Tooth Growth Analysis

Americo

19 luglio 2015

Packages

Main packages been used are `datasets`, `ggplot2`, `grid` and `gridExtra`.

Goals

We're going to analyze the ToothGrowth data in the R datasets package to perform some basic exploratory data analyses. Goals are to:

- provide a basic summary of the data;
- use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose (we'll only use the techniques from class, even if there's other approaches worth considering);
- state conclusions and the assumptions needed for conclusions.

Dataset

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`).

Exploratory data analysis

How delivery methods and doses are distributed? 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.

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. 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 that you can find at the appendix (*plot*1). I also want to take a look at the **distribution of lengths** (*plot*2 of appendix) but 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. So probably the best plot to look at is *plot*3 of appendix, where 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. This is the result:

```
##   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
```

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).

Verification of normality assumptions

But first of all I need to perform some check on normality of differences, although the size is very very small. I will create a dataframe of the differences between all the vectors I aim to compare.Naming convention is: `VC_1_05 = supp VC dose 1 - supp Vc dose 0.5` and `VC_OJ_05 = supp OJ - supp VC for dose 2`. At the appendix *plot*4 shows differences between lenghts derived by different doses of same delivery method, and they are almost all quite normal, and considering that T test in quite robust to the normality assumtpions, we can be happy wity this results.

Now we should alo plot the differences for same dose but different delivery method. At the appendix you find *plot*5 where you can see that really only dose 0.5 is far away from normality. We can go through the test now.

Hypothesis testing

Now let's perform the tests. As R provide a function to perform t test for paired samples. I separate `t test`results in the ones related to different doses given by same delivery methods, and the ones related to same doses of different delivery methods.

Different doses but same delivery method Naming convention is: **tVC_2_05 = t test on the difference between lengths derivde from dose 2 and dose 0.5 given with ascorbic acid**

Always:

- *null hypothesis*: mean of differences equal to 0
- *alternative hypothesis*: mean of differences greater than 0

Level of confidence: 0.95

Here are p-value results of all the six tests:

```
##     tVC_2_05      tVC_2_1     tVC_1_05     tOJ_2_05      tOJ_2_1
## 2.132128e-06 2.323975e-04 8.575825e-05 1.862051e-05 4.191956e-02
##     tOJ_1_05
## 1.217570e-03
```

All p-values are significantly lower than level of confidence alpha = 0.5. I would not put in doubt the results of the test in light of the non-perfect normality of differences, so for all the six tests ***null hypothesis is rejected***. This means that as dose increase the lengths of teeth, on average, increases (we are confident on these at 0.95, of course). You can take a look at results of the test at *plot*6 where the red line represents the borders of rejection region and the blue one the actaul t statistic for that test.

Same dose of different delivery method Naming convention is: **t_VC_OJ_2 = t test on the difference between lengths derived from dose 2 given with orange juice and given with ascorbic acid**

Always:

- *null hypothesis*: mean of differences equal to 0
- *alternative hypothesis*: mean of differences not equal to 0

Level of confidence: 0.95

Here are p-value results of all the three tests:

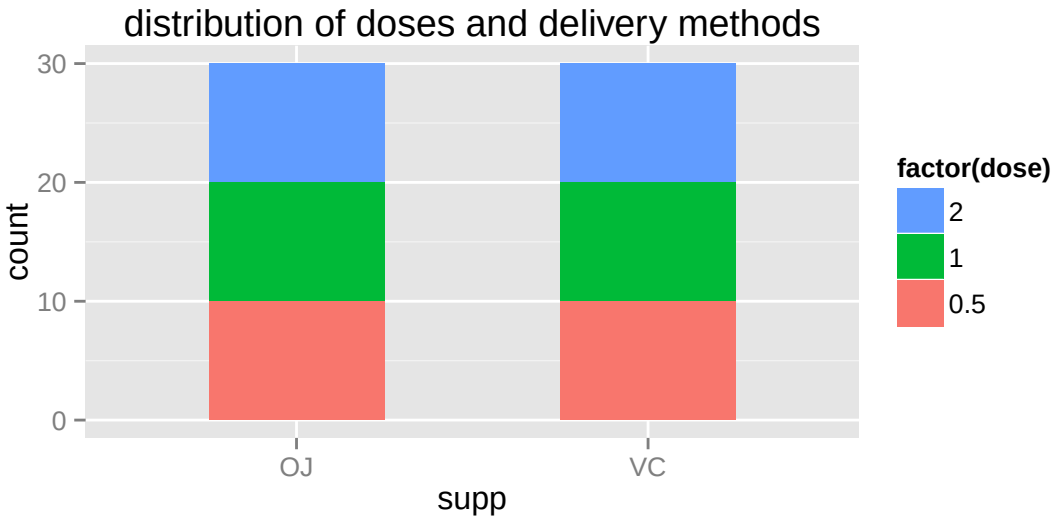
```
##   t_VC_OJ_2   t_VC_OJ_1  t_VC_OJ_05
## 0.966956704 0.008229248 0.015472048
```

For dose 2 t test has a huge p-value: of course **null hypothesis is not rejected** , and we can be confident in saying tha no difference (on average) in lenghts are due to different method if dose is two. This is not true for the other doses, where **null hypothesis is rejected**, although the non normality of differences for dose 0.5 make me doubtfoul about the results of the test.

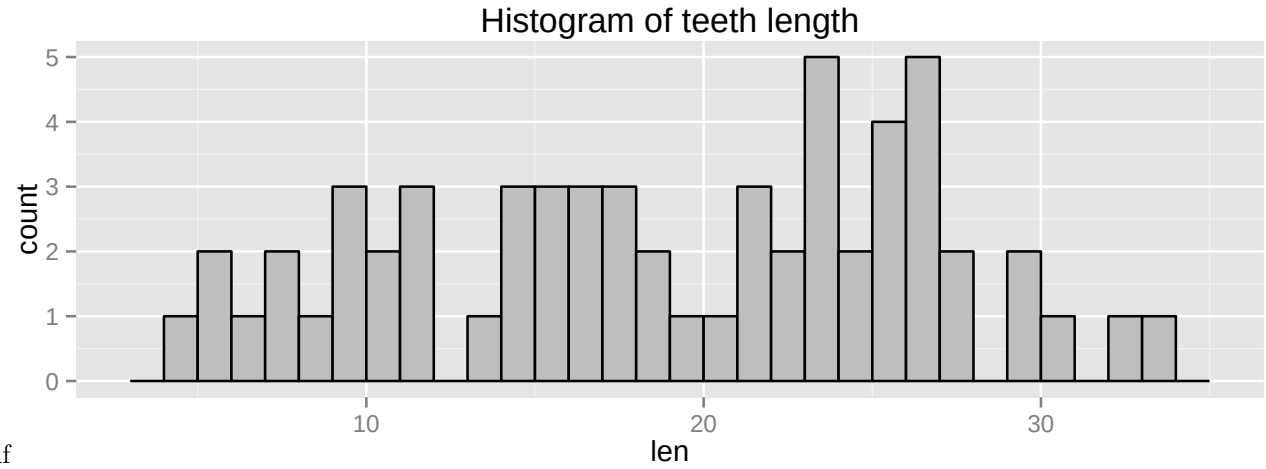
Conclusions

Appendix

Plot1

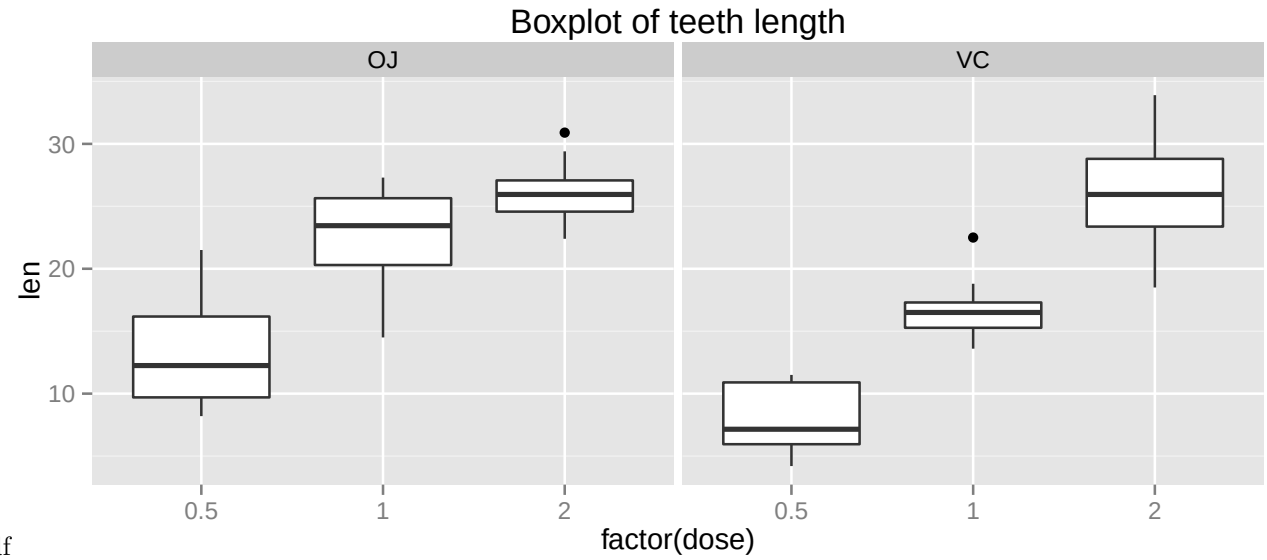


Plot2



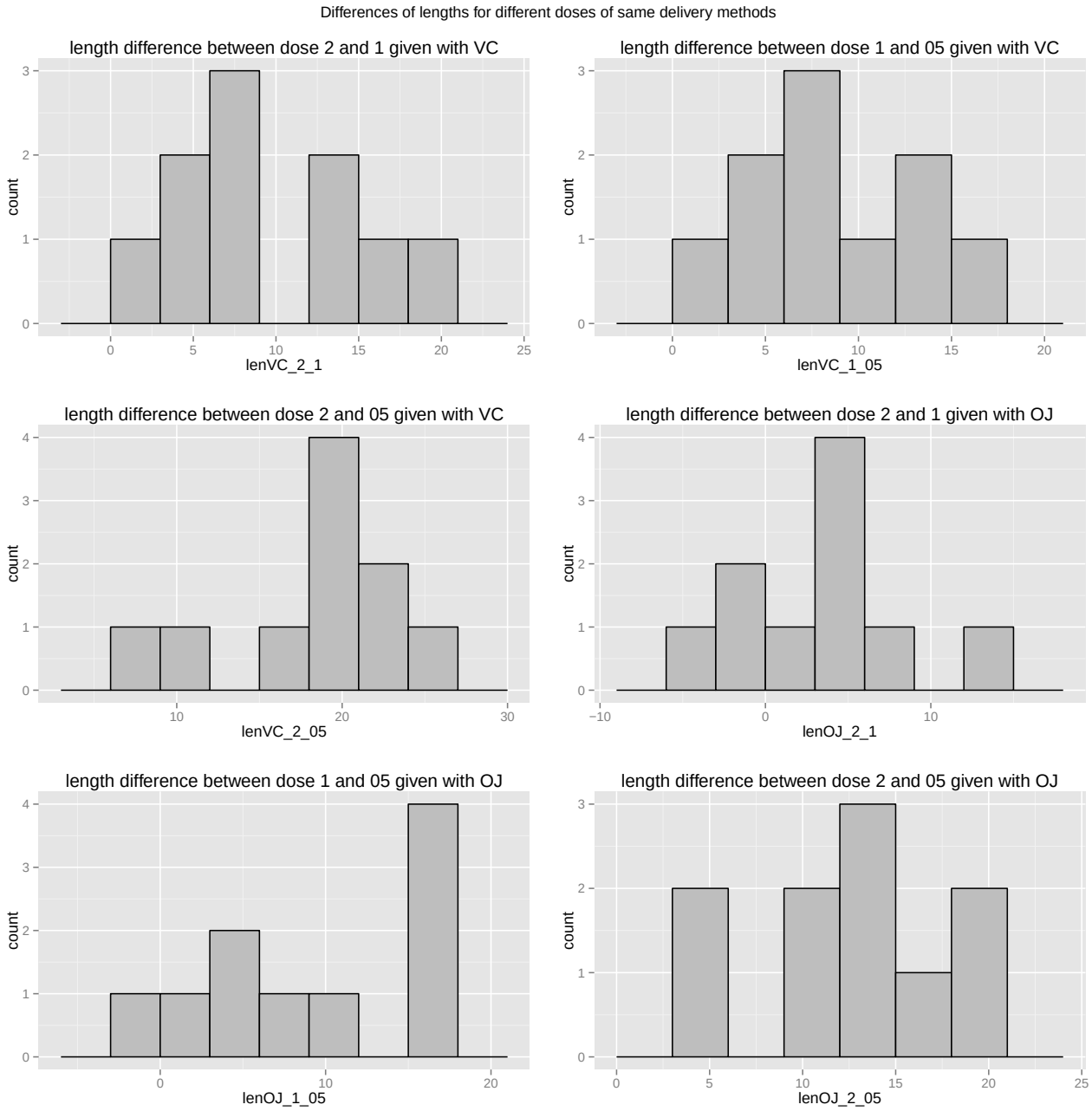
histogram-1.pdf

Plot3

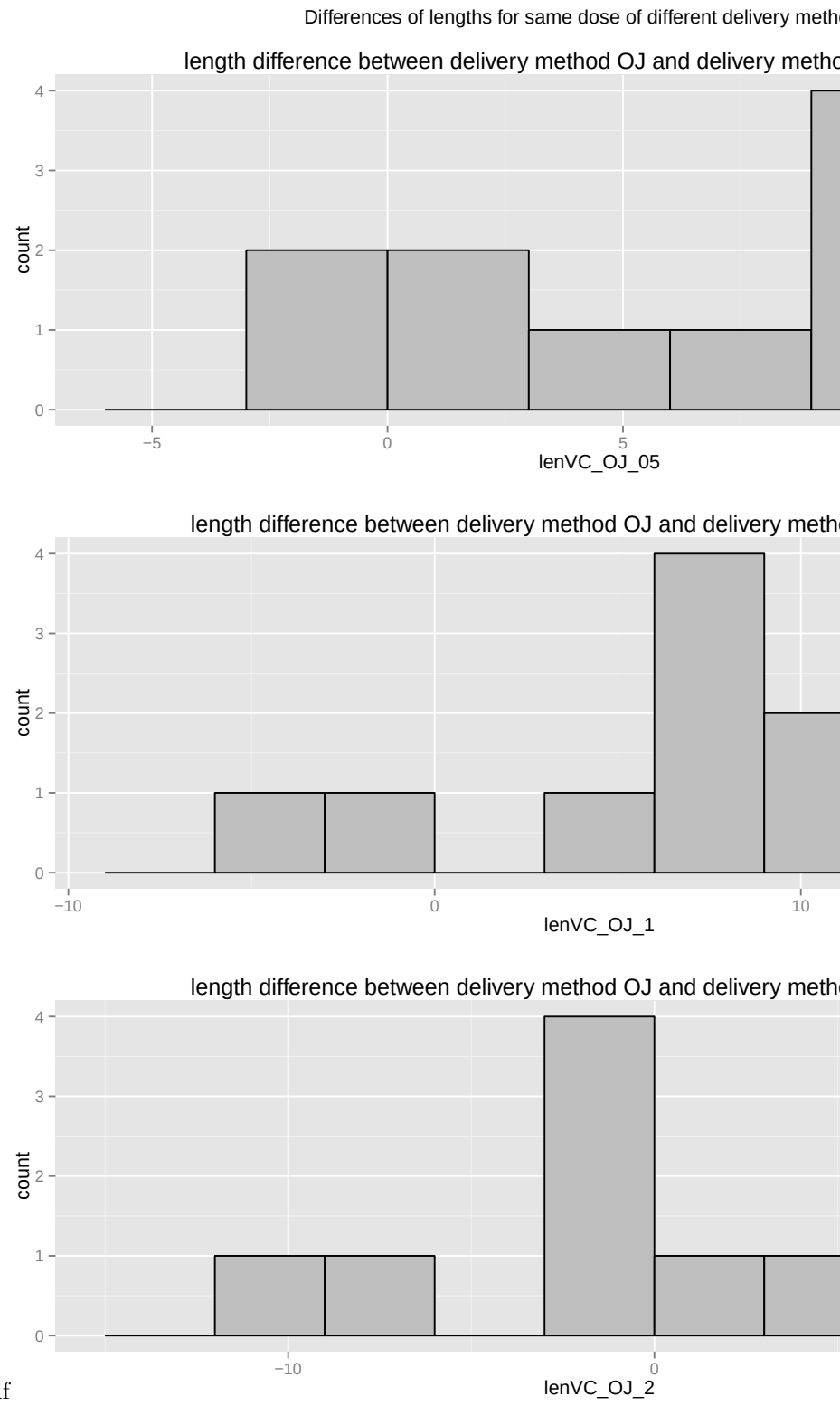


boxplots-1.pdf

Plot4


Differences of lengths for different doses of same delivery methods
length difference between dose 2 and 1 given with VC
lenVC_2_1
length difference between dose 1 and 05 given with VC
lenVC_1_05
length difference between dose 2 and 05 given with VC
lenVC_2_05
length difference between dose 2 and 1 given with OJ
lenOJ_2_1
length difference between dose 1 and 05 given with OJ
lenOJ_1_05
length difference between dose 2 and 05 given with OJ
lenOJ_2_05
count

Plot5



difference plots different delivery method-1.pdf

Plot6

vedi to do su github