Statisical Inference Course Project - Part 2

Alex McBride

Sunday, February 08, 2015

Contents

Summary of Project 1

Load the data and libraries 1

Explore the data 2

Test the data 2

Appendix **3**

Summary of Project

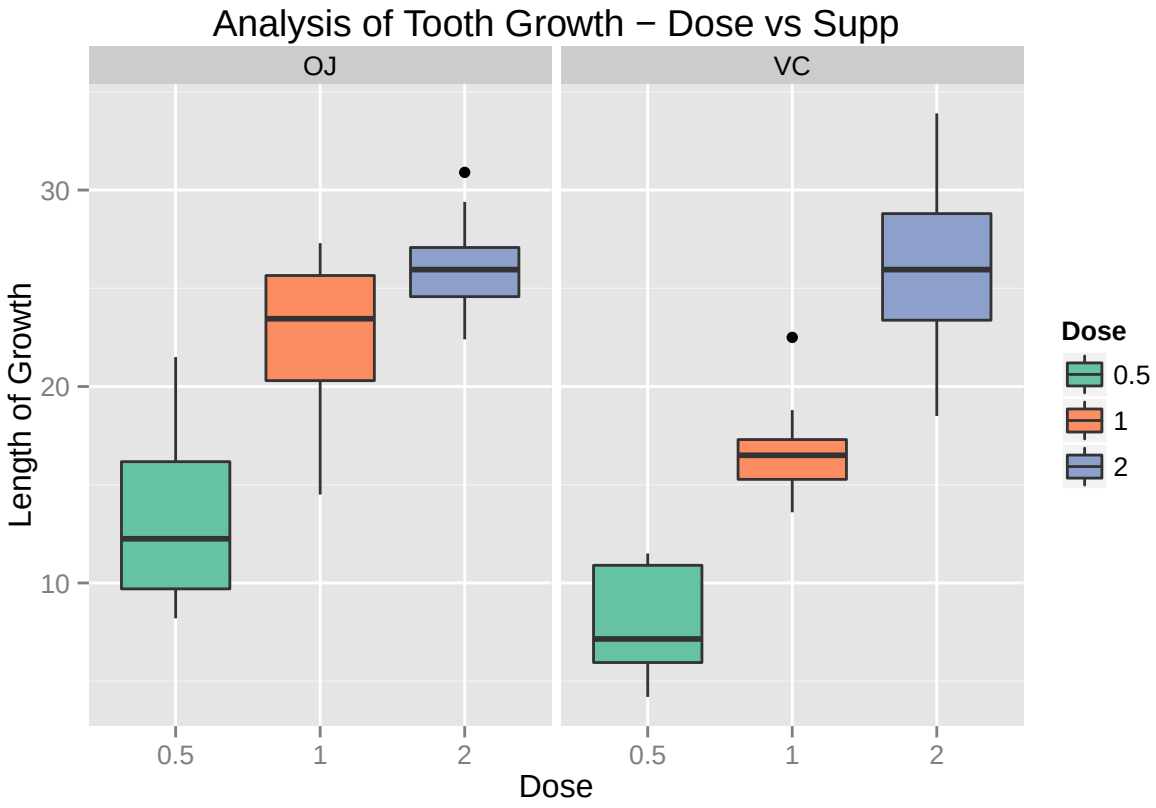
Analyze the ToothGrowth data in the R datasets package. Load the ToothGrowth data and perform some basic exploratory data analyses. Provide a basic summary of the data. Use confidence intervals and/or hypothesis tests, as taught in the Statistical Inference class, to compare tooth growth by "supp" and "dose". State conclusions, the reasoning behind the conclusions and the assumptions needed for the conclusions. For the purpose of this report, all code has been placed in the appendix at the end of the report

Load the data and libraries

```
##       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
```

From this summary we can see the overall maximum, minimum and average growth from all the tests but not split amongst the "dose" or "supp" type. Lets do that now and plot the results.

Explore the data



Looking at this simple exploratory analysis one might assume that the supp "OJ" has the better growth rates overall, regardless of dose. But if we look at dose size "2.0" we see that "VC" has a larger mean and maximum value than "OJ"

Table 1: Mean data

dose	supp	len
2	OJ	26.06
2	VC	26.14

Table 2: Max data

dose	supp	len
2	OJ	30.9
2	VC	33.9

Test the data

So, what does this mean for our analysis? We need to test comparisons of growth between "supp" and "dose" to find out whether "supp" type or "dose" rate has any effect on Tooth growth. Lets run some t-confidence tests and collate their p-values and confidence levels. We will test the two types of supplements overall and

then each as related to dose rates.

Test Assumptions

For these tests we will assume that the variance between each group is unequal, so use `var.equal = FALSE` in our `t.test`. Our hypothesis null (Ho) is that if the mean is 0 then the data groups being tested have no more effect on Tooth Growth than the other.

Run the tests

Table 3: Collated Test Results

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

Conclusions from the tests

Based of the results, we can say that for doses of 0.5 and 1.0, OJ has a greater effect on Tooth Growth than VC, we know this by the p.value indicators being less than 5% and the confidence intervals of the test do not contain `0`. For the test at `dose == 2.0` we cannot reject the Ho, as the p.value is greater than 5% and the confidence test contains `0`. For the test OJ vs VC we cannot reject the Ho either, as the p.value is 6% (greater than the 5% threshold used for statiscal analysis) and the confidence interval contains `0`
We can then conclude that to get greater tooth growth with low levels of dosage (0.5 & 1.0) one should use OJ instead of VC. At greater levels (2.0) of dosage it is uncertain whether there will be a greater effect from either OJ or VC.

Appendix

Code chunks for the report

- Load the libraries and data

```
library(ggplot2)
library(knitr)
data("ToothGrowth")
data <- as.data.frame(ToothGrowth)
summary(data)
```

- Create the plot

```
# Create Plot
ggplot(data, aes(x=factor(dose),y=len,fill=factor(dose))) +
        geom_boxplot(notch=F) + facet_grid(.~supp) +
```

```
     scale_x_discrete("Dose") +
     scale_y_continuous("Length of Growth") +
        scale_fill_brewer(name="Dose", palette = "Set2") +
     ggtitle("Analysis of Tooth Growth - Dose vs Supp")
```

- Show the Dose 2 tables

```
# get the max length per dose group
dmax <- aggregate(len~dose+supp, data=data, max)
# Get the mean length per dose group
dmean <- aggregate(len~dose+supp, data=data, mean)
# Mean data frame
dsub <- subset(dmean, dmean$dose >= 2)
rownames(dsub) <- NULL
kable(dsub, caption = "Mean data")
# Max data frame
dsub2 <- subset(dmax, dmax$dose >= 2)
rownames(dsub2) <- NULL
kable(dsub2, caption = "Max data")
```

- The Test code

```
# create dose rate subsets
dat2 <- subset(data, dose == 2)
dat1 <- subset(data, dose == 1)
dat.5 <- subset(data, dose == .5)
# run the tests and collate
tsupp <- t.test(len~I(relevel(supp, 2)), paired = FALSE, var.equal = FALSE, data = data)
t.5 <- t.test(len~I(relevel(supp, 2)), paired = FALSE, var.equal = FALSE, data = dat.5)
t1 <- t.test(len~I(relevel(supp, 2)), paired = FALSE, var.equal = FALSE, data = dat1)
t2 <- t.test(len~I(relevel(supp, 2)), paired = FALSE, var.equal = FALSE, data = dat2)
tcollate <- data.frame("p-value"=c(tsupp$p.value,t.5$p.value,t1$p.value,t2$p.value),
     "CI-Lower"=c(tsupp$conf[1],t.5$conf[1],t1$conf[1],t2$conf[1]),
     "CI-Upper"=c(tsupp$conf[2],t.5$conf[2],t1$conf[2],t2$conf[2]),
     "Mean VC" =c(tsupp$estimate[1],t.5$estimate[1],t1$estimate[1],t2$estimate[1]),
     "Mean OJ" =c(tsupp$estimate[2],t.5$estimate[2],t1$estimate[2],t2$estimate[2]),
     row.names=c("OJ vs VC: ", "0.5 dose: ", "1.0 dose: ","2.0 dose: "))
kable(round(tcollate, 3), caption = "Collated Test Results")
```