StatInf_prog2

jjperez

Sunday, Septembre 27, 2015

Title: ToothGowth Dataset Analysis

0. Sinopsis

The objetive of this part of the project is analize the ToothGowth data frame. This data frame has 60 observations on 3 variables:
[,1] len numeric Tooth length
[,2] supp factor Supplement type (VC or OJ)
[,3] dose numeric Dose in milligrams/day

1. Loading data and exploratory data analysis

The first step will be load all the necesary libraries and the data frame ToothGrowth into R.

```
library(lubridate, quietly = TRUE,warn.conflicts=FALSE)
library(RColorBrewer, quietly = TRUE,warn.conflicts=FALSE)
library(lattice, quietly = TRUE,warn.conflicts=FALSE)
library(xtable, quietly = TRUE,warn.conflicts=FALSE)
library(knitr, quietly = TRUE,warn.conflicts=FALSE)
library(dplyr, quietly = TRUE,warn.conflicts=FALSE)

mns <- NULL
sds <- NULL

data(ToothGrowth)
```

ToothGroth has two numerical field, len and dose, and one alphanumerical field. For the two first I will graph the histogram and check if there is pattern.

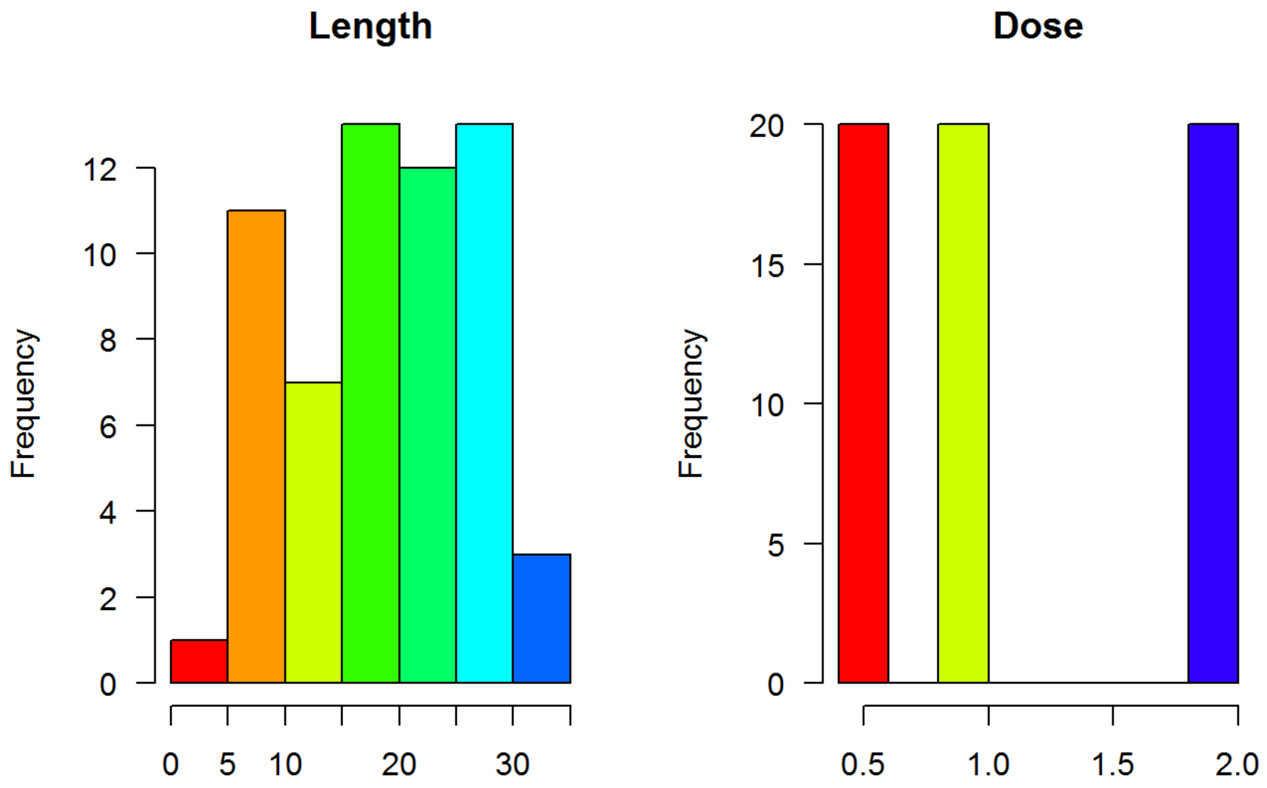
```
# How it looks like.
# First we convert supp into factors to count them.

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

# Now we plot the histograms on a combined panel.

par(mfrow = c(1,2))

h1<- hist(ToothGrowth$len, breaks = 10, xlab = "", main = "Length", col=rainbow(10
), las=1)
h2<- hist(ToothGrowth$dose, breaks = 10, xlab = "", main = "Dose", col=rainbow(10)
, las=1)
```



For the last one I will convert these values into factors and use the function **table** to check how many different values we have.

OJ, VC
30, 30

We can conclude that:
- **len**. Doesn´t have a clear structure but all the data seem to be grouped between 5 and 25
- **supp**. As we expected from the data frame description it only takes 2 different values: OJ, VC
- **dose**. This variable only takes 3 different values: 0.5, 1, 2

According to that we should analize **6** different cathegories. The last step is check if there is missing data into our data frame. To do that we will compare the number of complete cases with the lenght of the columns. If the numbers doesn´t match then there are rows with missing data.

```
# We check if there are some missing values on the given data
complete <- ifelse((sum(complete.cases(ToothGrowth))==length(ToothGrowth$len)), "N
o missing data", "There is missing data")
```

The result of that query shows: **No missing data**

This chunk of code remove the rows with missing data, in case they exist. Also we convert **dose** into a factor. It will be useful to perform the analysis.

```
if(complete == "There is missing data")
{
  ToothGrowth <- ToothGrowth[complete.cases(ToothGrowth),]
}
#ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

2. Analysis

Because the amount of data is not really big I will use the **t** function to estimate the confidence intervals. For each *__cathegory__ I will define a **95% confidence interval**. Recall we have **6** different cathegories.

2.1. Sample mean and standard deviations

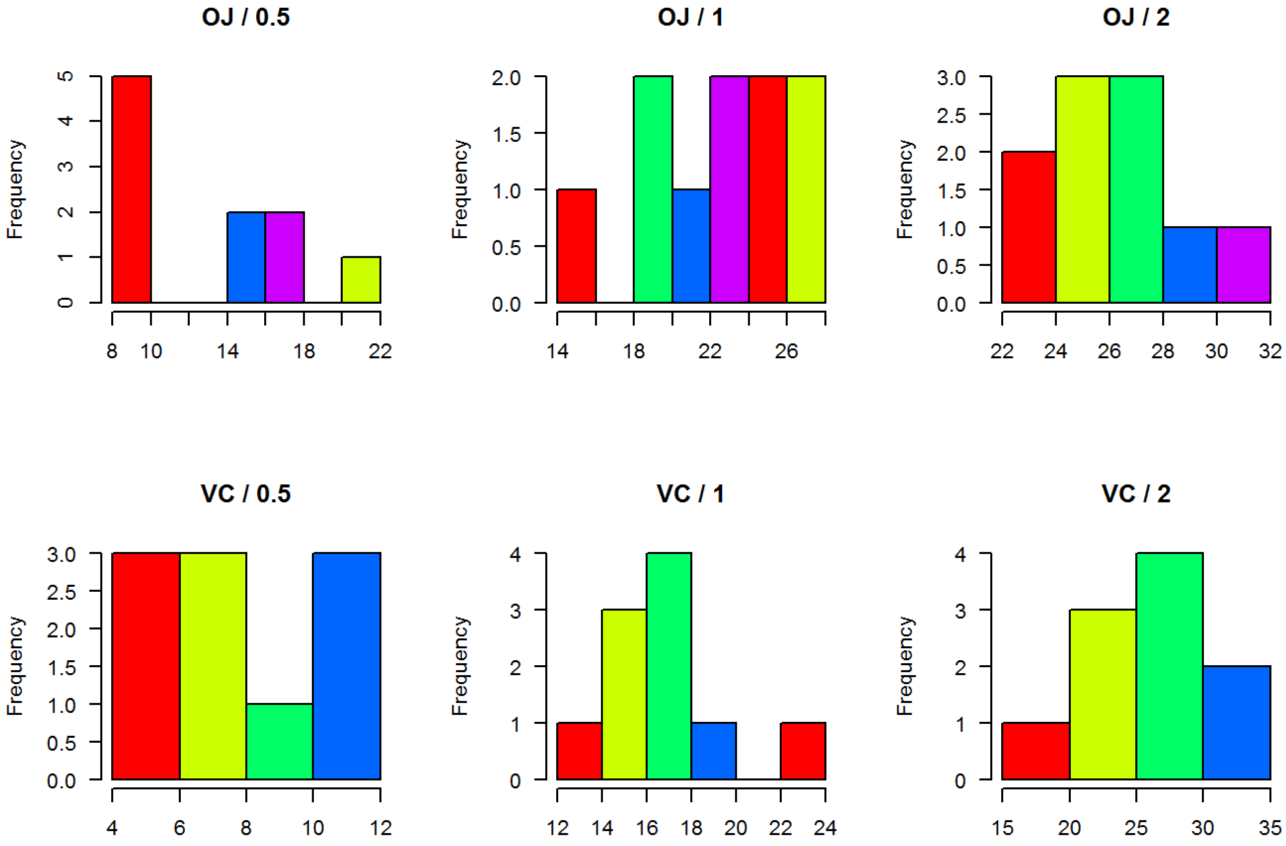
On this first step I will check how each cathegory looks like and I will calculate the mean and the standard deviation for each of them.

```
# We start calculating the expected values.

dat1 <- ToothGrowth[ ToothGrowth[,2]=="OJ" & ToothGrowth[,3]==0.5 , ]
dat2 <- ToothGrowth[ ToothGrowth[,2]=="OJ" & ToothGrowth[,3]==1 , ]
dat3 <- ToothGrowth[ ToothGrowth[,2]=="OJ" & ToothGrowth[,3]==2 , ]
dat4 <- ToothGrowth[ ToothGrowth[,2]=="VC" & ToothGrowth[,3]==0.5 , ]
dat5 <- ToothGrowth[ ToothGrowth[,2]=="VC" & ToothGrowth[,3]==1 , ]
dat6 <- ToothGrowth[ ToothGrowth[,2]=="VC" & ToothGrowth[,3]==2 , ]

par(mfrow = c(2,3))

h1<- hist(x = dat1[,1], main = "OJ / 0.5", xlab = " ", col=rainbow(5))
h2<- hist(x = dat2[,1], main = "OJ / 1", xlab =" ", col=rainbow(5), las=1)
h3<- hist(x = dat3[,1], main = "OJ / 2", xlab =" ", col=rainbow(5), las=1)
h4<- hist(x = dat4[,1], main = "VC / 0.5", xlab =" ", col=rainbow(5), las=1)
h5<- hist(x = dat5[,1], main = "VC / 1", xlab =" ", col=rainbow(5), las=1)
h6<- hist(x = dat6[,1], main = "VC / 2", xlab =" ", col=rainbow(5), las=1)
```



```
data <- matrix(NA, 6, 5)
colnames(data) <- c("supp","dose","mean","sd","n")

data[1,]<-c("OJ","0.5", round(mean(dat1[,1]),2), round(sd(dat1[,1]),2), length(dat1[,1]))
data[2,]<-c("OJ","1",round(mean(dat2[,1]),2), round(sd(dat2[,1]),2), length(dat2[,1]))
data[3,]<-c("OJ","2",round(mean(dat3[,1]),2), round(sd(dat3[,1]),2), length(dat3[,1]))
data[4,]<-c("VC","0.5",round(mean(dat4[,1]),2), round(sd(dat4[,1]),2), length(dat4[,1]))
data[5,]<-c("VC","1",round(mean(dat5[,1]),2), round(sd(dat5[,1]),2), length(dat5[,1]))
data[6,]<-c("VC","2",round(mean(dat6[,1]),2), round(sd(dat6[,1]),2), length(dat6[,1]))
```

supp, dose, mean, sd, n
OJ, 0.5, 13.23, 4.46, 10
OJ, 1, 22.7, 3.91, 10
OJ, 2, 26.06, 2.66, 10
VC, 0.5, 7.98, 2.75, 10
VC, 1, 16.77, 2.52, 10
VC, 2, 26.14, 4.8, 10

We can see how the maximum values for the average tooth length correspond to the highest values on the doses. Next step will be calculate the confidence intervals for each of them.

2.2. Confidence intervals

Due to the low number of observations we cannot use normal quantiles. Instead I will define the **t-confidence intervals** for each cathegorie. This 95% confidence interval will show the of values where the 95% of the measures should be.

```
# Last step will be calculate the confidence intervals for each cathegory

tconf <- matrix(NA, 6, 5)
colnames(tconf) <- c("supp","dose","LL","mean","UL")

tconf[1,]<-c("OJ","0.5", round(t.test(dat1[,1])$conf[1],2), round(mean(dat1[,1]),2
), round(t.test(dat1[,1])$conf[2],2) )
tconf[2,]<-c("OJ","1",round(t.test(dat2[,1])$conf[1],2), round(mean(dat2[,1]),2),
round(t.test(dat2[,1])$conf[2],2) )
tconf[3,]<-c("OJ","2",round(t.test(dat3[,1])$conf[1],2), round(mean(dat3[,1]),2),
round(t.test(dat3[,1])$conf[2],2) )
tconf[4,]<-c("VC","0.5",round(t.test(dat4[,1])$conf[1],2), round(mean(dat4[,1]),2)
, round(t.test(dat4[,1])$conf[2],2) )
tconf[5,]<-c("VC","1",round(t.test(dat5[,1])$conf[1],2), round(mean(dat5[,1]),2),
round(t.test(dat5[,1])$conf[2],2) )
tconf[6,]<-c("VC","2",round(t.test(dat6[,1])$conf[1],2), round(mean(dat6[,1]),2),
round(t.test(dat6[,1])$conf[2],2) )
```

supp, dose, LL, mean, UL
OJ, 0.5, 10.04, 13.23, 16.42
OJ, 1, 19.9, 22.7, 25.5
OJ, 2, 24.16, 26.06, 27.96
VC, 0.5, 6.02, 7.98, 9.94
VC, 1, 14.97, 16.77, 18.57
VC, 2, 22.71, 26.14, 29.57

3. Results

Looking to the results we can conclude:
- The average tooth lenght grows with the doses.
- With the supplement **OJ** at the same time the average tooth length increases the **standard deviation reduces**

4. Extras

You will find the R script in my GitHub repostory. You can go there through that (link)[https://github.com/jjperez78/StatisticalInference.git (https://github.com/jjperez78/StatisticalInference.git)]