Project 2 Statistical Inference

Pablo Arias

Wednesday, December 17, 2014

Data Load and basic exploratory analyses

We will begin by loading the data and making sure it is a tidy dataset suitable for analysis. The dataset measured the response in the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

```
library(datasets)
data(ToothGrowth)
str(ToothGrowth)
```

```
## '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 see that this dataset has 60 observations arranged in rows with three variables: *len* for the tooth length as a numeric value, *supp* for the two supplements (OJ for Orange Juice and AC for Ascorbic Acid) used and *dose* for the dosage used as a number. Since the dosage was done in fixed amounts, we can convert to a factor for easier analysis.

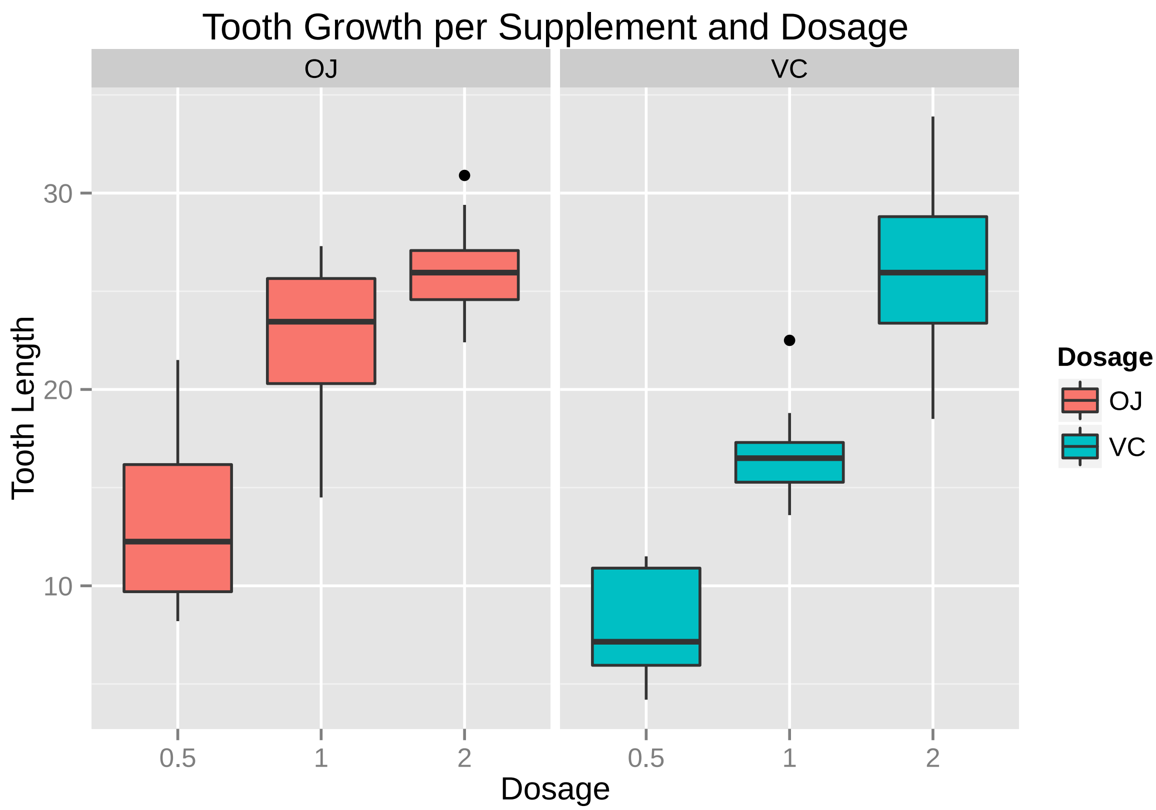
```
ToothGrowth$dose <- factor(ToothGrowth$dose)
summary(ToothGrowth)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

We can see there are 30 observations for each supplement and 10 observations for a particular dosage within each supplement.

We are interested in looking at each supplement separate and the impact of dosage has on tooth growth. We can use a boxplot to visualize

```
library(ggplot2)
ggplot(ToothGrowth, aes(x=dose, y=len)) +
        geom_boxplot(aes(fill=supp)) +
        facet_wrap( ~ supp) +
        xlab("Dosage") + ylab("Tooth Length") + ggtitle("Tooth Growth per Supplement and Dosage") +
        guides(fill=guide_legend(title="Dosage"))
```



Provide a basic summary of the data.

The boxplot shows that both supplements and dosage amounts have a positive (larger growth) impact in tooth growth. The OJ supplement appears to have a higher impact across all dosages. It is not clear enough if the data can sustain the claim that one supplement is better than the other at a specific dosage amount.

Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

From this data we can infer that what the researchers wanted to investigate if there was any impact of each supplement as a function of dosage and which supplement (if any) is the better than the other. We can do the the following:

1. Unequal and equal variance T test comparing supplements OJ and VC

We can use the *t.test()* to test the Null Hypothesis that the means for each supplement are equal (Ho: mu1 = mu2) and different under the alternative (Ha: mu1 != mu2). We use *paired = FALSE* because the set of guinea pigs that were given OJ was a complete different set than the ones that were given VC. We can check for equal and unequal variances

```
t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data=ToothGrowth)$conf.int
```

```
## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(len ~ supp, paired = FALSE, var.equal = TRUE, data=ToothGrowth)$conf.int
```

```
## [1] -0.1670064  7.5670064
## attr(,"conf.level")
## [1] 0.95
```

Whether the variances are equal or not, 0 is within the confidence intervals of 95%, so we can not rule 0 as the mean difference between the two supplements, so there is no significant evidence that one supplement is better than the other and we failed to reject Ho.

2. Unequal variance T test comparing dosages for both supplement

We can use *split()* function to split the data into dosages

```
tgdose <- split(ToothGrowth, ToothGrowth$dose)
```

We can now compare the groups 0.5 to 1 and 1 to 2 and test the null hypothesis that the difference between their means are equal to 0. Assuming unequal variances for all

```
t.test(tgdose[["1"]]$len, tgdose[["0.5"]]$len, paired = FALSE, var.equal = FALSE)$p.value
```

```
## [1] 1.268301e-07
```

```
t.test(tgdose[["2"]]$len, tgdose[["1"]]$len, paired = FALSE, var.equal = FALSE)$p.value
```

```
## [1] 1.90643e-05
```

Since the P-value is the probability under null hypothesis of obtaining evidence as extreme or more extreme than the obtained and on both cases the p-value is lower than 5%, we can reject our Null hypothesis and conclude that the dosage has an effect when both supplements considered together.

Conclusions and the assumptions needed for your conclusions

We can not say that one supplement has a better effect than the other because we failed to reject our null hypothesis that both means were equal. There is definitely an increase in tooth length when comparing lower to higher dosages if both supplements are grouped together. 1. We have assumed that supplement and dose amount have no confunding effect on the measured length. 2. We have assumed there is nothing else that has affected the length of the tooth, like diet or genetic factors. This assumption is based on a random selection of subjects. 3. We assumed that both factors, dose and supplent are independent.