Statistical Inference Course Project Part 2 – Inferential Data Analysis

Michael J. Catalano

Report Summary

The following is a statistical analsysis of the ToothGrowth data from the R datasets package. The general structure of the data was assessed with an appropriate plot, then confidence interval testing was performed to assess the relationship between tooth growth and supplement/dose.

First, we will retrieve the ToothGrowth dataset and take a look at its contents.

```
library(datasets)
toothdata <- data.frame(ToothGrowth)
str(toothdata)
```

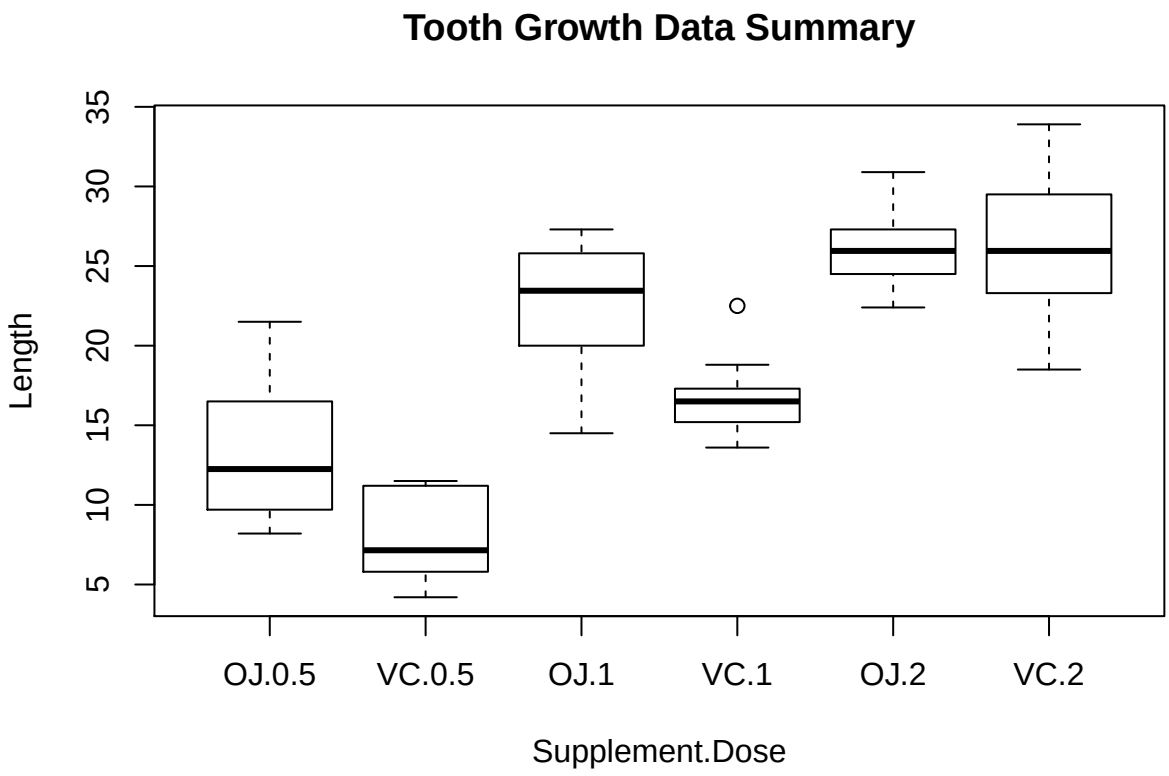
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
head(toothdata)
```

```
##    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
```

The dataset consists of 3 variables, length, supplement, and dose. We will be examining the effect of supplement and dose on tooth length. A box plot provides an illustration of the general relationship between variables.

```
boxplot(len ~ supp * dose, data=toothdata, main='Tooth Growth Data Summary', ylab='Length', xlab='Supple
```



There appears to be a direct relationship between dose and tooth length. Additionally, the supplement type OJ apparently resulted in greater tooth length than type VC for doses 0.5 and 1.0. However, we need to test whether these differences are statistically significant. Two-sample t tests will be used to evaluate whether the means from each group on the basis of supplement and dose are different or from the same population.

Our null hypothesis in each case will be that the difference in sample means is equal to 0, or in other words, that the sample means are the same.

Below we test the data based on supplement type.

```
t.test(len~supp, paired=F, var.equal=F, data=toothdata)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

The 95% confidence interval contains 0; therefore, we cannot say that the means of the supplement groups are in fact different with 95% confidence.

Now the data will be tested on the basis of dose. First, the data is subsetted into 3 tables comparing only two different doses per table. Then, the t-test data is generated.

```
s1 <- subset(toothdata, dose == c(0.5, 1.0))
s2 <- subset(toothdata, dose == c(0.5, 2.0))
s3 <- subset(toothdata, dose == c(1.0, 2.0))
```

```
t.test(len~dose, paired=F, var.equal=F, data=s1)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.4725, df = 17.976, p-value = 0.0002952
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -14.43327  -5.20673
## sample estimates:
## mean in group 0.5   mean in group 1
##             10.63             20.45
```

```
t.test(len~dose, paired=F, var.equal=F, data=s2)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -7.3335, df = 17.635, p-value = 9.362e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.72833 -10.93167
## sample estimates:
## mean in group 0.5   mean in group 2
##             10.63             25.96
```

```
t.test(len~dose, paired=F, var.equal=F, data=s3)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -3.6827, df = 17.949, p-value = 0.00171
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.899993  -2.980007
## sample estimates:
## mean in group 1 mean in group 2
##           19.02           25.96
```

Above we can see that the alternative hypothesis can be accepted for each test because the 95% confidence intervals do not contain 0. The largest p-value is obtained in the comparison of doses 1.0 and 2.0, meaning that there is a greater probability that the means for these particular groups are the same.

Conclusions For the above analysis it was assumed that the data are normally distributed and variances for each group are unequal. Accordingly, the results indicate that tooth length is likely dependent on dose, but not on supplement type. A two sample t-test suggested that the means by supplement are too close to be considered different with 95% confidence, as the confidence interval contained the value 0. The means by dose were tested in the same fashion, and it was determined that all three confidence intervals were exclusive of 0. Therefore, tooth length is affected by dose.