Statistical Inference Project Part 2: Basic Inferential Data Analysis

Joselle Abagat

5/27/2018

Criteria

1. Did you show where the distribution is centered at and compare it to the theoretical center of the distribution?
2. Did you show how variable it is and compare it to the theoretical variance of the distribution?
3. Did you perform an exploratory data analysis of at least a single plot or table highlighting basic features of the data?
4. Did the student perform some relevant confidence intervals and/or tests?
5. Were the results of the tests and/or intervals interpreted in the context of the problem correctly?
6. Did the student describe the assumptions needed for their conclusions?

Load libraries

The R libraries used in this report are: ggplot2, datatable, DT

Part 2: Basic Inferential Data Analysis - ToothGrowth Data

1. Load the ToothGrowth data and perform some basic exploratory data analyses

Description

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).

Format

A data frame with 60 observations on 3 variables. [,1] len numeric Tooth length [,2] supp factor Supplement type (VC or OJ). [,3] dose numeric Dose in milligrams/day

```
tg <- data.table(ToothGrowth)
# check the dimensions
dim(tg)
```

```
## [1] 60  3
```

```
# look at the first set of data
head(tg)
```

```
##     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
```

```
# look at the last set of data
tail(tg)
```

```
##     len supp dose
## 1: 24.8   OJ    2
## 2: 30.9   OJ    2
## 3: 26.4   OJ    2
## 4: 27.3   OJ    2
## 5: 29.4   OJ    2
## 6: 23.0   OJ    2
```

```
# check the column names
colnames(tg)
```

```
## [1] "len"  "supp" "dose"
```

```
# check the length of data
nrow(tg)
```

```
## [1] 60
```

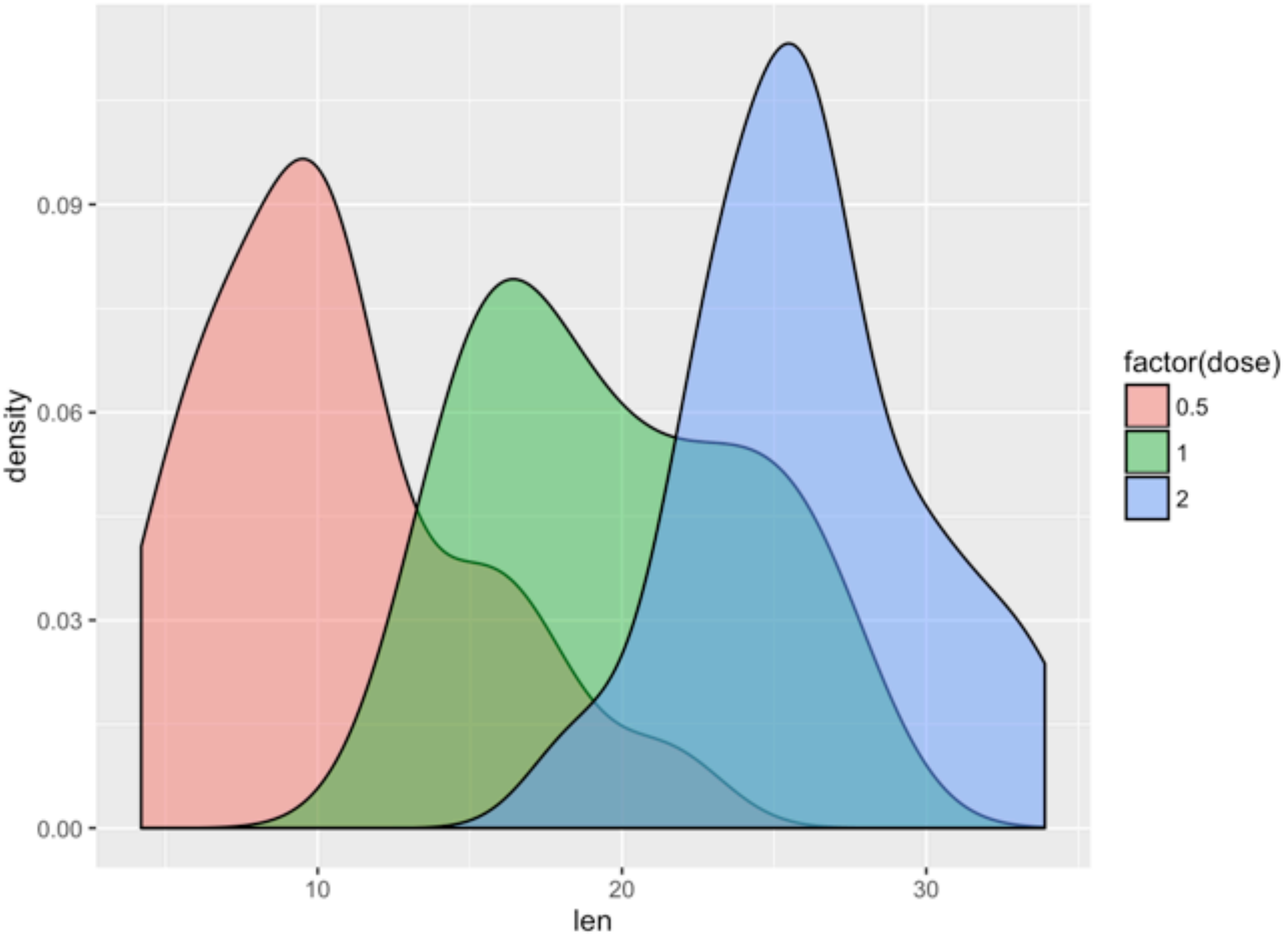
```
# factor delivery methods
levels(factor(tg$supp))
```

```
## [1] "OJ" "VC"
```

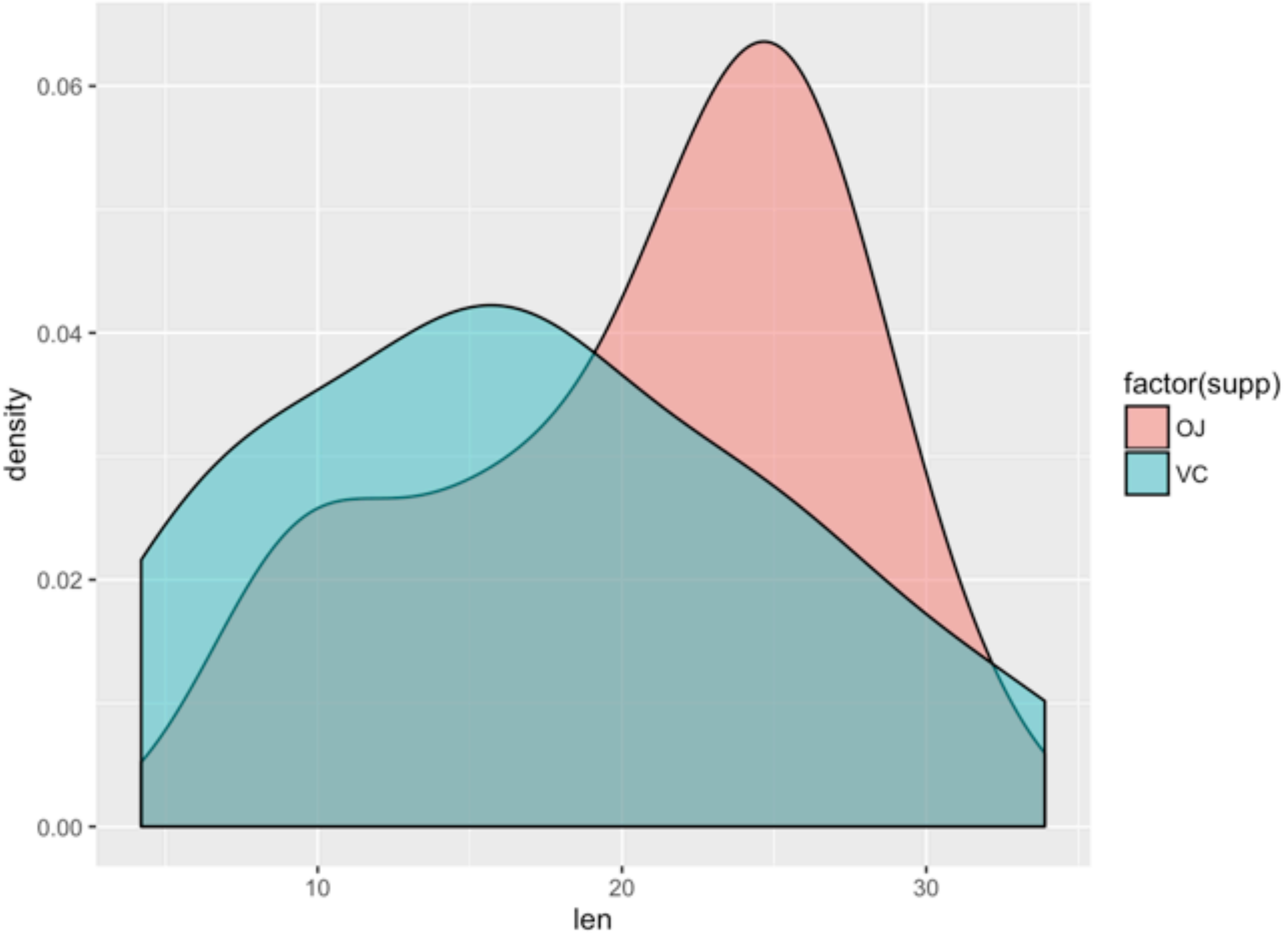
```
# factor doses
levels(factor(tg$dose))
```

```
## [1] "0.5" "1"   "2"
```

```
# look at the density distributions by doses
ggplot(tg, aes(x = len, fill = factor(dose))) + geom_density(alpha = 0.5)
```



```
# look at the density distributions by delivery methods
ggplot(tg, aes(x = len, fill = factor(supp))) + geom_density(alpha = 0.5)
```



Looking at the density plots, 2 ml/day seems to lead to the largest tooth growth. Using orange juice as a delivery method also seems to lead to the largest tooth growth.

2. Provide a basic summary of the data.

```
# overall summary
summary(tg)
```

```
##       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  
```

The basic summary doesn’t really make sense since it combines the means for all doses and all delivery methods, so let’s summarize the data grouped by delivery method and by dose.

```
# summary by delivery methods
tgSummary <- tg[, .(min(len), quantile(len, .25), median(len), mean(len), quantile(le
n, .75), max(len)), keyby = c("supp", "dose")]
colnames(tgSummary) <- c("supp", "dose", "min", "percentile25", "median", "mean", "pe
rcentile75", "max")
datatable(tgSummary)
```

Show 10 **entries** **Search:**

	supp	dose	min	percentile25	median	mean	percentile75	max
1	OJ	0.5	8.2	9.7	12.25	13.23	16.175	21.5
2	OJ	1	14.5	20.3	23.45	22.7	25.65	27.3
3	OJ	2	22.4	24.575	25.95	26.06	27.075	30.9
4	VC	0.5	4.2	5.95	7.15	7.98	10.9	11.5
5	VC	1	13.6	15.275	16.5	16.77	17.3	22.5
6	VC	2	18.5	23.375	25.95	26.14	28.8	33.9

Showing 1 to 6 of 6 entries

Previous 1 Next

3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there’s other approaches worth considering)

```
# tooth growth using OJ at 0.5 ml/day
ojHalf <- tg[supp == "OJ" & dose == 0.5]
nOJHalf <- nrow(ojHalf)
sOJHalf <- sd(ojHalf$len)
# tooth growth using OJ at 1.0 ml/day
oj1 <- tg[supp == "OJ" & dose == 1]
nOJ1 <- nrow(oj1)
sOJ1 <- sd(oj1$len)
# tooth growth using OJ at 2.0 ml/day
oj2 <- tg[supp == "OJ" & dose == 2]
nOJ2 <- nrow(oj2)
sOJ2 <- sd(oj2$len)
# tooth growth using VC at 0.5 ml/day
vcHalf <- tg[supp == "VC" & dose == 0.5]
nVCHalf <- nrow(vcHalf)
sVCHalf <- sd(vcHalf$len)
# tooth growth using VC at 1.0 ml/day
vc1 <- tg[supp == "VC" & dose == 1]
nVC1 <- nrow(vc1)
sVC1 <- sd(vc1$len)
# tooth growth using VC at 2.0 ml/day
vc2 <- tg[supp == "VC" & dose == 2]
nVC2 <- nrow(vc2)
sVC2 <- sd(vc2$len)

# Calc using 95th Confidence Intervals
# Compare Confidence Intervals at Dose = .5 using OJ
ciOJHalf <- round(mean(ojHalf$len) + c(-1,1)*qnorm(0.975)*sOJHalf/sqrt(nOJHalf),2)
# Compare Confidence Intervals at Dose = .5 using VC
ciVCHalf <- round(mean(vcHalf$len) + c(-1,1)*qnorm(0.975)*sVCHalf/sqrt(nVCHalf),2)
# Compare Confidence Intervals at Dose = 1 using OJ
ciOJ1 <- round(mean(oj1$len) + c(-1,1)*qnorm(0.975)*sOJ1/sqrt(nOJ1),2)
# Compare Confidence Intervals at Dose = 1 using VC
ciVC1 <- round(mean(vc1$len) + c(-1,1)*qnorm(0.975)*sVC1/sqrt(nVC1),2)
# Compare Confidence Intervals at Dose = 2 using OJ
ciOJ2 <- round(mean(oj2$len) + c(-1,1)*qnorm(0.975)*sOJ2/sqrt(nOJ2),2)
# Compare Confidence Intervals at Dose = 2 using VC
ciVC2 <- round(mean(vc2$len) + c(-1,1)*qnorm(0.975)*sVC2/sqrt(nVC2),2)
```

Let’s look at the summary of the data containing lower/upper limits with 95th Confidence Interval

```
cOJHalf <- c(mean(ojHalf$len), round(sOJHalf, 2), ciOJHalf)
cVCHalf <- c(mean(vcHalf$len), round(sVCHalf, 2), ciVCHalf)
cOJ1 <- c(mean(oj1$len), round(sOJ1, 2), ciOJ1)
cVC1 <- c(mean(vc1$len), round(sVC1, 2), ciVC1)
cOJ2 <- c(mean(oj2$len), round(sOJ2, 2), ciOJ2)
cVC2 <- c(mean(vc2$len), round(sVC2, 2), ciVC2)

dtOJ <- data.table(rbind(cOJHalf, cOJ1, cOJ2))
dtOJ <- cbind("OJ", c(.5, 1, 2), dtOJ)

dtVC <- data.table(rbind(cVCHalf, cVC1, cVC2))
dtVC <- cbind("VC", c(.5, 1, 2), dtVC)

dt <- rbind(dtOJ, dtVC)
colnames(dt) <- c("supp", "dose", "mean", "sd", "lower", "upper")

datatable(dt)
```

Show 10 **entries** **Search:**

	supp	dose	mean	sd	lower	upper
1	OJ	0.5	13.23	4.46	10.47	15.99
2	OJ	1	22.7	3.91	20.28	25.12
3	OJ	2	26.06	2.66	24.41	27.71
4	VC	0.5	7.98	2.75	6.28	9.68
5	VC	1	16.77	2.52	15.21	18.33
6	VC	2	26.14	4.8	23.17	29.11

Showing 1 to 6 of 6 entries Previous 1 Next

Now, let’s use the hypothesis tests to compare tooth growth by delivery method and dose. Using this will allow us to compare both delivery methods by dose. In this way, we may be able to tell more about the data.

```
# t.test at 0.5ml dose
t.test(ojHalf$len, vcHalf$len, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  ojHalf$len and vcHalf$len
## 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
```

```
# t.test at 1.0ml dose
t.test(oj1$len, vc1$len, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  oj1$len and vc1$len
## 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
```

```
# t.test at 2.0ml dose
t.test(oj2$len, vc2$len, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  oj2$len and vc2$len
## 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
```

4. State your conclusions and the assumptions needed for your conclusions.

- Based on the initial exploration of the ToothGrowth data, a dose of 2 ml/day seems to lead to the largest tooth growth. Using orange juice as a delivery method also seems to lead to the largest tooth growth. The summary validates this since at 2ml/day, both delivery methods show the highest mean: 26.06 using OJ and 26.14 using VC.

- The confidence intervals shown in each of the delivery method and dose shows that if we were to repeat the experiment, 95% of the data are expected to fall within these upper and lower limits. We can see that all the means by supp and dose are within the calculated limits.
- t.tests provided the most meaningful look at the data
- Assuming that the null hypothesis is true, we would reject the null hypothesis if the p-value is less than alpha (at alpha = 0.05).
    - at dose = 0.5, the p-value = 0.006 < 0.05, therefore, we would reject the null hypothesis. This means that the difference is statistically significant using either delivery methods at this dosage. This makes sense since the means of OJ and VC at 0.5 dose are quite different. The confidence interval range says that 95% of the time, there could be as little as 1.7 or as large as 8.8 difference in tooth growth length when using either OJ or VC.
    - at dose = 1, the p-value = 0.001 < 0.05. The same conclusion can be derived here as when the dose = 0.5. The confidence interval range says that 95% of the time, there could be as little as 2.8 or as large as 9.1 difference in tooth growth length when using either OJ or VC.
    - at dose = 2, the p-value = 0.964 > 0.05, therefore, we would not reject the null hypothesis. This means that the difference is not statistically significant using either delivery methods at this dosage. This is further proven by the fact that 0 is contained within the confidence interval. This means that at some point, there would be no difference at all in growth length if one uses either OJ or VC.

Based on the t.tests performed, while it’s true that the 2ml/day dosage leads to the largest tooth growth, it doesn’t matter as much which delivery method we use at this dosage. However, if we decide to use 0.5ml/day or 1ml/day, then tooth growth will be affected depending on whether OJ or VC is used as the delivery method.