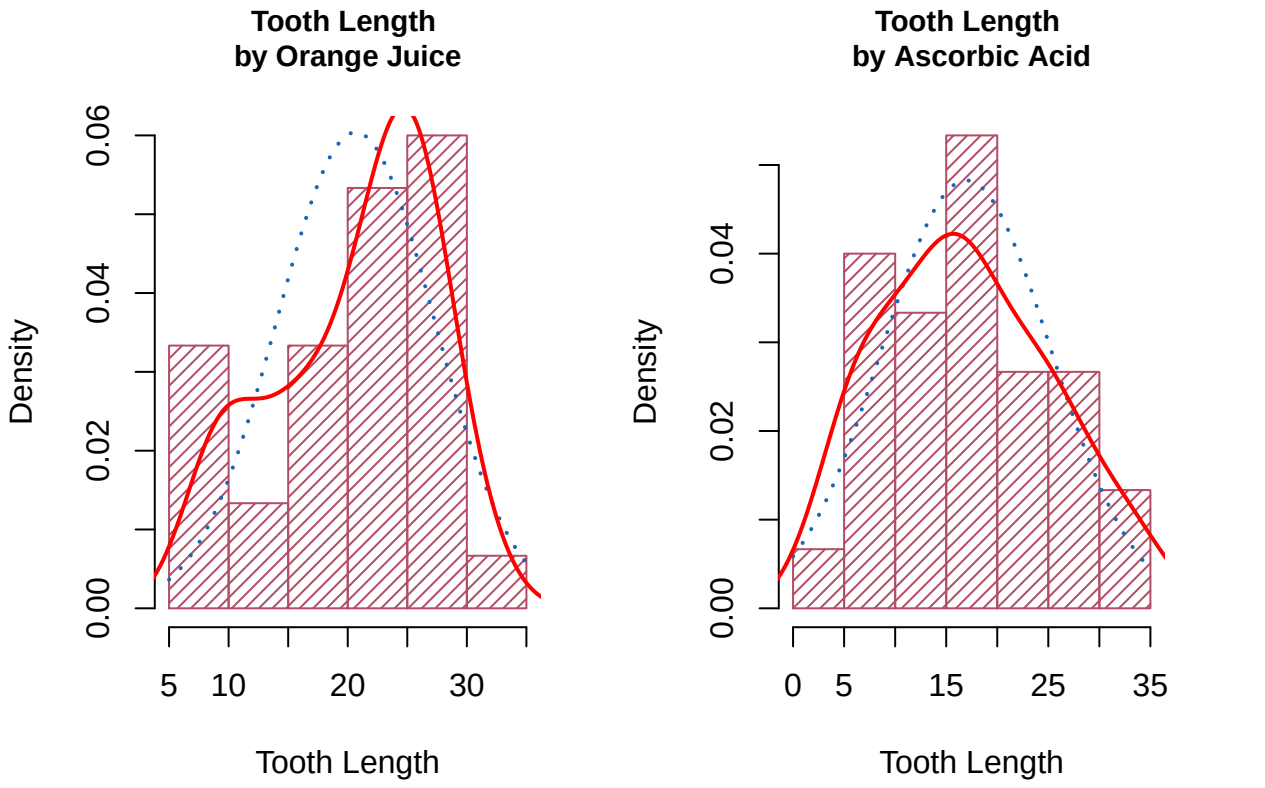
```
## [1] 30  3
```

We have two samples, each of size 30. These sample sizes are not large enough to use CLT (Central Limit Theorem). Then we will use the t statistics for our hypothesis. But first, let's check the conditions for using t

statistic.

- We assume data in ToothGrowth are independent from each other, then the **independence condition is met**.
- We also need to check that the data are **not very skewed** and **don't have outlyers**. For that let's observe the graphs of OJ and VC. We will look at their distributions along with their estimated density curves and the overlayed normal curves.

```
par(mfrow = c(1, 2))
hist(oj.df$len, density = 20, col = rgb(0.7,0.3,0.4),
     prob = TRUE, xlab = "Tooth Length",
     main = "Tooth Length \nby Orange Juice", cex.main = 0.9)
curve(dnorm(x, mean=mean(oj.df$len), sd=sd(oj.df$len)),
      col = rgb(0.1,0.4,0.7),
      lty = "dotted", lwd=2, add=TRUE)
lines(density(oj.df$len, adjust = 1), col = "red", lwd = 2)
hist(vc.df$len, density = 20, col = rgb(0.7,0.3,0.4),
     prob = TRUE, xlab = "Tooth Length",
     main = "Tooth Length \nby Ascorbic Acid", cex.main = 0.9)
curve(dnorm(x, mean=mean(vc.df$len), sd=sd(vc.df$len)),
      col = rgb(0.1,0.4,0.7),
      lty = "dotted", lwd=2, add=TRUE)
lines(density(vc.df$len, adjust = 1), col = "red", lwd = 2)
```



The red curves are the density estimates of our data, while the blue dotted curves are the normal curves. The data are not **very skewed** and **don't have outlyers**, which means that the second contition is also met and we can move on to the hypothesis testing.

- We will consider $\alpha = 0.05$
- We are testing $H_0 : \mu_1 - \mu_2 = 0$ versus $H_a : \mu_1 - \mu_2 \neq 0$
  - μ_1 is the mean (population mean) tooth growth based on OJ

- μ_2 is the mean (population mean) tooth growth based on VC

```
p.value.oj_vc <- t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = ToothGrowth)$p.value

ifelse(p.value.oj_vc < 0.05,
       paste("P value, OJ vs VC (ignoring dose):",
             round(p.value.oj_vc, 10),"<",0.05),
       paste("P value, OJ vs VC (ignoring dose):",
             round(p.value.oj_vc, 10))
       )
```

```
## [1] "P value, OJ vs VC (ignoring dose): 0.0606345079"
```

The p-value is greater than (although close to) 0.05 α level, so we fail to reject the null hypothesis that there is a difference in tooth growth based on the type of the supplement (when we ignore the dose).

b. Toothe Growth by dose Ignoring supp

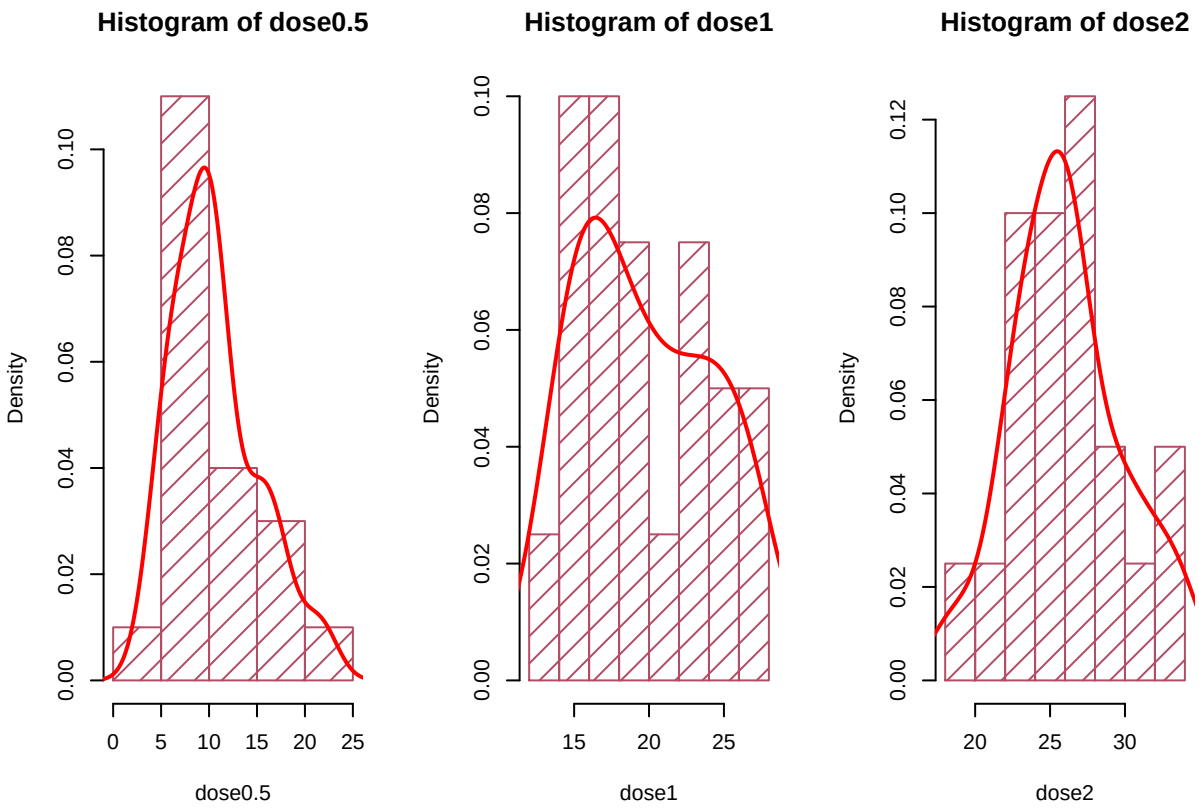
We are going to make 3 comparisons.

1. dose 0.5 vs dose 1,
2. dose 0.5 vs dose 2,
3. dose 1 vs dose 2.

We will separate doses to look at their distributions and decide whether or not we can use t confidence intervals/t tests.

```
dose0.5_df <- ToothGrowth %>% filter(dose == 0.5)
dose1_df <- ToothGrowth %>% filter(dose == 1)
dose2_df <- ToothGrowth %>% filter(dose == 2)
dose0.5 <- dose0.5_df$len
dose1 <- dose1_df$len
dose2 <- dose2_df$len
```

```
par(mfrow = c(1, 3))
hist(dose0.5, density = 10, col = rgb(0.7,0.3,0.4), prob = TRUE)
lines(density(dose0.5, adjust = 1), col = "red", lwd = 2)
hist(dose1, density = 10, col = rgb(0.7,0.3,0.4), prob = TRUE)
lines(density(dose1, adjust = 1), col = "red", lwd = 2)
hist(dose2, density = 10, col = rgb(0.7,0.3,0.4), prob = TRUE)
lines(density(dose2, adjust = 1), col = "red", lwd = 2)
```



From the plots we see that in all three cases there are no outlyers and not very skewness. And as before we assume that the data in ToothGrowth are independent from each other. This means that we can do t tests in all three cases.

1. dose 0.5 vs dose 1 (ignoring supp)

We will show the results after all three comparisons.

```
ToothGrowth.0.5_1 <- ToothGrowth %>% filter(dose == 0.5 | dose == 1)

p.value.0.5_1 <- t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = ToothGrowth.0.5_1)$p.valu
```

2. dose 0.5 vs dose 2 (ignoring supp)

```
ToothGrowth.0.5_2 <- ToothGrowth %>% filter(dose == 0.5 | dose == 2)

p.value.0.5_2 <- t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = ToothGrowth.0.5_2)$p.valu
```

3. dose 1 vs dose 2 (ignoring supp)

```
ToothGrowth.1_2 <- ToothGrowth %>% filter(dose == 1 | dose == 2)

p.value.1_2 <- t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = ToothGrowth.1_2)$p.value
```

Bonferroni Correction

As we make 3 comparisons, we will do a Bonferroni correction to avoid a big type 1 error. Bonferroni correction is pretty simple. We are just going to divide $\alpha = 0.05$ by the number of comparisons, 3. This new level is going to be our corrected α level.

```
alpha_bonferroni_1 <- 0.05/3
alpha_bonferroni_1
```

```
## [1] 0.01666667
```

P values

```
ifelse(p.value.0.5_1 < alpha_bonferroni_1,
       paste("P value, dose 0.5 vs 1 (ignoring supp):",
             round(p.value.0.5_1, 10),"<",round(alpha_bonferroni_1,4)),
      paste("P value, dose 0.5 vs 1 (ignoring supp):",
            round(p.value.0.5_1, 10))
      )
```

```
## [1] "P value, dose 0.5 vs 1 (ignoring supp): 1.268e-07 < 0.0167"
```

```
ifelse(p.value.0.5_2 < alpha_bonferroni_1,
       paste("P value, dose 0.5 vs 2 (ignoring supp):",
             round(p.value.0.5_2, 10),"<",round(alpha_bonferroni_1,4)),
       paste("P value, dose 0.5 vs 2 (ignoring supp):",
             round(p.value.0.5_2, 10))
       )
```

```
## [1] "P value, dose 0.5 vs 2 (ignoring supp): 0 < 0.0167"
```

```
ifelse(p.value.1_2 < alpha_bonferroni_1,
       paste("P value, dose 1 vs 2 (ignoring supp):",
             round(p.value.1_2, 10),"<",round(alpha_bonferroni_1,4)),
       paste("P value, dose 1 vs 2 (ignoring supp):",
             round(p.value.1_2, 10))
       )
```

```
## [1] "P value, dose 1 vs 2 (ignoring supp): 1.90643e-05 < 0.0167"
```

It turns out that in all three cases we can reject the null hypothesis, which states that there are no differences in doses.

c. Tooth Growth by dose When Accounting for supp

We are going to make 6 comparisons:

1. dose 0.5 (OJ) vs dose 1 (OJ)
2. dose 0.5 (OJ) vs dose 2 (OJ)
3. dose 1 (OJ) vs dose 2 (OJ)
4. dose 0.5 (VC) vs dose 1 (VC)

5. dose 0.5 (VC) vs dose 2 (VC)

6. dose 1 (VC) vs dose 2 (VC)

From ToothGrowth data let's create two data frames, one for each supp.

```
oj.df <- ToothGrowth %>% filter(supp == "OJ")
vc.df <- ToothGrowth %>% filter(supp == "VC")
```

As before the conditions for using t test are met, because:

- We assume data in ToothGrowth are independent.
- We saw that there were no outlyers and not very skweness in the distributions of doses when we didn't separate the supps. And as the doses we are considering now are the subsets of the doses we saw before, hence we will assume the condition is also met here.

1. dose 0.5 (OJ) vs dose 1 (OJ)

We are going to do all 6 tests and show the results in the end.

```
oj.df.0.5_1 <- oj.df %>% filter(dose == 0.5 | dose == 1)

p.oj.0.5_1 <- t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = oj.df.0.5_1)$p.value
```

2. dose 0.5 (OJ) vs dose 2 (OJ)

```
oj.df.0.5_2 <- oj.df %>% filter(dose == 0.5 | dose == 2)

p.oj.0.5_2 <- t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = oj.df.0.5_2)$p.value
```

3. dose 1 (OJ) vs dose 2 (OJ)

```
oj.df.1_2 <- oj.df %>% filter(dose == 1 | dose == 2)

p.oj.1_2 <- t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = oj.df.1_2)$p.value
```

4. dose 0.5 (VC) vs dose 1 (VC)

```
vc.df.0.5_1 <- vc.df %>% filter(dose == 0.5 | dose == 1)

p.vc.0.5_1 <- t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = vc.df.0.5_1)$p.value
```

5. dose 0.5 (VC) vs dose 2 (VC)

```
vc.df.0.5_2 <- vc.df %>% filter(dose == 0.5 | dose == 2)

p.vc.0.5_2 <- t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = vc.df.0.5_2)$p.value
```

6. dose 1 (VC) vs dose 2 (VC)

```
vc.df.1_2 <- vc.df %>% filter(dose == 1 | dose == 2)

p.vc.1_2 <- t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = vc.df.1_2)$p.value
```

Bonferroni Correction

As we have 6 comparisons, we will make a Bonferroni correction to avoid a big type 1 error. Bonferroni correction is pretty simple. We are just going to divide $\alpha = 0.05$ by the number of comparisons, 6. This new level is going to be our corrected α level.

```
alpha_bonferroni_2 <- 0.05/6
alpha_bonferroni_2
```

```
## [1] 0.008333333
```

P Values

```
ifelse(p.oj.0.5_1 < alpha_bonferroni_2,
       paste("p value, dose 0.5 (OJ) vs 1 (OJ):",
             round(p.oj.0.5_1,6), "<", round(alpha_bonferroni_2,6)),
       paste("p value, dose 0.5 (OJ) vs 1 (OJ):",
             round(p.oj.0.5_1,6))
       )
```

```
## [1] "p value, dose 0.5 (OJ) vs 1 (OJ): 8.8e-05 < 0.008333"
```

```
ifelse(p.oj.0.5_2 < alpha_bonferroni_2,
       paste("p value, dose 0.5 (OJ) vs 2 (OJ):",
             round(p.oj.0.5_2,6), "<", round(alpha_bonferroni_2,6)),
       paste("p value, dose 0.5 (OJ) vs 2 (OJ):",
             round(p.oj.0.5_2,6))
       )
```

```
## [1] "p value, dose 0.5 (OJ) vs 2 (OJ): 1e-06 < 0.008333"
```

```
ifelse(p.oj.1_2 < alpha_bonferroni_2,
       paste("p value, dose 1 (OJ) vs 2 (OJ):",
             round(p.oj.1_2,6), "<", round(alpha_bonferroni_2,6)),
       paste("p value, dose 1 (OJ) vs 2 (OJ):",
             round(p.oj.1_2,6))
       )
```

```
## [1] "p value, dose 1 (OJ) vs 2 (OJ): 0.039195"
```

```
ifelse(p.vc.0.5_1 < alpha_bonferroni_2,
       paste("p value, dose 0.5 (VC) vs 1 (VC):",
             round(p.vc.0.5_1,6), "<", round(alpha_bonferroni_2,6)),
       paste("p value, dose 0.5 (VC) vs 1 (VC):",
             round(p.vc.0.5_1,6))
       )
```

```
## [1] "p value, dose 0.5 (VC) vs 1 (VC): 1e-06 < 0.008333"
```

```
ifelse(p.vc.0.5_2 < alpha_bonferroni_2,
       paste("p value, dose 0.5 (VC) vs 2 (VC):",
             round(p.vc.0.5_2,10), "<", round(alpha_bonferroni_2,6)),
       paste("p value, dose 0.5 (VC) vs 2 (VC):",
             round(p.vc.0.5_2,10))
       )
```

```
## [1] "p value, dose 0.5 (VC) vs 2 (VC): 4.68e-08 < 0.008333"
```

```
ifelse(p.vc.1_2 < alpha_bonferroni_2,
       paste("p value, dose 1 (VC) vs 2 (VC):",
             round(p.vc.1_2,10), "<", round(alpha_bonferroni_2,6)),
       paste("p value, dose 1 (VC) vs 2 (VC):",
             round(p.vc.1_2,10))
       )
```

```
## [1] "p value, dose 1 (VC) vs 2 (VC): 9.1556e-05 < 0.008333"
```

Well, in almost all cases we reject the null hypothesis that there are no differences in the compared groups. Only for orange juice supplement of dose 1 and dose 2 we don't have enough evidence to reject the null hypothesis.

d. Tooth Growth by supp When Accounting for dose

We are going to make three comparisons.

1. dose 0.5 (OJ) vs dose 0.5 (VC)
2. dose 1 (OJ) vs dose 1 (VC)
3. dose 2 (OJ) vs dose 2 (VC)

Let's do a quick preprocessing of data

```
oj_vc.df.0.5 <- ToothGrowth %>% filter(dose == 0.5)
oj_vc.df.1 <- ToothGrowth %>% filter(dose == 1)
oj_vc.df.2 <- ToothGrowth %>% filter(dose == 2)
```

1. dose 0.5 (OJ) vs dose 0.5 (VC)

We will show the results after doing all three comparisons.

```
p.oj_vc.0.5 <- t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = oj_vc.df.0.5)$p.value
```

2. dose 1 (OJ) vs dose 1 (VC)

```
p.oj_vc.1 <- t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = oj_vc.df.1)$p.value
```

3. dose 2 (OJ) vs dose 2 (VC)

```
p.oj_vc.2 <- t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = oj_vc.df.2)$p.value
```

Bonferroni Correction

We have 3 comparisons, so our corrected α level will be $\alpha/3$.

```
alpha_bonferroni_3 <- 0.05/3
alpha_bonferroni_3
```

```
## [1] 0.01666667
```

P Values

```
ifelse(p.oj_vc.0.5 < alpha_bonferroni_3,
       paste("p value, dose 0.5 (OJ) vs 0.5 (VC):",
             round(p.oj_vc.0.5,6), "<",round(alpha_bonferroni_3,6)),
       paste("p value, dose 0.5 (OJ) vs 0.5 (VC):",
             round(p.oj_vc.0.5,6))
       )
```

```
## [1] "p value, dose 0.5 (OJ) vs 0.5 (VC): 0.006359 < 0.016667"
```

```
ifelse(p.oj_vc.1 < alpha_bonferroni_3,
       paste("p value, dose 1 (OJ) vs 1 (VC):",
             round(p.oj_vc.1,6), "<",round(alpha_bonferroni_3,6)),
       paste("p value, dose 1 (OJ) vs 1 (VC):",
             round(p.oj_vc.1,6))
       )
```

```
## [1] "p value, dose 1 (OJ) vs 1 (VC): 0.001038 < 0.016667"
```

```
ifelse(p.oj_vc.2 < alpha_bonferroni_3,
       paste("p value, dose 2 (OJ) vs 2 (VC):",
             round(p.oj_vc.2,6), "<",round(alpha_bonferroni_3,6)),
       paste("p value, dose 2 (OJ) vs 2 (VC):",
             round(p.oj_vc.2,6))
       )
```

```
## [1] "p value, dose 2 (OJ) vs 2 (VC): 0.963852"
```

As we see, we don't have enough evidence to deny that dose 2 of orange juice is resulting in different tooth length than dose 2 of Ascorbic Acid.
In two other cases we will reject the null hypothesis.

Note: Originally the report was required to be 3 pages, plus appendix but I extended it after finishing the course