Tooth Growth Analysis

Americo

19 luglio 2015

Utils

```
require(dplyr)
```

```
## Loading required package: 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
```

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
require(datasets)
require(tidyr)
```

```
## Loading required package: tidyr
```

```
options(digits = 3)
setwd("/Users/Americo/Documents/Education/Data_science/Coursera/Statinference/project/ass2_statinference
```

Goals

Now in the second portion of the class, we're going to analyze the ToothGrowth data in the R datasets package. Load the ToothGrowth data and perform some basic exploratory data analyses Provide a basic summary of the data. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering) State your conclusions and the assumptions needed for your conclusions. Some criteria that you will be evaluated on Did you perform an exploratory data analysis of at least a single plot or table highlighting basic features of the data? Did the student perform some relevant confidence intervals and/or tests? Were the results of the tests and/or intervals interpreted in the context of the problem correctly? Did the student describe the assumptions needed for their conclusions?

Dataset

Loading the dataset and studying the variables through the codebook and some initial exploration.

```
data(ToothGrowth)
?ToothGrowth
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
tail(ToothGrowth)
```

```
##     len supp dose
## 55 24.8   OJ    2
## 56 30.9   OJ    2
## 57 26.4   OJ    2
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

```
dim(ToothGrowth)
```

```
## [1] 60  3
```

```
str(ToothGrowth)
```

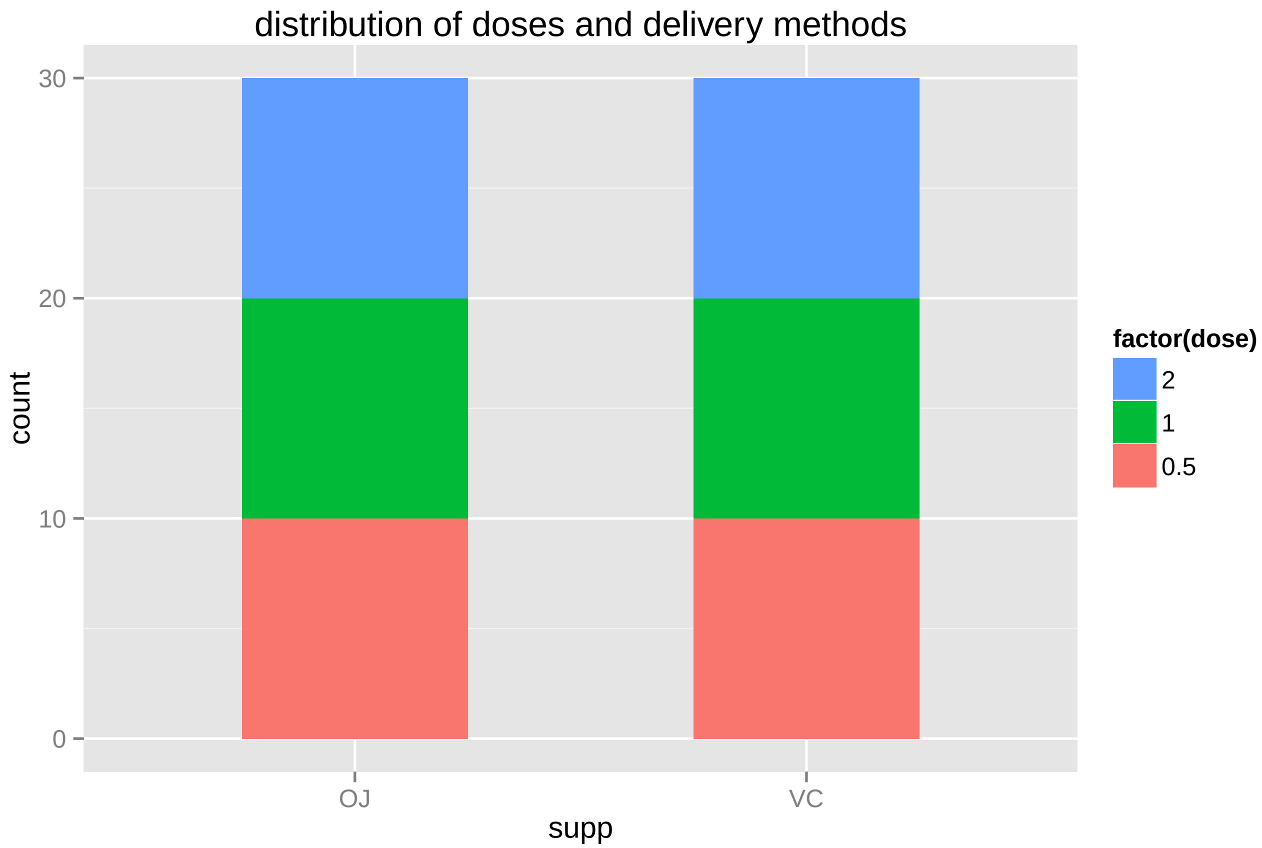
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

The dataset is composed by measurements on 10 pigs regarding the length of teeth (variable `len`) after the somministration of three different dose of vitamin (the variable `dose`) and two delivery methods (orange juice or ascorbic acid - the variable `supp`). There is not a column dedicated to the ID of the ten pigs, so I assume that the measurements are ordered this way: each ten observations represents a dose of vitamin, the first 30 with a delivery methods and the second 30 with the other one.

Exploratory data analysis

How delivery methods and doses are distributed?

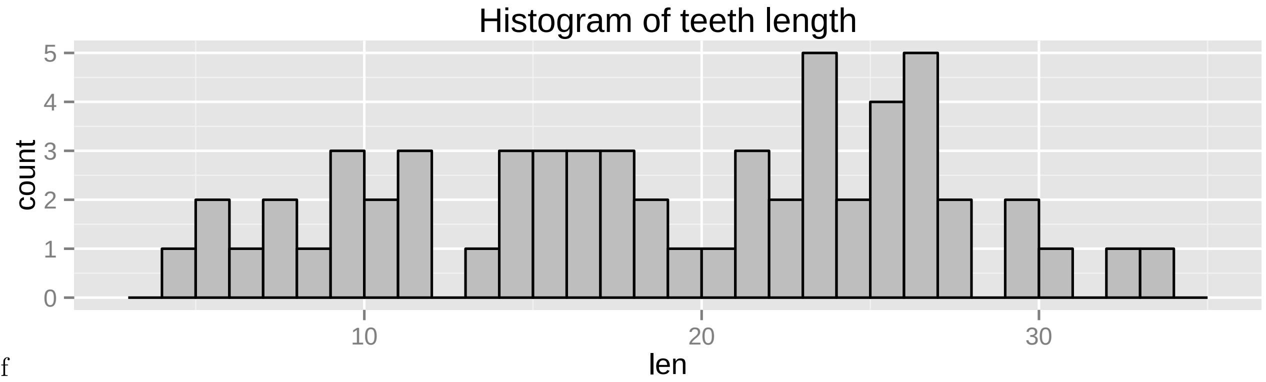
```
ggplot(data = myToothGrowth, aes(x = supp, fill = factor(dose))) +
        geom_bar(width = 0.5) +
        guides(fill=guide_legend(reverse=TRUE)) +
        ggtitle("distribution of doses and delivery methods")
```



My assumption (*I assume that the measurements are ordered this way: each ten observations represents a dose of vitamin, the first 30 with a delivery methods and the second 30 with the other one*) seem confirmed by the plot.

Let'see the distribution of teeth's lengths.

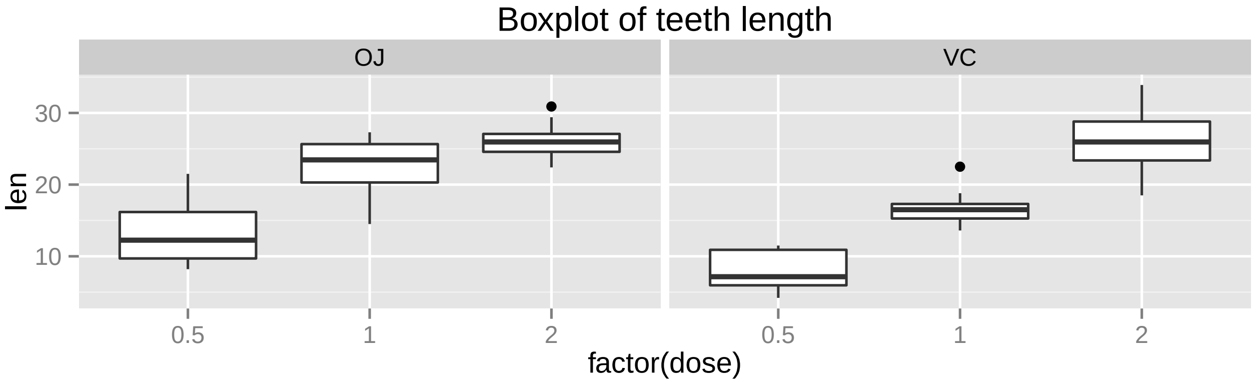
```
ggplot(data = myToothGrowth, aes(x = len)) +
             geom_histogram(binwidth = 1, colour = "black", fill = "gray") +
       ggtitle("Histogram of teeth length")
```



histogram-1.pdf

I don't know if this kind of exploration makes sense, because this 60 observations are not really *a population*, are much more the sum of 6 different “snapshot”" of a population. Let's plot them so.

```r
ggplot(data = myToothGrowth, aes(x = factor(dose), y = len)) +
        geom_boxplot() +
        facet_wrap(~supp) +
        ggtitle("Boxplot of teeth length")
```



boxplots-1.pdf

It seems that dose influences the growth of lenghts (very expected!) and that ascorbic acide is more effective than orange juice in facilitate this. Almost all population are quiete symmetrical, which is good for our further analysis. **Infact we are deailing with paired small samples**, so it's likely we are going to use t-test to compare groups, and t-test needs quite normal population to be accurate. Of course we can't check the normality of population (unless we use some tools that were not taugth in the class, but students have been explicitly discouraged in doing so), so we must perform analysis on these small samples. Furthermore, being paired samples, it's better to plot the differences between lengths, as we will test them with the `t distribution`. I will do this in the `statistical analysis` chapter, after some data manipulation.

Data manipulation

I want to create a dataframe with 7 variables, the 6 measurements and the ID of the subjects. It seems to me the better way to flexibly perform t-tests.

```r
myToothGrowth <- ToothGrowth
mydfs <- list()
j <- 1:10
for (i in 1:6) {
        mydfs[[i]] <- myToothGrowth[j,]
        names(mydfs[[i]]) <- c(paste("len", i), paste("supp", i),paste("dose", i))
        j <- j + 10
}
myToothGrowth2 <- bind_cols(mydfs)
myToothGrowth2$pig_id <- 1:10

#renaming variables to have a smaller dataframe
myToothGrowth3 <- myToothGrowth2[, seq(1, 19, 3)]
names(myToothGrowth3)  <- c("lenVC_05", "lenVC_1", "lenVC_2", "lenOJ_05", "lenOJ_1", "lenOJ_2", "pig_id
head(myToothGrowth3)
```

```
##   lenVC_05 lenVC_1 lenVC_2 lenOJ_05 lenOJ_1 lenOJ_2 pig_id
## 1      4.2    16.5    23.6     15.2    19.7    25.5      1
## 2     11.5    16.5    18.5     21.5    23.3    26.4      2
## 3      7.3    15.2    33.9     17.6    23.6    22.4      3
## 4      5.8    17.3    25.5      9.7    26.4    24.5      4
## 5      6.4    22.5    26.4     14.5    20.0    24.8       5
## 6     10.0    17.3    32.5     10.0    25.2    30.9       6
```

Statistical analysis

As seen before, we are dealing witg 6 measurements on 10 subjects. These measurements were taken in 2 different conditions, let'say: *Orange juice* and *Ascorbic acid*. So we must be very careful in comparing two measurements belonging to different conditions.

My idea is:

- to compare lengths derived from different doses given with the same delivery method, to see if bigger dose (as we expect) related to bigger length;
- to compare same doses given with different delivery methods, to see if different delivery method relates to different length (looking at the boxplots, i think it worth considering the dose 0.5 and 1, which seems the ones with bigger difference between lengths distributions and their averages).

But first of all I need to perform some check on normality of differences, although the size is very very small.

```
##Conclusions
```