Exploratory Analysis and Confidence Interval testing of ToothGrowth data

Overview:

We analyse the ToothGrowth dataset with 1. Exploratory Graphs and 2. t-tests comparing the means of the outcome (len) for various combinations of the covariates (supp, dose)

Exploratory Data Anaysis

```
summary(ToothGrowth)
```

```
##       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  
```

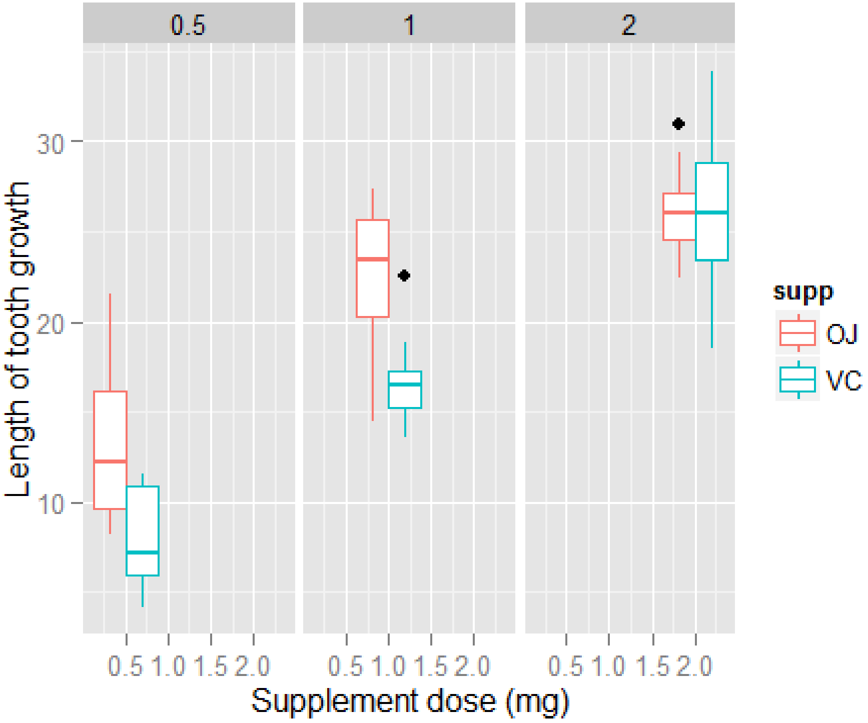
From the above summary, we see that 'len' is a numeric variable which is continuously distributed. 'dose' and 'supp' are factor variables having three and two levels repectively.

We plot a box chart of the variable 'len', separated and faceted by the 6 unique combinations of the other two (predictor) variables 'supp' & 'dose'

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.1.3
```

```
g <- ggplot(ToothGrowth, aes(x = dose, y = len)) + geom_boxplot(aes(y =
len, colour = supp)) + facet_grid(.~dose)
g <- g + xlab("Supplement dose (mg)")
g <- g + ylab("Length of tooth growth")
g
```



Data subsetting

We separate the dataset into 6 subsets - each corresponding to a unique combination of 'supp' ('OJ', 'VC') and 'dose' (0.5, 1.0, 2.0) For convenience, we then retain only the 'len' column.

```
library(dplyr)

##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

toogro_VC0.5 <-
filter(ToothGrowth,(ToothGrowth$dose==0.5)&(ToothGrowth$supp=="VC"))
toogro_VC1.0 <-
filter(ToothGrowth,(ToothGrowth$dose==1.0)&(ToothGrowth$supp=="VC"))
toogro_VC2.0 <-
filter(ToothGrowth,(ToothGrowth$dose==2.0)&(ToothGrowth$supp=="VC"))
toogro_OJ0.5 <-
filter(ToothGrowth,(ToothGrowth$dose==0.5)&(ToothGrowth$supp=="OJ"))
toogro_OJ1.0 <-
filter(ToothGrowth,(ToothGrowth$dose==1.0)&(ToothGrowth$supp=="OJ"))
toogro_OJ2.0 <-
filter(ToothGrowth,(ToothGrowth$dose==2.0)&(ToothGrowth$supp=="OJ"))
```

```
toogro_VC0.5 <- toogro_VC0.5[,1]
toogro_VC1.0 <- toogro_VC1.0[,1]
toogro_VC2.0 <- toogro_VC2.0[,1]
toogro_OJ0.5 <- toogro_OJ0.5[,1]
toogro_OJ1.0 <- toogro_OJ1.0[,1]
toogro_OJ2.0 <- toogro_OJ2.0[,1]
```

Finally, we also create 3 subsets corresponding to each level of 'dose':

```
toogro_0.5 <- as.numeric(rbind(toogro_OJ0.5, toogro_VC0.5))
toogro_1.0 <- as.numeric(rbind(toogro_OJ1.0, toogro_VC1.0))
toogro_2.0 <- as.numeric(rbind(toogro_OJ2.0, toogro_VC2.0))
```

Hypothesis testing

The idea (basis: the exploratory chart + theory) is that Vitamin C at higher dosage and delivered through orange juice rather than ascorbic acid leads to higher tooth growth in Guinea pigs.

We use Student's t-tests to compare the means of our data subsets. 1. H0: The means of 'len' of group (OJ, 0.5) and group (VC, 0.5) are equal

```
t.test(toogro_OJ0.5, toogro_VC0.5)

## 
##  Welch Two Sample t-test
## 
## data:  toogro_OJ0.5 and toogro_VC0.5
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean of x mean of y 
##     13.23      7.98
```

Interpretation: The 95% confidence interval for the difference between the two means is (1.72, 8.78). Zero does not lie in this interval. Further we have a large value of the t-statistic (3.1697) with a small p-value (0.006359). In other words, the probability under the null hypothesis of obtaining such a high value of the t-statistic is 0.6%. Considering an alpha (type 1 error) threshold of 5%, we can reject the null hypothsis H0. Hence, the means of 'len' of group (OJ, 0.5) and group (VC, 0.5) are not equal. The mean of the first group is larger than the second and the difference is between 1.72 & 8.78 with a 95% confidence level.

2. H0: The means of 'len' of group (OJ, 1.0) and group (VC, 1.0) are equal

```
t.test(toogro_OJ1.0, toogro_VC1.0)

## 
##  Welch Two Sample t-test
## 
```

```
## data:  toogro_OJ1.0 and toogro_VC1.0
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean of x mean of y
##     22.70     16.77
```

Result: H0 rejected. The means of 'len' of group (OJ, 1.0) and group (VC, 1.0) are not equal. The mean of the first group is larger than the second and the difference is between 2.8 & 9.06 with a 95% confidence level.

3. H0: The means of 'len' of group (OJ, 1.0) and group (VC, 1.0) are equal

```
t.test(toogro_OJ2.0, toogro_VC2.0)

## 
##  Welch Two Sample t-test
## 
## data:  toogro_OJ2.0 and toogro_VC2.0
## t = -0.0461, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

Result: H0 cannot be rejected - p-value is 0.9 = 90%.

4. H0: The means of 'len' of group (0.5) and group (1.0) are equal

```
t.test(toogro_0.5, toogro_1.0)

## 
##  Welch Two Sample t-test
## 
## data:  toogro_0.5 and toogro_1.0
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

Result: H0 rejected. The means of 'len' of group (0.5) and group (1.0) are not equal. The mean of the first group is smaller than the second and the difference is between -11.98 & -6.28 with a 95% confidence level.

5. H0: The means of 'len' of group (1.0) and group (2.0) are equal

```
t.test(toogro_1.0, toogro_2.0)
```

```
##
##  Welch Two Sample t-test
##
## data:  toogro_1.0 and toogro_2.0
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

Result: H0 rejected. The means of 'len' of group (OJ, 1.0) and group (VC, 2.0) are not equal. The mean of the first group is smaller than the second and the difference is between -9 & -3.73 with a 95% confidence level.

6. H0: The means of 'len' of group (0.5) and group (2.0) are equal

```
t.test(toogro_0.5, toogro_2.0)

##
##  Welch Two Sample t-test
##
## data:  toogro_0.5 and toogro_2.0
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean of x mean of y
##    10.605    26.100
```

Result: H0 rejected. The means of 'len' of group (OJ, 0.5) and group (VC, 2.0) are not equal. The mean of the first group is smaller than the second and the difference is between -18.15 & -12.83 with a 95% confidence level.

Conclusions & Assumptions

1. At dosage levels 0.5 mg and 1.0 mg of vitamin C, Orange Juice as the delivery medium is associated with greater efficacy as regards tooth growth. At dosage level 2.0 mg, the difference in tooth growth between the two delivery methods is statistically insignificant.

2. Higher dosage of vitamin C is associated with higher tooth growth.

Assumptions: 1. The sample, although small, is representative and unbiased 2. There are no other predictor (confounding) variables apart from 'supp' and 'dose' 3. The extent of this analysis is to establish correlation - not causality