Statistical Inference Course Project

Cesar Espitia, June 18, 2015

Purpose

This is the project for the statistical inference class. In it, you will use simulation to explore inference and do some simple inferential data analysis. The project consists of basic inferential data analysis.

Problem 2

Now in the second portion of the class, we're going to analyze the ToothGrowth data in the R datasets package.

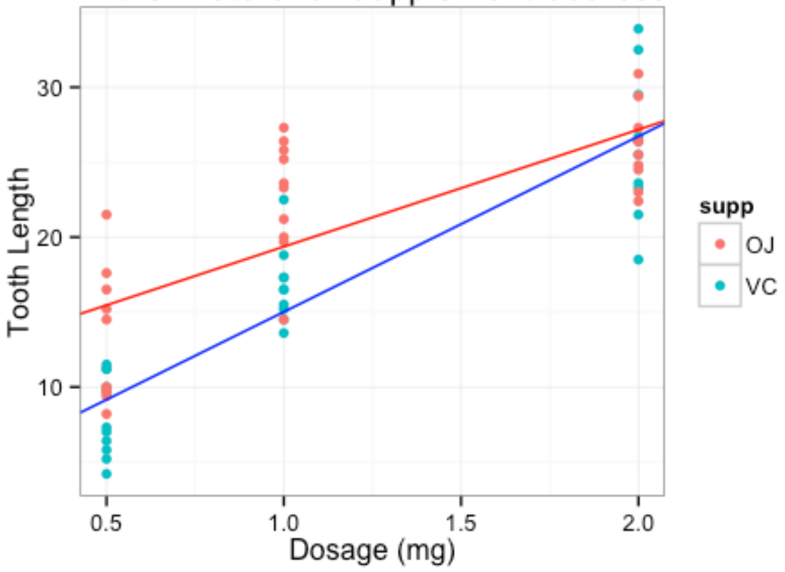
1. Load the ToothGrowth data and perform some basic exploratory data analyses

As can be seen from Figure 1, both cases of Tooth Length experienced positives growth as the dosage of the Supplement Source increased from 0.5 to 2.0 mg.

The code for creating Figure 1 is in Appendix A.

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

Figure 1. Tooth Growth vs. Calcium Dosage from Either Natural or Supplement Sources



Fitting a linear model to each supplement also shows that the Vitamin C group had more pronounced growth from the lower to higher dosage.

```
fits <- lmList(formula = len ~ dose | supp, data = Tooth)
fits

## Call: lmList(formula = len ~ dose | supp, data = Tooth)
## Coefficients:
```

```
##    (Intercept)      dose
## OJ      11.550  7.811429
## VC       3.295 11.715714
##
## Degrees of freedom: 60 total; 56 residual
## Residual standard error: 4.083142
```

2. Provide a basic summary of the data.

The data shows that the Tooth Length shows differences across the supplement source and the dosage.

```
#summary of the Tooth Growth data
Supp <- Tooth %>% group_by(supp) %>% summarise(mean = mean(len), sd = sd(len),
var = var(len))
SuppDose <- Tooth %>% group_by(supp, dose) %>% summarise(mean = mean(len), sd =
sd(len), var = var(len))
```

When comparing the Tooth Length soleley between Supplement Sources, we notice that the Vitamin C overall has a lower mean than Orange Juice and it has a higher standard deviation.

```
Supp

## Source: local data frame [2 x 4]
##
##   supp     mean       sd      var
## 1   OJ 20.66333 6.605561 43.63344
## 2   VC 16.96333 8.266029 68.32723
```

When we compare, Tooth Length by Supplement Source and Dosage, we notice that the means at 2.0 mg are nearly identical with difference in variance and standard deviation. If we look at the 0.5 mg group, the mean of Tooth Length was nearly 50% less for the Vitamin C (VC) group vs. the Orange Juice group (OJ).

```
SuppDose

## Source: local data frame [6 x 5]
## Groups: supp
##
##   supp dose  mean       sd       var
## 1   OJ  0.5 13.23 4.459709 19.889000
## 2   OJ  1.0 22.70 3.910953 15.295556
## 3   OJ  2.0 26.06 2.655058  7.049333
## 4   VC  0.5  7.98 2.746634  7.544000
## 5   VC  1.0 16.77 2.515309  6.326778
## 6   VC  2.0 26.14 4.797731 23.018222
```

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).

This analysis will focus on conducting hypothesis test due to the nature of the data which can be sliced to be solely based upon two groups (OJ vs. VC, 0.5 mg vs. 1.0mg, etc). The tests to be conducted will be Two Sample T-tests that are centered around the following hypothesis:

Comparing Supplement Groups ignoring Dosage

Ho: µ(VC) = µ(OJ)

Assumptions: groups aren't paired, variance is not equal

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

##
##  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 confidence interval includes 0 which means that we cannot reject the null hypothesis of both means being equal.

Comparing Dosage Groups ignoring Supplement

Ho: μ(VC) = μ(OJ)

Assumptions: groups aren't paired, variance is not equal Also note that the code for this section is in the Appendix B.

```
d

##      [,1]                 [,2]                 [,3]
## [1,] "Tooth_0510"         "Tooth_0520"         "Tooth_1020"
## [2,] "-11.9837812579016" "-18.1561665388306" "-8.99648051689202"
## [3,] "-6.27621874209841" "-12.8338334611694" "-3.73351948310799"
```

For all three cases in pairing dosage, 0 is not in the confidence interval. This means that we must reject the null hypothesis that there is no difference in Tooth Length amongst the dosage pairs. In face, the difference is quite large when comparing the interval between 0.5 and 2.0 mg which we can also see from the Figure 1 earlier in this document.

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

At first glance it would seem that there is no signifanct difference in Tooth Length when ignoring dosage levels, but when you do incorporate dosage levels, the varibaility and the difference begin to be more apparent.

In our original t.test comparing OJ vs. VC ignoring dosage we couldn't reject the null hypothesis at 95% confidence interval. If we lowered it by 1% then we would reject the null hypothesis, this means that if the means between both groups was different even by a little it would have also been reject.

Also, it is of interest to note that all dosage comparisons do show significant differences in Tooth Length between them and we had to reject all the null hypotheses stating that there wouldn't be.

Assumptions A few assumptions that had to be made.
- data is iid normal - data is not skewed - variances are different - the sample guinea pigs represent the genus guinea pigs as a whole, no bias in dosage, and no guinea pig was used in a different random trial (duplication).

APPENDIX A. Figure 1 Code.

```
dim(ToothGrowth)

## [1] 60  3

#store data into new var
Tooth <- tbl_df(ToothGrowth)

#create linear model comparing length vs. dosage by group supp (supplements)
fits <- lmList(formula = len ~ dose | supp, data = Tooth)

#plot data to view patterns
#qplot(dose, len, colour=supp, data=Tooth, main="Figure 1. Tooth Growth vs.
Calcium Dosage from \n Either Natural or Supplement Sources", ylab="Tooth
Length", xlab = "Dosage (mg)") + geom_abline(intercept=coef(fits)[1,1],
slope=coef(fits)[1,2], color="red") + geom_abline(intercept=coef(fits)[2,1],
slope=coef(fits)[2,2], color="blue") + theme_bw()
```

APPENDIX B. CODE FOR QUESTION 3.

```
#subset into three test (0.5 vs 1.0, 0.5 vs. 1.0, 1.0 vs. 2.0)
Tooth_0510 <- Tooth %>% filter(dose %in% c(0.5, 1.0))
Tooth_0520 <- Tooth %>% filter(dose %in% c(0.5, 2.0))
Tooth_1020 <- Tooth %>% filter(dose %in% c(2.0, 1.0))
a <- t.test(len ~ dose, paired=F, var.equal=F, data = Tooth_0510)
b <- t.test(len ~ dose, paired=F, var.equal=F, data = Tooth_0520)
c <- t.test(len ~ dose, paired=F, var.equal=F, data = Tooth_1020)

d <- cbind(c("Tooth_0510",a$conf.int))
d <- cbind(d, c("Tooth_0520",b$conf.int))
d <- cbind(d, c("Tooth_1020",c$conf.int))

d

##      [,1]                [,2]                [,3]
## [1,] "Tooth_0510"        "Tooth_0520"        "Tooth_1020"
## [2,] "-11.9837812579016" "-18.1561665388306" "-8.99648051689202"
## [3,] "-6.27621874209841" "-12.8338334611694" "-3.73351948310799"
```