Brief analysis of the ToothGrowth dataset in R

Author: D.Aiazzi

Overview

For the second part of the course work of the Statistical Inference class, we're going to analyze the ToothGrowth data in the R datasets package. These are the project instructions:

1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.
3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

Exploratory analyses

The ToothGrowth dataset shows the growth of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid). The data frame is composed of 60 observations on 3 variables:
- **len**: Tooth length (numeric) - **supp**: Supplement type (VC for abosrbis acid, OJ for orange juice) (factor) - **dose**: Dose in milligrams (factor)

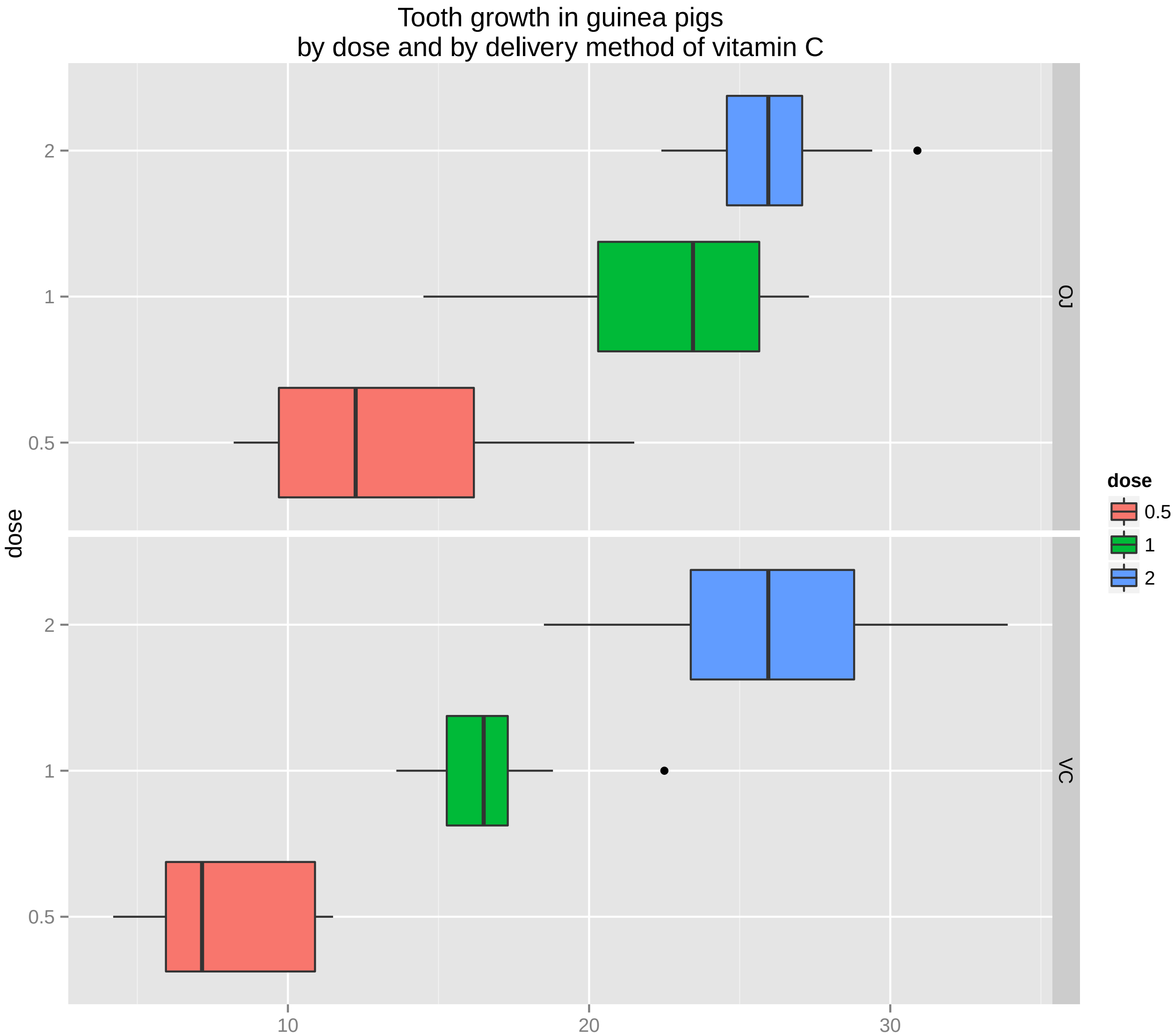
```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
## '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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

A quick summary shows the data range that goes from a minimum of 4.2 to a maximum of 33.9, with a mean 18.81 and median 19.25.

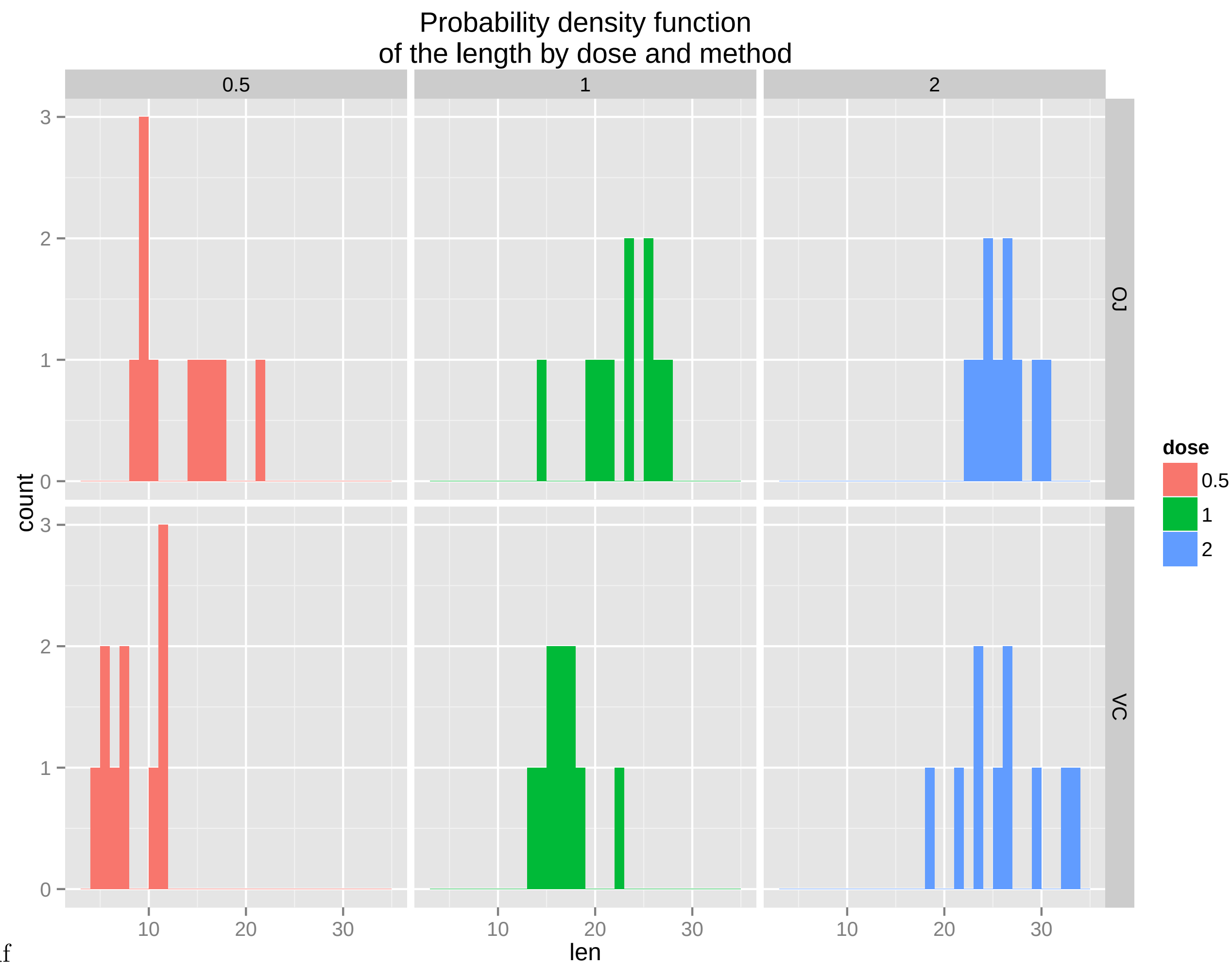
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   13.08   19.25   18.81   25.28   33.90
```

The following graphs show the measured length divided by delivery methods and by dose.



graph 1-1.pdf

From the first graphs, it seems that there is a positive correlation between the dose and the length. It also seems that OJ permforms better overall, although with higher variance for dose 0.5 and 1.
It seems iteresting to test the difference in response to the two different methods and the difference in response to different dosages.



graph 2-1.pdf

The second graph shows the density of the length mesurements for each pair dose/method. It is assumed that the data distribution is normal and given the small size of the sample, t-intervals will be used for the hypothesis tests.

Hypothesis test

We want to first understand if the difference in tooth growth is explained by higher levels of vitamin C delivered.
We set our null hypothesis that a lower supply gives the same length of a higher one and we test separately the two methods. We calculate the t-interval for each pair of dose, divided by group.

```
##
##  One Sample t-test
##
## data:  OJ1.0 - OJ0.5
## t = 4.1635, df = 9, p-value = 0.002435
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   4.324616 14.615384
## sample estimates:
## mean of x
##      9.47
```

```
##
##  One Sample t-test
##
## data:  OJ2.0 - OJ1.0
## t = 1.9435, df = 9, p-value = 0.08384
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.5509376  7.2709376
## sample estimates:
## mean of x
##      3.36
```

```
##
##  One Sample t-test
##
## data:  OJ2.0 - OJ0.5
## t = 7.4919, df = 9, p-value = 3.724e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   8.956042 16.703958
## sample estimates:
## mean of x
##     12.83
```

```
##
##  One Sample t-test
##
## data:  VC1.0 - VC0.5
## t = 6.1364, df = 9, p-value = 0.0001715
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   5.549601 12.030399
## sample estimates:
## mean of x
##      8.79
```

```
##
##  One Sample t-test
##
## data:  VC2.0 - VC1.0
## t = 5.346, df = 9, p-value = 0.0004648
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   5.405082 13.334918
## sample estimates:
## mean of x
##      9.37
```

```
##
##  One Sample t-test
##
## data:  VC2.0 - VC0.5
## t = 9.7912, df = 9, p-value = 4.264e-06
```

```
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  13.9643 22.3557
## sample estimates:
## mean of x
##     18.16
```

The null hypothesis that the difference is zero is rejected for all of the cases except for OJ2.0 - OJ1.0, which means that over a dose of 1 the difference in tooth growth when the dose is increased is not statistically significant.

We then test the hypothesis that there is a statistical difference between the two supply methods, taking as null hypothesis that OJ performs the same as VC.

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

The t test over the two samples fails to reject the null hypothesis. The two methods overall performs the same.

Let's see in detail by different doses.

```
##
##  One Sample t-test
##
## data:  OJ0.5 - VC0.5
## t = 2.9791, df = 9, p-value = 0.01547
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.263458 9.236542
## sample estimates:
## mean of x
##      5.25
```

```
##
##  One Sample t-test
##
## data:  OJ1.0 - VC1.0
## t = 3.3721, df = 9, p-value = 0.008229
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.951911 9.908089
## sample estimates:
## mean of x
##      5.93
```

```
##
##  One Sample t-test
##
## data:  OJ2.0 - VC2.0
## t = -0.0426, df = 9, p-value = 0.967
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -4.328976  4.168976
## sample estimates:
## mean of x
##     -0.08
```

OJ performs better at smaller doses (0.5 and 1) but for higher doses (2), the difference between the two methods is not statistically significant.

Conclusions

The hypothesis tests have been conducted under the assumptions that the data distribution is normal and that the variance is not constant amongst the groups.

Based on the assumptions that we want to maximise the tooth growth and minimise the dose, the method OJ with dose 1 is the best performer.

Code

```r
## Overlook at the data ##
data(ToothGrowth)
df <- ToothGrowth
rm(ToothGrowth)
df$dose <- as.factor(df$dose)
head(df)
str(df)

## Summary of the data ##
summary(df$len)

## Explanatory graph 1 ##
library(ggplot2)
ggplot(df, aes(dose, len)) +
    geom_boxplot(aes(fill = dose)) +
    facet_grid(supp ~ .) +
    ggtitle("Tooth growth in guinea pigs\nby dose and by delivery method of vitamin C") +
    coord_flip()

## Explanatory graph 2 ##
ggplot(df, aes(x = len)) +
    geom_histogram(aes(fill = dose), binwidth = 1) +
    ggtitle("Probability density function\nof the length by dose and method") +
    facet_grid(supp ~ dose)
```

```r
## Set data and variables for the test ##
OJ0.5 <- df[df$supp == "OJ" & df$dose == 0.5, 1]
OJ1.0 <- df[df$supp == "OJ" & df$dose == 1, 1]
OJ2.0 <- df[df$supp == "OJ" & df$dose == 2, 1]
VC0.5 <- df[df$supp == "VC" & df$dose == 0.5, 1]
VC1.0 <- df[df$supp == "VC" & df$dose == 1, 1]
VC2.0 <- df[df$supp == "VC" & df$dose == 2, 1]

var.equal <- FALSE
conf.level <- .95

## Do bigger doses correspond to higher tooth growth ##
# Method OJ #
t.test(OJ1.0-OJ0.5, paired = FALSE, var.equal = var.equal)
t.test(OJ2.0-OJ1.0, paired = FALSE, var.equal = var.equal)
t.test(OJ2.0-OJ0.5, paired = FALSE, var.equal = var.equal)
# Method VC #
t.test(VC1.0-VC0.5, paired = FALSE, var.equal = var.equal)
t.test(VC2.0-VC1.0, paired = FALSE, var.equal = var.equal)
t.test(VC2.0-VC0.5, paired = FALSE, var.equal = var.equal)

## Do OJ and VC perform differently overall? ##
t.test(len ~ supp, paired = FALSE, var.equal = var.equal, data = df, conf.level = conf.level)

## Do OJ and VC perform differently by dose? ##
t.test(OJ0.5-VC0.5, paired = FALSE, var.equal = var.equal)
t.test(OJ1.0-VC1.0, paired = FALSE, var.equal = var.equal)
t.test(OJ2.0-VC2.0, paired = FALSE, var.equal = var.equal)
```