Tooth Growth Data Analysis

This is my reproducible markdown script for peer assignment 1 part B of the Coursera course Statistical In ference.

By Robert Merriman

Course reference URL

The ToothGrowth data reference contains "*The response is 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).*"

In this project I will analyze the ToothGrowth data in the R datasets package and:

1. Load the ToothGrowth data and perform some basic exploratory data analyses.
2. Provide a basic summary of the data.
3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.
4. State my conclusions and the assumptions needed for my conclusions.

First, read the data and modify some basic data structure

```
# Load the package "datasets"
library(datasets)
# Load the ToothGroum data
data(ToothGrowth)
# Change the dose to a factor
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

Explore the data

When exploring data I like to start with the basics such str, head, summary, and as we're working with two variables use a boxplot showing mean, quantiles, and outliers. I used and modified the code sample from the boxplot help content.

```
# Explore the data
# look at the structure of the data
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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

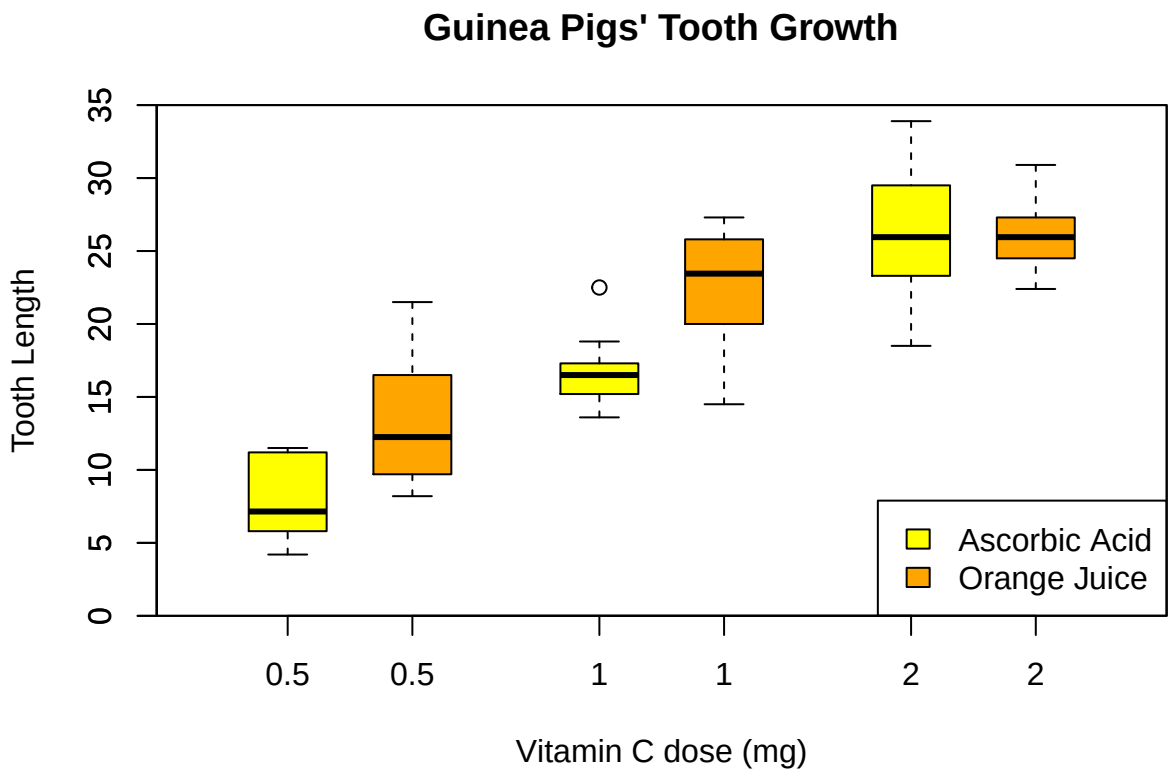
```
# look at a few rows of data
head(ToothGrowth, 3)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
```

```r
# look at basic statistics
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
```

```r
# as we have two variables, use a boxplot
boxplot(len ~ dose, data = ToothGrowth,
        boxwex = 0.25, at = 1:3 - 0.2,
        subset = supp == "VC", col = "yellow",
        main = "Guinea Pigs' Tooth Growth",
        xlab = "Vitamin C dose (mg)",
        ylab = "Tooth Length",
        xlim = c(0.5, 3.5), ylim = c(0, 35), yaxs = "i")
boxplot(len ~ dose, data = ToothGrowth, add = TRUE,
        boxwex = 0.25, at = 1:3 + 0.2,
        subset = supp == "OJ", col = "orange")
legend("bottomright", c("Ascorbic Acid", "Orange Juice"),
       fill = c("yellow", "orange"))
```



Observations from the boxplot:

- There is an outlier for ascorbic acid at 1 mg dosage (length = 22.5). As it was a recorded instance I will keep it in for analysis.
- For orange juice increasing dosage had increasing affect on the median of the length of the teeth.
- For ascorbic acid increasing dosage had increasing affect on the median of the length of the teeth.
- The median increase of the length of the teeth is higher for orange juice at dosage of 0.5 and 1.0 mg when compared to the median increase when ascorbid acid was the supplement.
- At 0.5 dosage, orange juice appears to have significantly more affect however it is not clear if it is statistically significant as the interquartiles for both orange juice and ascorbic acid overlap.
- At 1.0 dosage, orange juice appears to have significantly more affect that may be due to more than random chance.

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

Reference r-bloggers post on two sample Student t-Test

First I will split the data by supplement and dose. I will start with a null hypothesis to confirm if I should do the sample Student t-Test with equal or unequal variance. My null hypothesis is that the variance within each of the populations is equal and will be tested with a alpha level of 0.05%. I will use the Fisher's F-test to verify the homogeneity of variances.

```
# Split the data up by dose and supp
splitData <- split(ToothGrowth, list(ToothGrowth$supp, ToothGrowth$dose))
# test the variance of tooth growth length in each pair of supplement x dosage
VD <- NULL
VarData <- data.frame(Supp1=character(), Dosage1=character(),
      Supp2=character(), Dosage2=character(), p.value=double(),
      ComputedF=double(), TabValueF=double(), stringsAsFactors = F)

for (i in 1:5) {
    for (j in (i+1):6) {
        VD <- var.test(splitData[[i]]$len, splitData[[j]]$len)
        RD <- c(as.character(splitData[[i]]$supp[1]), as.character(splitData[[i]]$dose[1]),
                as.character(splitData[[j]]$supp[1]), as.character(splitData[[j]]$dose[1]),
                as.double(round(VD$p.value,4)), as.double(round(VD[1]$statistic[1],4)),
                round(qf(0.95, VD$parameter[1] , VD$parameter[2]),4))
        VarData[nrow(VarData)+1,] <- RD}}
```

As shown in the following table, for every comparison, the computed F value is less than the tabulated value of F and the p-value is > 0.05 therefore the null hypothesis of homogeneity of variance is accepted. Therefore when doing the sample Student t-Tests I will use the parameter var.equal = TRUE.

```
# show the data from the Fisher's F-test
VarData
```

```
##    Supp1 Dosage1 Supp2 Dosage2 p.value ComputedF TabValueF
## 1     OJ     0.5    VC     0.5  0.1649    2.6364    3.1789
```

```
## 2     OJ     0.5    OJ       1   0.702    1.3003    3.1789
## 3     OJ     0.5    VC       1  0.1032    3.1436    3.1789
## 4     OJ     0.5    OJ       2  0.1383    2.8214    3.1789
## 5     OJ     0.5    VC       2  0.8313    0.8641    3.1789
## 6     VC     0.5    OJ       1  0.3072    0.4932    3.1789
## 7     VC     0.5    VC       1  0.7975    1.1924    3.1789
## 8     VC     0.5    OJ       2  0.9212    1.0702    3.1789
## 9     VC     0.5    VC       2   0.112    0.3277    3.1789
## 10    OJ       1    VC       1  0.2046    2.4176    3.1789
## 11    OJ       1    OJ       2   0.264    2.1698    3.1789
## 12    OJ       1    VC       2  0.5523    0.6645    3.1789
## 13    VC       1    OJ       2  0.8747    0.8975    3.1789
## 14    VC       1    VC       2  0.0678    0.2749    3.1789
## 15    OJ       2    VC       2  0.0927    0.3063    3.1789
```

Compute the sample Student t-Tests and state my conclusions and the assumptions needed for my conclusions.

Now I can run the sample Student t-Tests using var.equal = TRUE. My null hypothesis for each pair is that the means are significantly similar with an alternate hypothesis that the means are statisitcally significant.

```
# Run the sample Student t-Tests for each pair of supplement x dosage
TD <- NULL
RD <- NULL
TData <- data.frame(Supp1=character(), Dosage1=character(),
        Supp2=character(), Dosage2=character(), p.valueRounded=double(),
        CI95Lower=double(), CI95Upper=double(), stringsAsFactors = F)

for (i in 1:5) {
    for (j in (i+1):6) {
        TD <- t.test(splitData[[i]]$len, splitData[[j]]$len, var.equal = T)
        RD <- c(as.character(splitData[[i]]$supp[1]), as.character(splitData[[i]]$dose[1]),
                as.character(splitData[[j]]$supp[1]), as.character(splitData[[j]]$dose[1]),
                as.double(round(TD$p.value,6)), round(TD$conf.int[1],4), round(TD$conf.int[2],4))
        TData[nrow(TData)+1,] <- RD}}
```

My Conclusion and assumptions

As shown in the following table for all but two comparisons the p-value was less than 0.05 and the 95th confidence intervals (CIs) did not contain 0 therefore the means are statistically significant. For the two comparisons where the p-value was > 0.05 therefore the null hypothesis of the means are significantly similar is accepted.

```
# show the data from the sample Students t-Test
TData
```

```
##    Supp1 Dosage1 Supp2 Dosage2 p.valueRounded CI95Lower CI95Upper
## 1     OJ     0.5    VC     0.5       0.005304    1.7703    8.7297
## 2     OJ     0.5    OJ       1        8.4e-05  -13.4108   -5.5292
## 3     OJ     0.5    VC       1        0.04224   -6.9417   -0.1383
## 4     OJ     0.5    OJ       2              0  -16.2782   -9.3818
## 5     OJ     0.5    VC       2          7e-06  -17.2619   -8.5581
```

```
## 6     VC     0.5    OJ       1             0  -17.8951  -11.5449
## 7     VC     0.5    VC       1          1e-06  -11.2643   -6.3157
## 8     VC     0.5    OJ       2             0   -20.618   -15.542
## 9     VC     0.5    VC       2             0  -21.8328  -14.4872
## 10    OJ       1    VC       1      0.000781    2.8407    9.0193
## 11    OJ       1    OJ       2      0.037363   -6.5005   -0.2195
## 12    OJ       1    VC       2      0.095837   -7.5523    0.6723
## 13    VC       1    OJ       2             0  -11.7198   -6.8602
## 14    VC       1    VC       2       3.4e-05   -12.969    -5.771
## 15    OJ       2    VC       2       0.96371    -3.723     3.563
```

My assumptions are:

- That the guinea pigs used in this study are repesentative of the population of guinea pigs.
- That the guinea pigs in this sample were independently and randomly chosen from a normally distributed population of guinea pigs with unknown but equal variance.
- That the dosage and supplement were randomly assigned to each guinea pig.