ToothGrowth Data Analysis

Ricardo Fernandez

Overview

The ToothGrowth dataset explains the relation between the growth of teeth of 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 and ascorbic acid).

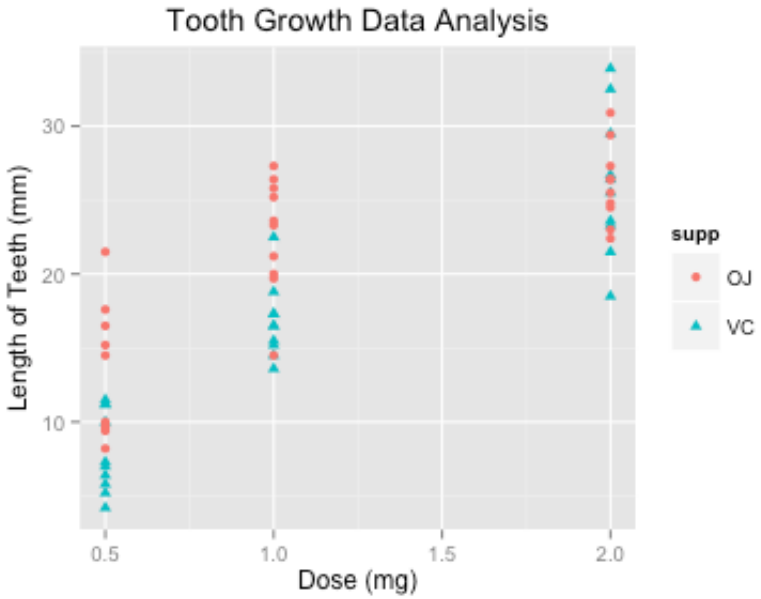
Basic exploratory data analysis

Has been performed a basic exploratory data analysis in order to overview the data structure of the ToothGrowth data library. We will check the 3 first an last lines of the data and the basic structure of the data set.

```
> head(ToothGrowth, n = 3)
   len supp dose
1  4.2   VC  0.5
2 11.5   VC  0.5
3  7.3   VC  0.5
> tail(ToothGrowth, n = 3)
    len supp dose
58 27.3   OJ    2
59 29.4   OJ    2
60 23.0   OJ    2
```

```
> 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 ...
```

As we can see the data is divided by lenght of teeth `len`, dose level `dose` and delivery method `supp`.



Basic summary of the data.

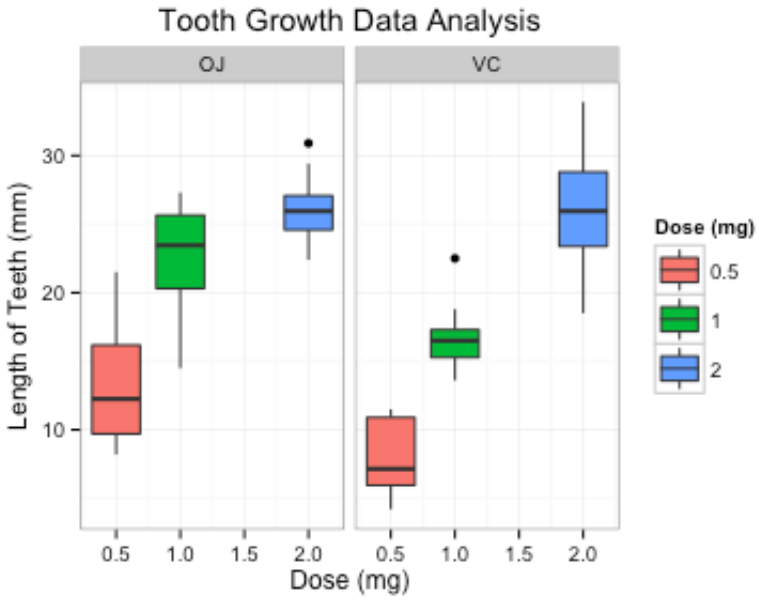
Statistical summary of the data, table summary by delivery method `supp` and dose level `dose` and statistical summary grouped by delivery method and dose level.

```
> 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
> table(ToothGrowth$supp, ToothGrowth$dose)

    0.5  1  2
  OJ  10 10 10
  VC  10 10 10
```

```
> by(ToothGrowth$len, INDICES = list(ToothGrowth$supp, ToothGrowth$dose), summary)
: OJ
: 0.5
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.20    9.70   12.25   13.23   16.18   21.50
------------------------------------------------
: VC
: 0.5
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.20    5.95    7.15    7.98   10.90   11.50
------------------------------------------------
: OJ
: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.50   20.30   23.45   22.70   25.65   27.30
------------------------------------------------
: VC
: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.60   15.27   16.50   16.77   17.30   22.50
------------------------------------------------
: OJ
: 2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22.40   24.58   25.95   26.06   27.08   30.90
------------------------------------------------
: VC
: 2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.50   23.38   25.95   26.14   28.80   33.90
```

Also a graphic overview of the summary data:



Confidence intervals and hypothesis tests.

```
suppT1 <- t.test(len~supp, paired=F, var.equal=T, data=ToothGrowth)
suppT2 <- t.test(len~supp, paired=F, var.equal=F, data=ToothGrowth)
suppRes <- data.frame("ConfLow"=c(suppT1$conf[1],suppT2$conf[1]),
                      "ConfHigh"=c(suppT1$conf[2],suppT2$conf[2]),
                      "PValue"=c(suppT1$p.value, suppT2$p.value),
                      row.names=c("Eq. Var.","Uneq. Var."))
```

```
> suppRes
              ConfLow ConfHigh     PValue
Eq. Var.   -0.1670064 7.567006 0.06039337
Uneq. Var. -0.1710156 7.571016 0.06063451
```

Considering the results obtained we cannot conclude that there are any differences between both delivery methods orange juice and ascorbic acid. The p-value for both variances, equal and unequal, are large and the confidence intervals contain 0.

> If the confidence interval includes 0 we can say that there is no significant difference between the means of the two populations, at a given level of confidence.

Test by supplement.

```
> t.test(len ~ supp, data = ToothGrowth)

Welch Two Sample t-test

data:  len by supp
t = 1.9153, df = 55.309, p-value = 0.06063
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1710156  7.5710156
sample estimates:
mean in group OJ     mean in group VC
        20.66333             16.96333
```

From the results of the 95% confidence interval `[-0.1710156 7.5710156]` for both means (group OJ `20.66333` and group VC `16.96333`) we cannot reject the null hipothesis and conclude if there is a significant difference in the two supplement types.

Test by dose.

```
> tgDose0.5_1 <- subset(ToothGrowth, dose %in% c(0.5, 1.0))
> t.test(len ~ dose, data = tgDose0.5_1)

    Welch Two Sample t-test

data:  len by dose
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 in group 0.5   mean in group 1
           10.605            19.735
```

```
> tgDose1_2 <- subset(ToothGrowth, dose %in% c(1.0, 2.0))
> t.test(len ~ dose, data = tgDose1_2)

    Welch Two Sample t-test

data:  len by dose
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 in group 1 mean in group 2
         19.735          26.100
```

From results obtained for the intervals `[0.5, 1.0]` and `[1.0, 2.0]` we can reject the null hypothesis and conclude that there is difference of teeth length depending of the dose administered. From the results of the 95% confidence interval `[-11.983781, -6.276219]` for both means group 0.5 `10.605` and group 1 `19.735` ; and `[-8.996481 -3.733519]` for both means group 1 `19.735` and group 2 `26.100` .

Concusions

This conclusions are based in some statistical limitations, as for example, the data is randomly generated and the results only can be generalized if the samples population is representative of the entry population. Of this two assumptions are considered in the experiment we can conclude that:

- The supplement type do no effect the tooth growth.
- If we increase the dose level the tooth growth tend to be increased.

Apendix

Code

```
library(ggplot2); library(datasets);
data(ToothGrowth)
# Check the first and last 5 lines of the data and the basic structure.
head(ToothGrowth, n = 5)
tail(ToothGrowth, n = 5)
str(ToothGrowth)
# Plot the data
ggp <- ggplot(aes(x = dose, y = len), data = ToothGrowth) +
  geom_point(aes(color=supp, shape = supp)) +
  labs(title=expression("Tooth Growth Data Analysis"),
     x = "Dose (mg)", y = expression("Length of Teeth (mm)"))
print(ggp)
# Provide a basic summary of the data.
summary(ToothGrowth)
table(ToothGrowth$supp, ToothGrowth$dose)
by(ToothGrowth$len, INDICES = list(ToothGrowth$supp, ToothGrowth$dose), summary)
ggpsupp <- ggplot(aes(x = dose, y = len), data = ToothGrowth) +
  geom_boxplot(aes(fill = factor(dose))) + facet_grid(.~supp) + theme_bw() +
  guides(fill=guide_legend(title="Dose (mg)")) +
  labs(title=expression("Tooth Growth Data Analysis"),
       x = "Dose (mg)", y = expression("Length of Teeth (mm)"))
print(ggpsupp)
# Confidence intervals and hypothesis test compare tooth growth by supp and dose
suppT1 <- t.test(len~supp, paired=F, var.equal=T, data=ToothGrowth)
suppT2 <- t.test(len~supp, paired=F, var.equal=F, data=ToothGrowth)
suppRes <- data.frame("ConfLow"=c(suppT1$conf[1],suppT2$conf[1]),
                      "ConfHigh"=c(suppT1$conf[2],suppT2$conf[2]),
                      "PValue"=c(suppT1$p.value, suppT2$p.value),
                      row.names=c("Eq. Var.","Uneq. Var."))
suppRes
# T-Test by supplemant
t.test(len ~ supp, data = ToothGrowth)
# T-test by dose
tgDose0.5_1 <- subset(ToothGrowth, dose %in% c(0.5, 1.0))
t.test(len ~ dose, data = tgDose0.5_1)
tgDose1_2 <- subset(ToothGrowth, dose %in% c(1.0, 2.0))
t.test(len ~ dose, data = tgDose1_2)
# T test for supplement by dose
tgDose0.5 <- subset(ToothGrowth, dose == 0.5)
t.test(len ~ supp, data = tgDose0.5)
tgDose1 <- subset(ToothGrowth, dose == 1.0)
t.test(len ~ supp, data = tgDose1)
tgDose2 <- subset(ToothGrowth, dose == 2.0)
t.test(len ~ supp, data = tgDose2)
```