Project 2: a study on the ToothGrowth database

0) Introduction

In this short note, we will do a simple statistical analysis of the ToothGrowth database contained in the datasets library. Note for the reviewers: since it is quite useless to go back and forth between the code put in appendix and the core of the project, I merged the two.

I) Exploratory analysis

We start by loading the datasets library

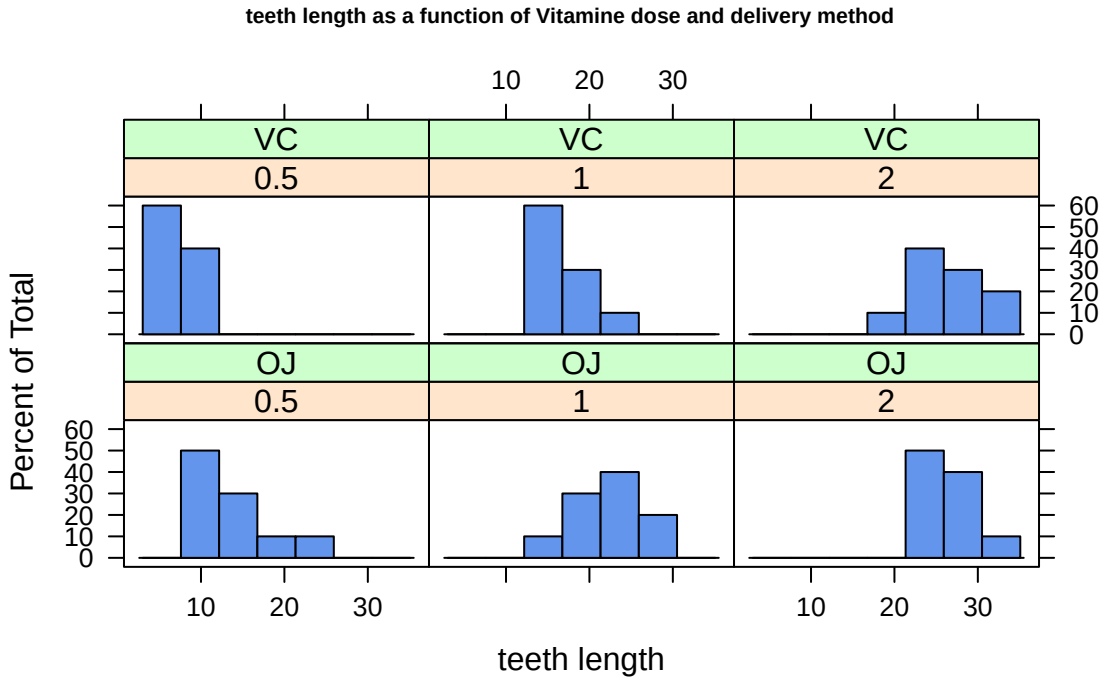
```
library(datasets)
 dim(ToothGrowth)
```

we see that ToothGrowth is a 60×3 dataframe. The first column (using `?ToothGrowth`) corresponds to the length of teeth in each of 10 guinea pigs, the second column to two different delivery methods (orange juice or ascorbic acid). The thirs column corresponds to three different dose levels of Vitamin C (0.5, 1, and 2 mg). Since the Vitamine dose and the delivery method are qualitative variables, we will convert them into factors

```
ToothGrowth$supp<-as.factor(ToothGrowth$supp)
 ToothGrowth$dose<-as.factor(ToothGrowth$dose)
```

Let us consider following the exploratory graph

```
 library(lattice)
histogram(~len|dose*supp,data=ToothGrowth, xlab="teeth length",
    main=list(label="teeth length as a function of Vitamine dose and delivery method"
              ,cex=0.7),col="cornflowerblue")
```



and a summary of the data frame gives

```
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
```

The histograms represent the length of the teeth as a function of both the delivery method and the vitamine dose. As one can see, as the Vitamine dose increases, the teeth length seems to increas as well. It also seems that the Orange Juice delivery method leads to slightly longer teeth than with ascorbic acid. This will be our working hypotheses for the next section.

I) Does the teeth length depends on the delivery method?

Firstly, does the teeth length depends on the delivery method? We can compute the mean and the standard deviation of both classes with

```
meanOJ<-mean(ToothGrowth$len[ToothGrowth$supp=="OJ"])
meanVC<-mean(ToothGrowth$len[ToothGrowth$supp=="VC"])
sdOJ<-sd(ToothGrowth$len[ToothGrowth$supp=="OJ"])
sdVC<-sd(ToothGrowth$len[ToothGrowth$supp=="VC"])
```

Our hypothesis H_0 is : the teeth length is the same whatever the delivery method, while our H_a hypothesis is that the OJ delivery method gives different teeth length than the VC one. Given the small size of our sample, we will do a confidence interval analysis with the quantile of the student-t distribution. We will assume equal variance. The relevant formulas are

$$[\text{conf int}] = \mu_{\text{OJ}} - \mu_{\text{VC}} \pm qt(0.975, N_{\text{OJ}} + N_{\text{VC}} - 2) \times \Sigma \times \sqrt{\frac{1}{N_{\text{OJ}}} + \frac{1}{N_{\text{VC}}}} \tag{1}$$

where the pooled standard error being

$$\Sigma = \sqrt{\frac{(N_{\text{OJ}} - 1)\sigma_{\text{OJ}}^2 + (N_{\text{VC}} - 1)\sigma_{VC}^2}{N_{\text{OJ}} + N_{\text{VC}} - 2}} \tag{2}$$

the R calculation gives

```
NOJ<-sum(ToothGrowth$supp=="OJ")
NVC<-sum(ToothGrowth$supp=="VC")
Sigma<-sqrt(((NOJ-1)*sdOJ^2+(NVC-1)*sdVC^2)/(NOJ+NVC-2))
meanOJ-meanVC+c(-1,1)*qt(0.975,NOJ+NVC-2)*Sigma*sqrt(1/NOJ+1/NVC)
```

```
## [1] -0.1670064  7.5670064
```

the same result can be obtained with

```
t.test(ToothGrowth$len[ToothGrowth$supp=="OJ"],
       ToothGrowth$len[ToothGrowth$supp=="VC"],paired=FALSE,var.equal=TRUE)$conf
```

```
## [1] -0.1670064  7.5670064
## attr(,"conf.level")
## [1] 0.95
```

Since 0 is in the 95% confidence interval, we **fail to reject the null Hypothesis** H_0.

I) Does the teeth length depends on the vitamine dose ?

Secondly, does the teeth length depends on the vitamine dose? This time we will just use the built-in R function. We will assume equal variance. Our hypothesis H_0 is : the mean of the length teeth does not depend on the vitamine dose, while our H_a hypothesis is that the vitamine dose affects the teeth length. Since the t-test is designed to compare between two groups, we will do three t-test: between the 0.5 and the 1mg doses, the 0.5 and the 2mg doses and finally between the 1 and the 2mg doses.

1. 0.5 vs 1mg dose:

```
t.test(ToothGrowth$len[ToothGrowth$dose=="1"],
       ToothGrowth$len[ToothGrowth$dose=="0.5"],paired=FALSE,var.equal=TRUE)$conf
```

```
## [1]  6.276252 11.983748
## attr(,"conf.level")
## [1] 0.95
```

2. 0.5 vs 2mg dose:

```
t.test(ToothGrowth$len[ToothGrowth$dose=="2"],
       ToothGrowth$len[ToothGrowth$dose=="0.5"],paired=FALSE,var.equal=TRUE)$conf
```

```
## [1] 12.83648 18.15352
## attr(,"conf.level")
## [1] 0.95
```

3. 1 vs 2mg dose:

```
t.test(ToothGrowth$len[ToothGrowth$dose=="2"],
       ToothGrowth$len[ToothGrowth$dose=="1"],paired=FALSE,var.equal=TRUE)$conf
```

```
## [1] 3.735613 8.994387
## attr(,"conf.level")
## [1] 0.95
```

In all cases, 0 is in not in the 95% confidence interval and we **reject the null Hypotheses** H_0.