Project part 2: tooth growth analysis

Sylvain Tenier

Overview

The ToothGrowth dataset describes the evolution in the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs, where each animal received one of three dose levels of vitamin C by one of two delivery methods. The complete description is provided at the R documentation site.

Exploratory Data Analysis

The dataset consists of a dataframe containing 60 observations composed of 3 attributes :

```
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 ...
```

- the Tooth length, described as a numeric variable named *len*,
- the Supplement type, a factor named *supp*. The levels of the factor are either *VC* or *OJ*,
- the Dose in milligrams/day expressed as a numeric variable named *dose*.

The documentation tells us that the tooth length is the output. For our Exploratory Data Analysis, we generate a boxplot for each dose and supplement type.

```
library(ggplot2)
ggplot(data=ToothGrowth, aes(x=factor(dose),y=len, color=supp))+
  facet_wrap(~ supp )+ geom_boxplot() +
    labs(title="Tooth length for a given dose, for each supplement type",
         x="Dose given", y="Length of tooth")
```
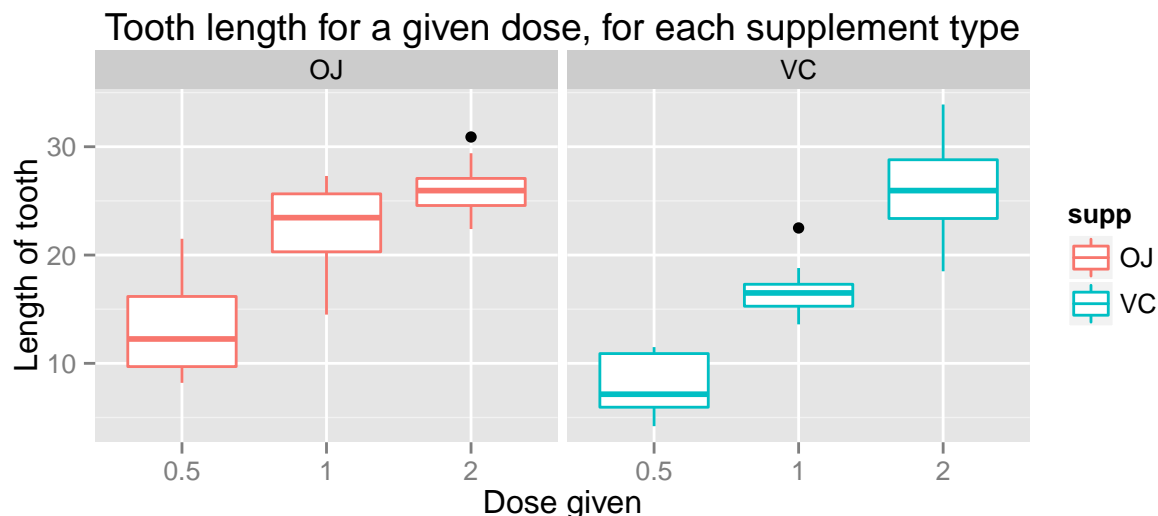
This figure shows that whatever the supplement type, a higher dose yields a higher growth. We can also see that the *OJ* supplement type is more efficient for small doses, but that the *OJ* and *VC* supplements type have similar mean when the highest dose (2.0) is used.

Summary of data

We group the data by dose and supplement type and calculate the minimum value, the mean, the maximum value and the standard deviation for each subgroup.

```
library(knitr)
suppressPackageStartupMessages(library(dplyr, quietly=TRUE))
tgsummary <- ToothGrowth %>%
    group_by(dose,supp) %>%
    summarise(count=n(), min=min(len), mean=mean(len), max=max(len), sd=sd(len))
kable(tgsummary)
```

dose	supp	count	min	mean	max	sd
0.5	OJ	10	8.2	13.23	21.5	4.459708
0.5	VC	10	4.2	7.98	11.5	2.746634
1.0	OJ	10	14.5	22.70	27.3	3.910953
1.0	VC	10	13.6	16.77	22.5	2.515309
2.0	OJ	10	22.4	26.06	30.9	2.655058
2.0	VC	10	18.5	26.14	33.9	4.797731

We can see that there are 10 measurements for each subgroup. The mean confirms that increasing the dose increases the result, and that for small doses the *OJ* is more efficient, although the maximum value is reached with the *VC* supplement at the highest dose.

Confidence interval

Test Assumptions

- the variance between each group is unequal,
- the groups are independant.

Calculations

We must now calculate the confidence intervals to expand on the validity of those preliminary results. We will make the calculations for all measurements, and then for each dose.

```
#calculate using t.test() for all, 0.5, 1.0 and 2.0
testall<- t.test(ToothGrowth[ToothGrowth$supp=='OJ',]$len,
                 ToothGrowth[ToothGrowth$supp=='VC',]$len,
                 var.equal = FALSE)
test05<- t.test(ToothGrowth[(ToothGrowth$supp=='OJ')&(ToothGrowth$dose==0.5),]$len,
        ToothGrowth[(ToothGrowth$supp=='VC')&(ToothGrowth$dose==0.5),]$len,
        var.equal = FALSE)
```

```
test10 <- t.test(ToothGrowth[(ToothGrowth$supp=='OJ')&(ToothGrowth$dose==1),]$len,
    ToothGrowth[(ToothGrowth$supp=='VC')&(ToothGrowth$dose==1),]$len,
    var.equal = FALSE)
test20 <- t.test(ToothGrowth[(ToothGrowth$supp=='OJ')&(ToothGrowth$dose==2),]$len,
    ToothGrowth[(ToothGrowth$supp=='VC')&(ToothGrowth$dose==2),]$len,
    var.equal = FALSE)
#create dataframe with results
dfall <- data.frame(row.names='all doses',pvalue=testall$p.value,
                confinf=testall$conf.int[1],confsup=testall$conf.int[2],
                OJmean=testall$estimate[1],VCmean=testall$estimate[2])
df05 <- data.frame(row.names='0.5 dose',pvalue=test05$p.value,
                confinf=test05$conf.int[1],confsup=test05$conf.int[2],
                OJmean=test05$estimate[1],VCmean=test05$estimate[2])
df10 <- data.frame(row.names='1.0 dose',pvalue=test10$p.value,
                confinf=test10$conf.int[1],confsup=test10$conf.int[2],
                OJmean=test10$estimate[1],VCmean=test10$estimate[2])
df20 <- data.frame(row.names='2.0 dose',pvalue=test20$p.value,
                confinf=test20$conf.int[1],confsup=test20$conf.int[2],
                OJmean=test20$estimate[1],VCmean=test20$estimate[2])
#display table with results
kable(total <- rbind(dfall, df05, df10, df20) )
```

	pvalue	confinf	confsup	OJmean	VCmean
all doses	0.0606345	-0.1710156	7.571016	20.66333	16.96333
O.5 dose	0.0063586	1.7190573	8.780943	13.23000	7.98000
1.0 dose	0.0010384	2.8021482	9.057852	22.70000	16.77000
2.0 dose	0.9638516	-3.7980705	3.638070	26.06000	26.14000

The t tests are carried with the H_0 hypothesis :

```
OJ and VL supplement types generate the same growth in teeth length
```

Conclusions

We can see that :

- the confidence interval for all results contains the 0 value,
- the confidence interval for small doses (0.5 and 1.0) does not contain the 0 value and have a small p value,
- the confidence interval for the highest dose (2.0) contains 0 and have a huge p value.

We can thus reject the H_0 hypothesis for the small doses, but not for the 2.0 and all doses.