Basic inferential data analysis on the tooth growth data in R

Sanja Stegerer

Thursday, November 19, 2015

A note in the beginning:
As stated in the introduction for the assignment, each report should not be longer than 3 pages + 3 pages of figures and code. I therefore went for the goal to not exceed 6 pages including all figures and code.

Load the ToothGrowth data and perform some basic exploratory data analyses

Load the data and suppress messages and warnings.

```r
data(ToothGrowth)
library(ggplot2)
suppressWarnings(suppressMessages(library(plyr)))
library(grid)
```

First step in the exploratory data analysis: get an overview of the dataframe

```r
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 ...
```

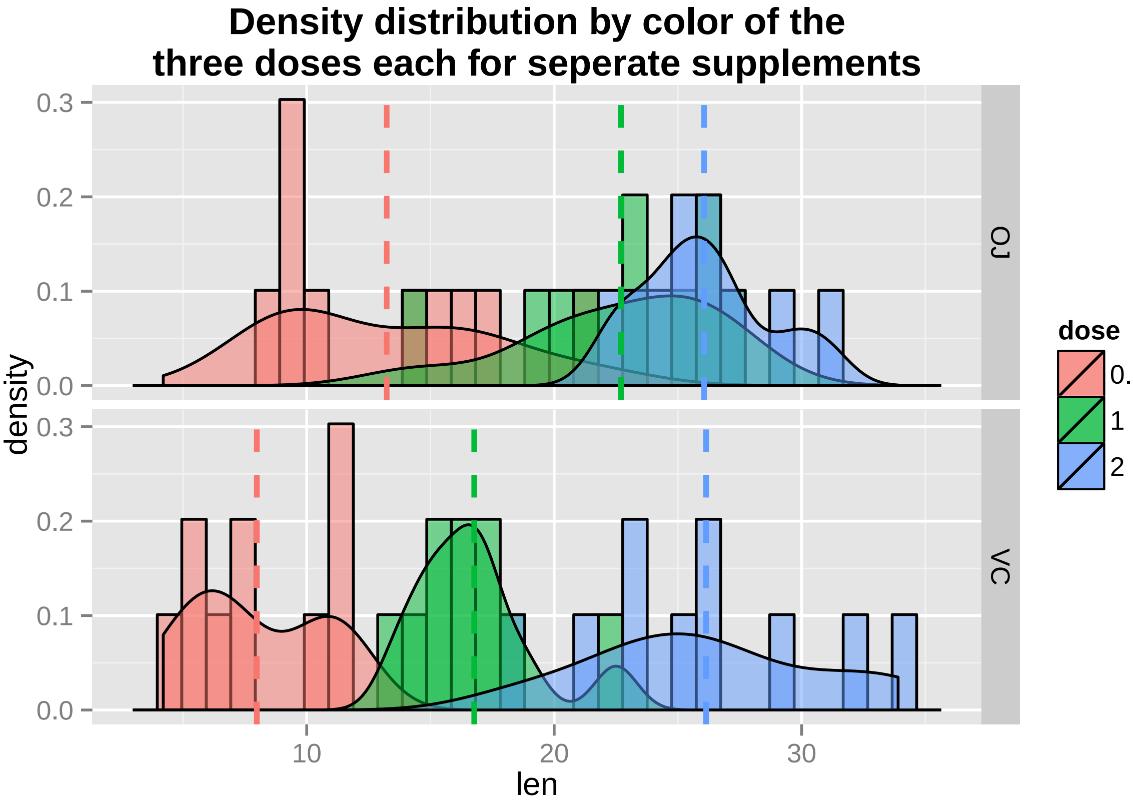
then (step 2) check how many examples from each supplementation and dose exists

```r
table(ToothGrowth$supp, ToothGrowth$dose, dnn=c("supp","dose"))
```

```
##     dose
## supp 0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

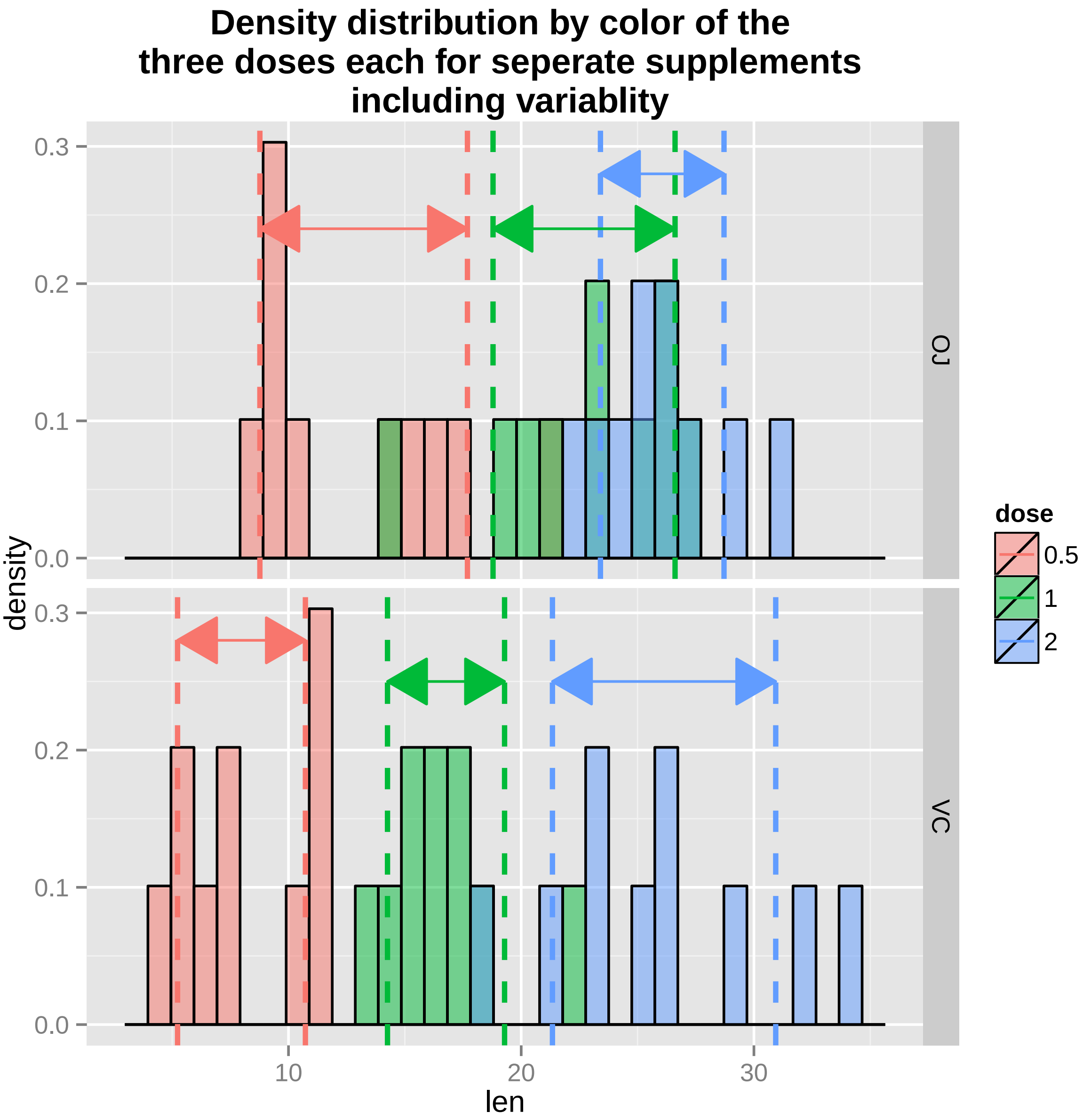
And as a third I step create a plot, that emphasises the differences in tooth growth per dose and supplementation

```r
ToothGrowth$dose <- factor(ToothGrowth$dose)
try_cat <- ddply(ToothGrowth,.(dose,supp),summarize,means=mean(len))
q = ggplot(ToothGrowth, aes(x=len,fill=dose)) +
    geom_histogram(alpha=0.5, position='identity', aes(y=..density..),
                   bin=(max(ToothGrowth$len)-min(ToothGrowth$len))/30, colour='black') +
    geom_density(alpha=0.5) + facet_grid(supp ~.) +
    geom_vline(data=try_cat, aes(xintercept=means, color = dose), linetype='dashed',
               size=1) +
    ggtitle('Density distribution by color of the
three doses each for seperate supplements') +
    theme(plot.title=element_text(face='bold'))

print(q)
```



where the dashed lines represent the mean values for each seperate distribution.
This plot already gives a great overview on the data and how it is structured. The following plot shows the range of $\mu \pm sd$ to give an overview over the variability of the data.

```
cat_sd <- ddply(ToothGrowth,.(dose,supp),summarize,sds=sd(len))
cat_sd <- merge(cat_sd, try_cat, by=c("dose","supp"))

q_sd = ggplot(data=ToothGrowth, aes(x=len,fill=dose)) +
    geom_histogram(alpha=0.5, position='identity', aes(y=..density..),color='black',
                   bin=(max(ToothGrowth$len)-min(ToothGrowth$len))/30) +
    facet_grid(supp ~.) +
    geom_vline(data=cat_sd, aes(xintercept=means+sds, color = dose),
               linetype='dashed', size=1) +
    geom_vline(data=cat_sd, aes(xintercept=means-sds, color = dose),
               linetype='dashed', size=1) +
    geom_segment(data=cat_sd, aes(x=means-sds, xend= means+sds,
                                  y=c(0.24,0.24,0.28,0.28,0.25,0.25),
                                  yend=c(0.24,0.24,0.28,0.28,0.25,0.25), color = dose),
                 arrow=arrow(ends='both',type='closed'),size=0.5)+
    ggtitle('Density distribution by color of the
three doses each for seperate supplements \n including variablity') +
    theme(plot.title=element_text(face='bold'))
print(q_sd)
```



These two plots allow me to develop 3 Nullhypotheses

1. 2 doses of the supplement OJ are as good for toothgrowth as 2 doses of the supplement VC.
2. 1 dose of the supplement OJ has a similar effect on toothsgrowths as 2 doses of the supplement VC.
3. 2 doses of OJ have the same effect or less on toothgrowth than all other supplementations and doses, except for 2 doses of the supplement VC.

Provide a basic summary of the data

In order to properly construct the plots for the exploratory data analysis, I already generated a summary of the data and stored it in the the variable cat_sd:

```
cat_sd
```

```
##   dose supp      sds means
## 1  0.5   OJ 4.459709 13.23
## 2  0.5   VC 2.746634  7.98
## 3    1   OJ 3.910953 22.70
## 4    1   VC 2.515309 16.77
## 5    2   OJ 2.655058 26.06
## 6    2   VC 4.797731 26.14
```

Use hypothesis testing to compare different tooth growths by supp and dose.

In the following I create a dataframe containing p values and the t confidence intervals for all combinations of supplementation and doses. The assumptions are

1. Each pair of dose and supplementation has a different variance than the other ones.
2. We use a two sided test, as two of the hypotheses test for equality for the mean value of two distributions.
3. We use a one sided test for the last hypothesis as our Null hypothesis is that all other dose-supplement pairs are as effective as the OC.2 pair or better.

```
suppressWarnings(library(data.table))
final_set <- data.frame(combination=character(),p.value=numeric(),
                        interval.lower=numeric(), interval.upper=numeric())
in_groups <- split(ToothGrowth, list(ToothGrowth$supp, ToothGrowth$dose))
names = names(in_groups)
for (i in 1:length(names)) {
    for (m in 1:length(names)){
        if (i < m) {
            temp <- t.test(in_groups[[i]][[1]], in_groups[[m]][[1]], var.equal = FALSE)
            temp2 <- c("",0,0,0)
            temp2[1] <- paste(names(in_groups)[i], names(in_groups)[m], sep="~")
            temp2[2] <- round(temp$p.value,7)
            temp2[3:4] <- round(temp$conf.int,7)
            final_set <- rbindlist(list(final_set,as.list(temp2)))
        }
    }
    }
final_set$contains0 <- final_set$interval.lower <= 0& final_set$interval.upper >= 0
print(final_set)
```

```
##       combination   p.value interval.lower interval.upper contains0
##  1: OJ.0.5~VC.0.5 0.0063586      1.7190573      8.7809427     FALSE
##  2:   OJ.0.5~OJ.1  8.78e-05    -13.4156344     -5.5243656     FALSE
##  3:   OJ.0.5~VC.1 0.0460103      -7.008109      -0.071891     FALSE
##  4:   OJ.0.5~OJ.2   1.3e-06    -16.3352406     -9.3247594     FALSE
##  5:   OJ.0.5~VC.2   7.2e-06    -17.2635219     -8.5564781     FALSE
##  6: VC.0.5~OJ.1         0    -17.9214925    -11.5185075     FALSE
##  7: VC.0.5~VC.1     7e-07     -11.265712      -6.314288     FALSE
##  8: VC.0.5~OJ.2         0    -20.6181832    -15.5418168     FALSE
##  9: VC.0.5~VC.2         0     -21.901512     -14.418488     FALSE
## 10:   OJ.1~VC.1 0.0010384      2.8021482      9.0578518     FALSE
## 11:   OJ.1~OJ.2 0.0391951     -6.5314425     -0.1885575     FALSE
## 12:   OJ.1~VC.2 0.0965261     -7.5643336      0.6843336      TRUE
## 13:   VC.1~OJ.2     2e-07    -11.7203326     -6.8596674     FALSE
## 14:   VC.1~VC.2  9.16e-05    -13.0542667     -5.6857333     FALSE
## 15:   OJ.2~VC.2 0.9638516     -3.7980705      3.6380705      TRUE
```

Discussion and conclusion

Going back to our Null-hypotheses:

1. 2 doses of the supplement OJ are as good for toothgrowth as 2 doses of the supplement VC.
2. 1 dose of the supplement OJ has a similar effect on toothsgrowths as 2 doses of the supplement VC.
3. 2 doses of OJ have the same effect or less on toothgrowth than all other supplementations and doses, except for 2 doses of the supplement VC.

and the corresponding H_A's are:

1. 2 doses of the supplement OJ(OC.2) have a different effect on toothgrowth as 2 doses of the supplement VC (VC.2).
2. 1 dose of the supplement OJ (OC.1) has a different effect on toothgrowth as 2 doses of VC (VC.2).
3. 2 doses of OJ (OC.2) have a greater effect on toothgrowth than all other supplementation, except for 2 doses of VC (VC.2).

Checking the first Null Hypothesis:

```
##    combination   p.value interval.lower interval.upper contains0
## 1:   OJ.2~VC.2 0.9638516     -3.7980705      3.6380705      TRUE
```

The upper subset makes it clear that we cannot reject the Null-Hypothethis and that the probability of the means being equal is 96%. The t-confidence interval backs this conclusion as it contains 0 and is actually centered around 0. In 95% of the time this intervall contains the true difference of the means.

Checking the second Null Hypothesis:

```
##    combination   p.value interval.lower interval.upper contains0
## 1:   OJ.1~VC.2 0.0965261     -7.5643336      0.6843336      TRUE
```

This result is way less clear, but we cannot reject the Null-Hypothesis either. 9% of data can be explained by the null hypothesis, which is not enough to reject it, but not a lot either. The confidence intervall contains 0, but is not centered around it, with a tendency towards negative values. This means, that the mean more often than not, is smaller than 0. Which can be interpreted as thoothgrowth is more often bigger under VC.2 than under OJ.1, but this difference is not big enough to be significant.

Checking the third Null Hypothesis: I have to redo the t-test as a one sided test:

```
final_set2 <- data.frame(combination=character(),p.value=numeric(),
                        interval.lower=numeric(), interval.upper=numeric())
    for (m in 1:length(names(in_groups))){
        if (names[m] != "OJ.2") {
            temp <- t.test(in_groups$OJ.2[[1]], in_groups[[m]][[1]], var.equal = FALSE,
                           alternative="greater")
            temp2[1] <- paste("OJ.2", names(in_groups)[m], sep="~")
            temp2[2] <- round(temp$p.value,7)
            temp2[3:4] <- round(temp$conf.int,7)
            final_set2 <- rbindlist(list(final_set2,as.list(temp2)))
        }
    }
print(final_set2)
```

```
##    combination   p.value interval.lower interval.upper
## 1: OJ.2~OJ.0.5     7e-07      9.9484504            Inf
## 2: OJ.2~VC.0.5         0      15.985071            Inf
## 3:   OJ.2~OJ.1 0.0195976      0.7486236            Inf
## 4:   OJ.2~VC.1     1e-07      7.2841492            Inf
## 5:   OJ.2~VC.2 0.5180742     -3.1334996            Inf
```

I set the alternative to 'greater' as our hypothesis states OJ.2 has a greater impact, than the other supplementations and doses, except for VC.2.

Going through the p-values we can safely reject all the Nullhypothesis (all doses and supplementations have the same or a greater effect than OJ.2) and conclude that OJ.2 has a greater effect on toothgrowth than the others (except VC.2). The significance level is always lower than 5%.