Statistical Inference Course Project:

Analysis of Tooth Growth by Supplement and Dosage

Erik Cornelsen

March 25, 2016

Synopsis / Overview

We will analyze the `ToothGrowth` data set that is available in R. The full data set name is 'The Effect of Vitamin C on Tooth Growth in Guinea Pigs'. It contains measurements of 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). Presumably the data was not collected in a pairwise fashion, meaning that the same guinea pig was not subjected to different supplements and doses over different time periods. For this analysis we want to compare the tooth growth of the guinea pigs by the supplement and dose levels.

Explore & Describe the Data

```
library(gridExtra)
library(dplyr)
library(ggplot2)
library(knitr)

# get data set
library(datasets)
data(ToothGrowth)
data <- ToothGrowth
```

Basic Data Features

The `len` variable gives the tooth growth. The `supp` variable gives the supplement type (OJ, VC). The `dose` variable gives the supplement dose. Both `supp` and `dose` are discrete, `len` is the only continous variable.

```
str(data)
```

```
## '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 can look at the stats for the `len` variable separated by supplement type or dosage.

```
statsSupp <- summarise(group_by(data, supp), count= n(), mean=mean(len), median=median(le
n), "Std Dev" = sd(len))
statsDose <- summarise(group_by(data, dose), count= n(), mean=mean(len), median=median(le
n), "Std Dev" = sd(len))
```

Statistics by Supplement Type

supp	count	mean	median	Std Dev
OJ	30	20.663	22.7	6.606
VC	30	16.963	16.5	8.266

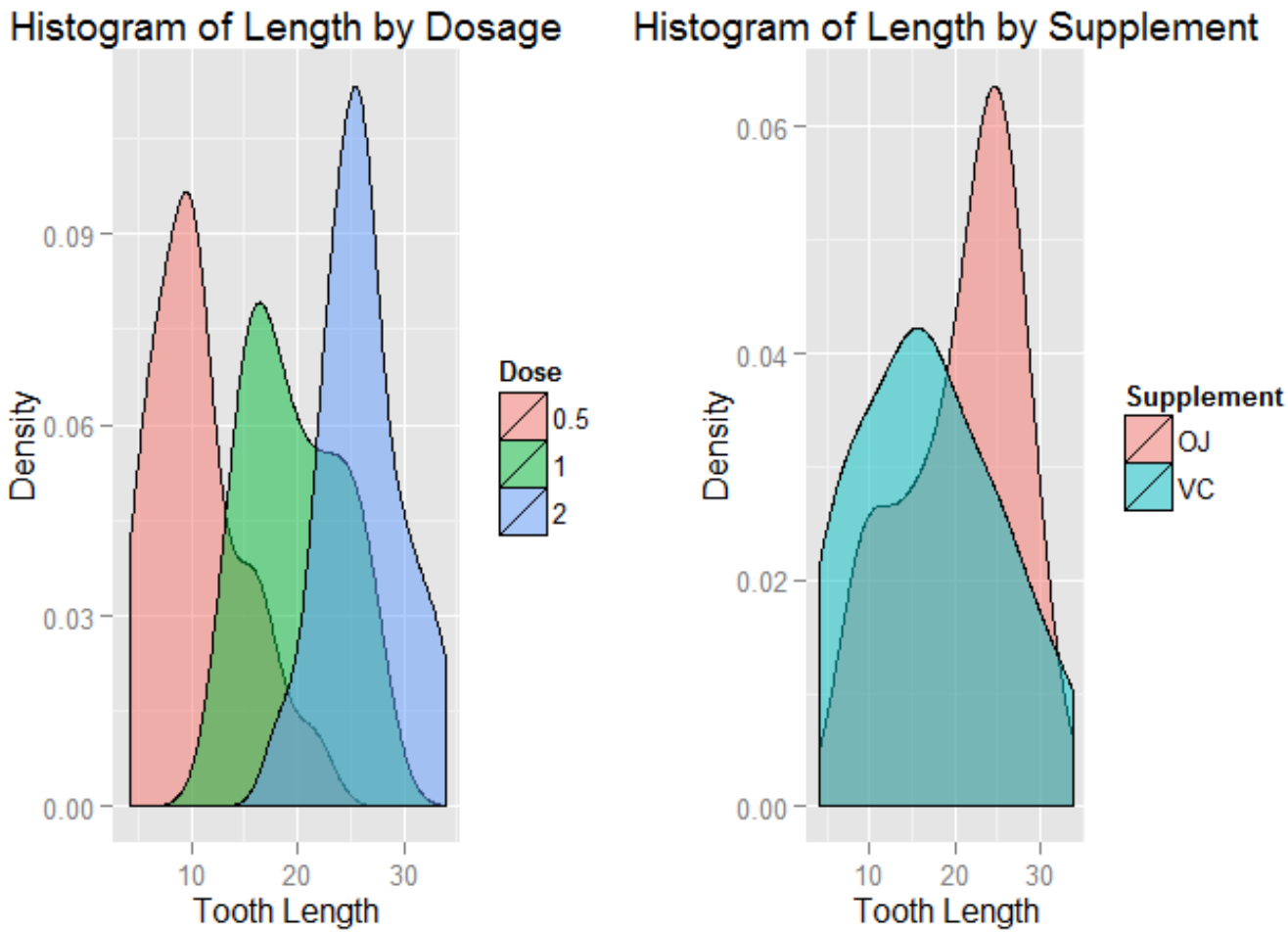
Statistics by Dosage

dose	count	mean	median	Std Dev
0.5	20	10.605	9.85	4.500
1.0	20	19.735	19.25	4.415
2.0	20	26.100	25.95	3.774

Visual Exploration

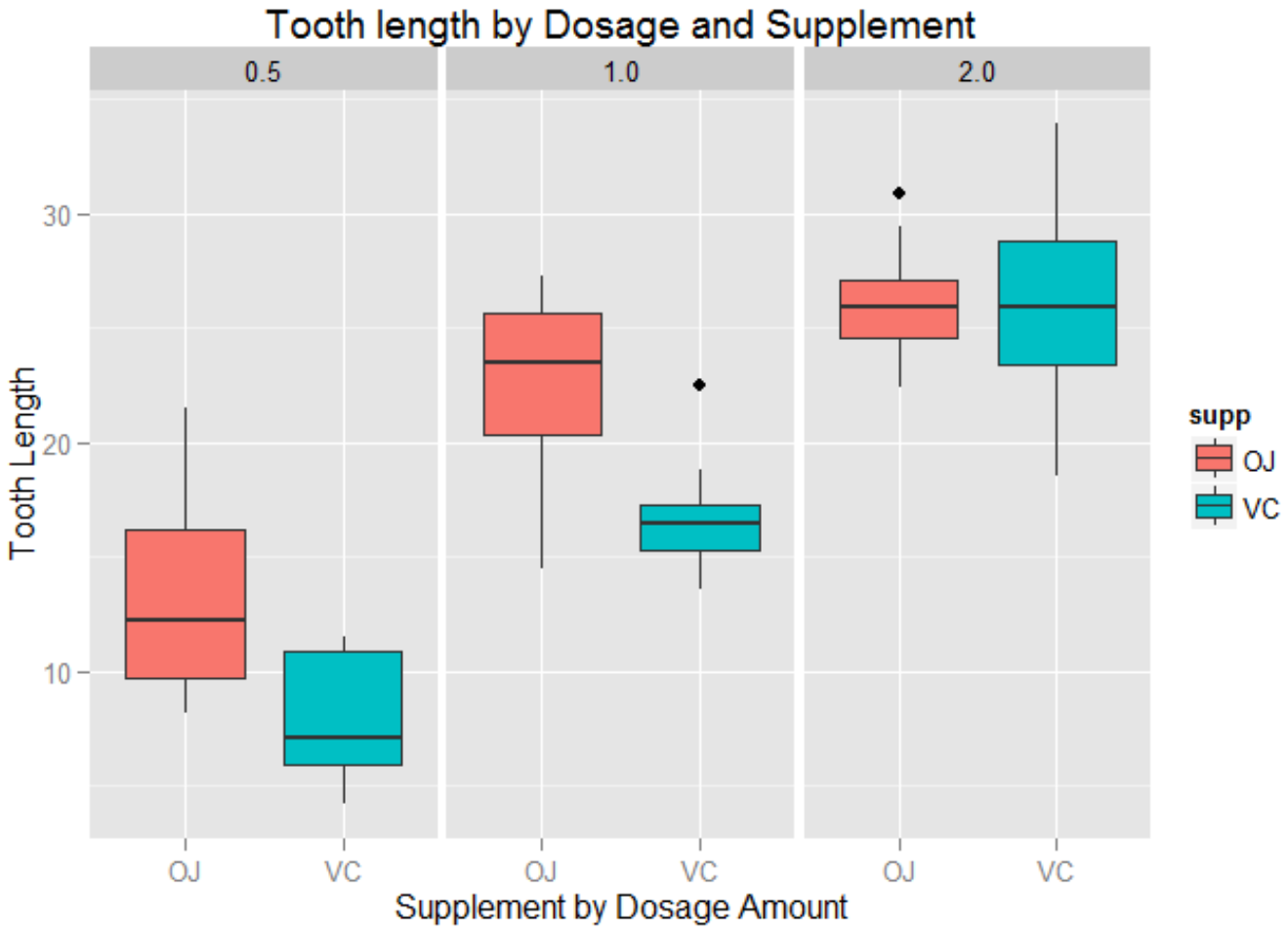
We can look at histograms comparing Supplement and Dosage.

```
g1 <- ggplot(data, aes(x=len, fill=as.factor(dose))) + geom_density(alpha = 0.5) +
    labs(x='Tooth Length', y='Density', title='Histogram of Length by Dosage') +
    scale_fill_discrete(name="Dose")

g2 <- ggplot(data, aes(x=len, fill=as.factor(supp))) + geom_density(alpha = 0.5) +
    labs(x='Tooth Length', y='Density', title='Histogram of Length by Supplement') +
    scale_fill_discrete(name="Supplement")

grid.arrange(g1, g2, ncol=2)
```



We will do a visual check on the potential relationships between delivery methods at each dose level in a boxplot.

```
ggplot(data = data, aes(x = supp, y = len)) +
    geom_boxplot(aes(fill = supp)) +
    facet_wrap(~ dose) +
    labs(x='Supplement by Dosage Amount', y='Tooth Length', title='Tooth length by Dosage
and Supplement')
```



The relationship between dosage amount and tooth length strongly suggests that the higher dosages may cause longer tooth length as the tooth length consistently increases as the dosage increases.

The relationship between the supplement type and tooth length is not obvious at this stage. At lower dosages it seems that orange juice correlates with longer teeth, but at the higest dosage (2mg) there is no significant difference.

Hypothesis Testing & Confidence Intervals

We will use confidence intervals and hypothesis tests to compare tooth growth by Supplement Type and Dosage. We will look at Supplement and Dosage separatley.

Assumptions

1. The poplulations are independent and a random population was used.
2. The data was not collected in a pairwise fashion, meaning that the same guinea pig was not subjected to different supplements and doses over different time periods.
3. Because the sample sizes are small it is appropriate to use the t-test to calculate confidence intervals.
4. The variances between the sample popluations are not equal (var.equal=FALSE)

5. The sample data is not paired (paired=FALSE)

Testing Supplement Type

- H0: The type of supplement DOES NOT increase tooth growth. In this case, the difference in means between OJ and VC should be close to 0.
- Ha: The type of supplement DOES increase tooth growth. In this case, the difference in means between OJ and VC should NOT be close to 0

Because we want to isolate any potential effects of the dosage ammount we'll split the data up into the three dosage groups (0.5,1.0,2.0). Then within those groups, we compare the difference in means of the supplements (OJ,VC).

```
D05 <- data[data$dose == 0.5,]
D10 <- data[data$dose == 1.0,]
D20 <- data[data$dose == 2.0,]

TXX <- t.test(len~supp, data=data, paired=FALSE, var.equal=FALSE)
T05 <- t.test(len~supp, data=D05, paired=FALSE, var.equal=FALSE)
T10 <- t.test(len~supp, data=D10, paired=FALSE, var.equal=FALSE)
T20 <- t.test(len~supp, data=D20, paired=FALSE, var.equal=FALSE)

results_supp <- data.frame(
    "p-value"  = c(TXX$p.value, T05$p.value, T10$p.value, T20$p.value),
    "CI.Lower" = c(TXX$conf.int[1], T05$conf.int[1], T10$conf.int[1], T20$conf.int[1]),
    "CI.Upper" = c(TXX$conf.int[2], T05$conf.int[2], T10$conf.int[2], T20$conf.int[2]),
    "OJ mean" = c(TXX$estimate[1], T05$estimate[1], T10$estimate[1], T20$estimate[1]),
    "VC mean" = c(TXX$estimate[2], T05$estimate[2], T10$estimate[2], T20$estimate[2]),
    row.names = c("Dose: ALL","Dose: 0.5","Dose: 1.0","Dose: 2.0" )
    )

kable(results_supp, digits=3, align='c', caption="95% t-test summary for Supplement Types
(OJ,VC)")
```

95% t-test summary for Supplement Types (OJ,VC)

	p.value	CI.Lower	CI.Upper	OJ.mean	VC.mean
Dose: ALL	0.061	-0.171	7.571	20.663	16.963
Dose: 0.5	0.006	1.719	8.781	13.230	7.980
Dose: 1.0	0.001	2.802	9.058	22.700	16.770
Dose: 2.0	0.964	-3.798	3.638	26.060	26.140

Supplement Conclusions

If we look at all the data together (Dose:ALL) we would accept H0 since the 95% confidence interval contains the case where the difference in means equals zero. But when we look at OJ and VC only within the Dosage groups we find that we would need to reject H0 for Dose=0.5 and Dose=1.0 and accept H0 for Dose=2.0.

There may be some additional factors (like absorption rates) that could be investigated to account for the differences between dosage ammounts. For now we will rely on the Dose=ALL data and therefore *accept the null hypothosis* that the type of supplement *DOES NOT* increase tooth growth.

Testing Dosage Ammount

- H0: An increase in dosage DOES NOT increase tooth growth. In this case, the difference between means of each dosage ammount should be closer to 0.
- Ha: An increase in dosage DOES increase tooth growth. so diff in means of each dosage amt should NOT be close to zero.

1. We will initially perform the test on the entire data withot splitting it up by supplement so that we have an additional referece to look at.
2. Because we want to isolage any potential effects of the supplement types, we'll also split the data up into the two supplement groups (OJ,VC) and then compare dosage amounts within those groups.
3. Because the t.test requires exactly 2 means to compare we must further split up the dosage comparisons so that each comparison has exactly 2 levels. [0.5>>1.0],[1.0>>2.0],[0.5>>2.0]

```
library(dplyr)

#Subset data for dosage amounts
D01 <- filter(data,dose==c(0.5,1))
D12 <- filter(data,dose==c(1,2))
D02 <- filter(data,dose==c(0.5,2))

#subset data for OJ and dosage amounts
OJ01 <- filter(data,supp=="OJ" & dose==c(0.5,1))
OJ12 <- filter(data,supp=="OJ" & dose==c(1,2))
OJ02 <- filter(data,supp=="OJ" & dose==c(0.5,2))

#subset data for VC and dosage amounts
VC01 <- filter(data,supp=="VC" & dose==c(0.5,1))
VC12 <- filter(data,supp=="VC" & dose==c(1,2))
VC02 <- filter(data,supp=="VC" & dose==c(0.5,2))

#t.test for dosage amounts
TD01 <- t.test(len~dose, data=D01, paired=FALSE, var.equal=FALSE)
TD12 <- t.test(len~dose, data=D12, paired=FALSE, var.equal=FALSE)
TD02 <- t.test(len~dose, data=D02, paired=FALSE, var.equal=FALSE)

#t.test for OJ and dosage amounts
TOJ01 <- t.test(len~dose, data=OJ01, paired=FALSE, var.equal=FALSE)
TOJ12 <- t.test(len~dose, data=OJ12, paired=FALSE, var.equal=FALSE)
TOJ02 <- t.test(len~dose, data=OJ02, paired=FALSE, var.equal=FALSE)

#t.test for VC and dosage amounts
TVC01 <- t.test(len~dose, data=VC01, paired=FALSE, var.equal=FALSE)
TVC12 <- t.test(len~dose, data=VC12, paired=FALSE, var.equal=FALSE)
TVC02 <- t.test(len~dose, data=VC02, paired=FALSE, var.equal=FALSE)

df <- data.frame()
LST <- list('All:0.5-1.0'=TD01,'All:1.0-2.0'=TD12,'All:0.5-2.0'=TD02,
            'OJ:0.5-1.0'=TOJ01,'OJ:1.0-2.0'=TOJ12,'OJ:0.5-2.0'=TOJ02,
            'VC:0.5-1.0'=TVC01,'VC:1.0-2.0'=TVC12,'VC:0.5-2.0'=TVC02)
i <- 0
for (x in LST) {
    i <- i+1
    df <- rbind(df, data.frame(p.value = x$p.value,
                               CI.lower = x$conf.int[1],
                               CI.upper = x$conf.int[2],
                               mean1 = x$estimate[1],
                               mean2 = x$estimate[2],
                               row.names = names(LST)[i]
))}

kable(df, digits=5, align='c', caption="95% t-test summary for Dosage Amts (0.5mg,1.0mg,
2.0mg)")
```

95% t-test summary for Dosage Amts (0.5mg,1.0mg,2.0mg)

	p.value	CI.lower	CI.upper	mean1	mean2
All:0.5-1.0	0.00030	-14.43327	-5.20673	10.63	20.45
All:1.0-2.0	0.00171	-10.89999	-2.98001	19.02	25.96
All:0.5-2.0	0.00000	-19.72833	-10.93167	10.63	25.96
OJ:0.5-1.0	0.00125	-14.94689	-5.61311	14.40	24.68
OJ:1.0-2.0	0.04510	-11.23644	-0.16356	20.72	26.42
OJ:0.5-2.0	0.00054	-16.94335	-7.09665	14.40	26.42
VC:0.5-1.0	0.00056	-12.65876	-6.06124	6.86	16.22
VC:1.0-2.0	0.03057	-15.32237	-1.03763	17.32	25.50
VC:0.5-2.0	0.00071	-25.62861	-11.65139	6.86	25.50

Dosage Conclusions

Looking at each of the t-tests performed we find that there is NO test where the confidence interval contains the null hypothosis case. This confirms what the graph strongly suggested that we should reject H0. This leads us to *accept the alternate hypothosis* that an increased dosage amount *DOES* increase tooth growth.

Appendix / Reference

- Description of ToothGrowth data set: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ToothGrowth.html (https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ToothGrowth.html)
- http://www.cookbook-r.com/Graphs/Legends_(ggplot2)/ (http://www.cookbook-r.com/Graphs/Legends_(ggplot2)/)
- https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf (https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf)
- http://www.rstudio.com/wp-content/uploads/2015/04/ggplot2-cheatsheet.pdf (http://www.rstudio.com/wp-content/uploads/2015/04/ggplot2-cheatsheet.pdf)
- https://drive.google.com/drive/u/0/folders/0ByIrJAE4KMTtcVBmdm1BOEZoeEk (https://drive.google.com/drive/u/0/folders/0ByIrJAE4KMTtcVBmdm1BOEZoeEk)

Class Assignment Verbaige

Analyze the ToothGrowth data in the R datasets package.

1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.

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

Some criteria that you will be evaluated on - Did you perform an exploratory data analysis of at least a single plot or table highlighting basic features of the data? - Did the student perform some relevant confidence intervals and/or tests? - Were the results of the tests and/or intervals interpreted in the context of the problem correctly? - Did the student describe the assumptions needed for their conclusions?

Basic concept of Hypothosis Testing

- Check Conditions
- Write H0 Ha
    - H0-boring hypothosis, nothing has changed,
    - Ha-Thing trying to prove, more people like X,
- Create Null Model
- Think, then calculate (pvalue, confinterval, etc)

Pvalue is likelyhood of what you observe if H0 is true. if pvalue is below .05 (coin filp) then H0 not likley, look at Ha.