Investigation of the ToothGrowth Dataset

Data Exploration

We load the data from the "ToothGrowth" dataset and have an initial look at it in order to decide what (if any) processing is necessary.

```
data("ToothGrowth")
head(ToothGrowth, n=5)
```

```
##    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
```

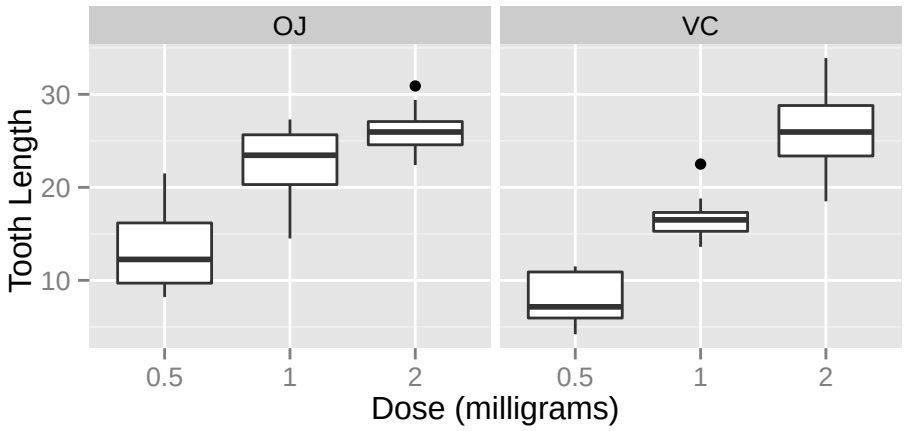
The help page for the ToothGrowth data set shows us that this dataset contains data about the length of teeth in Guinea Pigs. There are three colums. len - the ToothLength, supp - the type of supplement given to the animal (Vitamin C (VC) or Orange Jice (OJ)) and dose - the dose of the supplement given in milligrams.

```
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 summarise the data, grouping by sup and dose and computing mean lengths and use this to plot the mean length of teeth against dose as shown in the boxplots below. We see that there does appear to be a relationship between dose and the length of the teeth. There are too few data points to speculate on the nature of the relationship although naively it initially appears to be somewhat linear.

```
summary <- group_by(ToothGrowth,supp, dose) %>%
  summarise(MeanLength=mean(len))
ggplot(ToothGrowth, aes(y=len,x=factor(dose))) + geom_boxplot() + facet_grid(. ~ supp, scales="free_y") + 
```



The boxplots also suggest that Orange Juice has a greater effect than Vitamin C on the response but the spread of the data makes this far from certain and this bears more detailed examination.

Data Manipulation

We now need to reshape the data in order to facilitate computing differences (ie we want to go from a narrow to wide format).

```
# Using Reshape/dcast would be nice but lack of uniqueness in labels makes it PAINFUL!
VC <- ToothGrowth[ToothGrowth$sup == "VC",]
VCbyDose <- as.data.frame(with(VC, split(len, dose)))
names(VCbyDose) <- gsub("X", "Dose", names(VCbyDose), fixed=TRUE)

OJ <- ToothGrowth[ToothGrowth$sup == "OJ",]
OJbyDose <- as.data.frame(with(OJ, split(len, dose)))
names(OJbyDose) <- gsub("X", "Dose", names(OJbyDose), fixed=TRUE)
```

In-depth analysis

Assumptions

- We are dealing with independent and identically distributed variables.
- We have a large enough sample set for the Central Limit Theorem to be valid so that the means of our IIDs follow a normal distribution.

NB We do **not** assume that both populations have common variance.

Hypothesis -The response of the Guinea Pigs Tooth Length is proportional to the dose of the supplement that they receive.

We perform Students t tests for the difference in means where we subtract the length for lower dose from that of the higher dose. This means that if the 95% confidence interval is entirely above zero then we have a strong indicator that upping the dose increases the length of the teeth.

```
a <- t.test(VCbyDose$Dose1 - VCbyDose$Dose0.5, paired=FALSE, var.equal=FALSE)
c(a$conf.int, a$estimate)
```

```
##                      mean of x
##  5.549601 12.030399  8.790000
```

```
a <- t.test(VCbyDose$Dose2 - VCbyDose$Dose1, paired=FALSE, var.equal=FALSE)
c(a$conf.int, a$estimate)
```

```
##                      mean of x
##  5.405082 13.334918  9.370000
```

Whe the supplement used is Vitamin c, we see that for both tests the bottom of the confidence interval is well above zero which gives us a strong indication that increasing the dose, increases the length of the teeth.

```
a <- t.test(OJbyDose$Dose1 - OJbyDose$Dose0.5, paired=FALSE, var.equal=FALSE)
c(a$conf.int, a$estimate)
```

```
##                      mean of x
##  4.324616 14.615384  9.470000
```

```
a <- t.test(OJbyDose$Dose2 - OJbyDose$Dose1, paired=FALSE, var.equal=FALSE)
c(a$conf.int, a$estimate)
```

```
##                          mean of x
## -0.5509376  7.2709376  3.3600000
```

For Orange juice the story is not so clear. Between the 1ml and 0.5 ml doses things are similar to Vitamin C with the entire confidence interval above zero and the mean of the difference well above 0. However the difference between the 1ml and 2ml doses has confidence interval that dips under 0 which means that we cannot say that that is a 0.95 probability that the difference is positive. This can also be seen from the box plot - the top of the 1ml box overlaps the bottom of the 2ml one. Having said this, the mean os well above zero and the interval barely dips under zero so the indication still is that increasing dose, increases the teeth length.

Hypothesis - Orange juice has a greater effect on tooth length than Vitamin C

We now want to investigate which of the two supplements has the greater effect. We evaluate further t.tests of the difference of the means of the length for each of the two supplements for each dosage level - we subtract the Vitamin C value from the Orange Juice one so a positive value indicates increased response from Vitamin C.

```
a <- t.test(OJbyDose$Dose0.5 - VCbyDose$Dose0.5, paired=FALSE, var.equal=FALSE)
c(a$conf.int, a$estimate)
```

```
##                      mean of x
##  1.263458  9.236542  5.250000
```

```
a <- t.test(OJbyDose$Dose1 - VCbyDose$Dose1, paired=FALSE, var.equal=FALSE)
c(a$conf.int, a$estimate)
```

```
##                      mean of x
##  1.951911  9.908089  5.930000
```

```
a <- t.test(OJbyDose$Dose2 - VCbyDose$Dose2, paired=FALSE, var.equal=FALSE)
c(a$conf.int, a$estimate)
```

```
##                      mean of x
## -4.328976  4.168976 -0.080000
```

We see that for the 0.5ml and 1ml doses the confidence intervals are well above zero as is the mean which clearly indicates that there is a stronger response to Orange Juice than there is to Vitamin C.

However for the 2ml dose, the confidence interval is pretty much centred around 0 and the mean difference is approximately -0.1 which suggests that in he sample, for the 2ml does we see approximately the same response from both supplements and in fact it suggests very slightly that the stronger response is from Vitamin C.

Further data and analysis is necessary here to investigate whether this is a genuine effect or whether there is either some experimental error or error in the analysis. Special attention should be paid to whether the sample size is large enough to make the Central Limit Theorem valid and the distribution of means to truly be normal.

Conclusions

- Increasing the dose of both of the supplements generally increases the length of the Guinea Pigs teeth.
- Orange Juice appears to induce larger changes than Vitamin C when used as a supplement.
- Further analysis and/or data is needed to aacount for the anomolous nature of the data for 2ml dosage.