Analysis of the ToothGrowth data in R

RDSN
20 October 2015

Description

The response is the length 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).

1. Load the ToothGrowth data and perform some basic exploratory data analyses

Loading the ToothGrowth data

```
data("ToothGrowth")
```

Let's have a look at the first rows of this data set

```
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
```

And its structure

```
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 ...
```

So we've got here 60 observations and 3 variables.

```
table(ToothGrowth$dose)

## 
## 0.5   1   2 
##  20  20  20
```

So we see here that the numerical variable "dose" has only 3 values : 0.5, 1, 2.

By looking at the entire dataset, we can observe that indeed, this dataset is about 10 subjects. Every 10 rows, the numerical value of the dose change. Every 30 rows, the value of the supp variable changes.

To have an idea of the data, we can plot the evolution of the average length depending on the supplement and the dose of vitamin C.

```
av <- aggregate(len ~ supp + dose, data = ToothGrowth, mean)
library(ggplot2)

g <- ggplot(data = av, aes(x = dose, y = len, col = supp)) + geom_line()
g
```

See Appendix – Figure 1 for the plot

We can see here that the average length is larger with the method "OJ" (Orange Juice), than with the ascorbic acid. We can also observe that the more important the dose of Vitamin C, the more important the average length.

2. Provide a basic summary of the data

See Appendix – Figure 2 for the detail

3. 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)

3.1 Supp

First, we are going to compare the 2 levels of supp.

See Appendix – Figure 3 for the detail

Here, we have a confidence interval from -7.57 to 0.17, so 0 is inside the confidence interval. p-value = 0.06 What it means is that we don't have enough evidence to reject the null H0 hypothesis, the null hypothesis being that the mean of the 2 groups are equal. And so we can't say that the difference in the mean of the 2 groups is significant.

3.2 Dose

Let's do the same with the dose values. As we have 3 dose values, we must perform here 3 different t.test, one for each pair.

See Appendix – Figure 4 for the detail

Note that we have considered here that the values are paired, because each group is represented by the same pigs (10 ginea pigs). If we had considered it not paired, the values of the intervals would not have been much different :

See Appendix – Figure 5 for the detail

Furthermore, we see here that for each t.test, 0 is not included in the confidence interval, and each p-value is very small. This means that we can reject the null hypothesis (H0) and so we can say that the difference in the means of the different groups is significant.

4. State your conclusions and the assumptions needed for your conclusions.

Assumptions:

- To perform those t.tests,
  - For **Supp** : we have assumed that the groups were unpaired. Indeed, The subjects tested are the same for each value of dose, but not for each value of supp, the sample of size 30 for each supp being composed of 3 times the same 10 guinea pigs.
  - For **Dose** : we have assumed that the groups were paired. Indeed, The subjects tested are the same for each value of dose. As we have shown above, the t.test performed with unpaired values presents no significant differences in terms of confidence intervals.
- Then, we have assumed that the 2 groups, each time, don't have the same variance. We don't have any evidence that the variance may be the same. That's why we have not specified in the t.test the value "var.equal = TRUE", and so this value is set to FALSE by default.
- In order to use the t interval, we have also assumed that the data are iid normal.

4.1 Supp

Conclusions:

We have a confidence interval from -7.57 to 0.17, so 0 is inside the confidence interval. p-value = 0.06 What it means is that we don't have enough evidence to reject the null H0 hypothesis, the null hypothesis being that the means of the 2 groups are equal. And so we can't say that the difference in the means of the 2 groups is significant.

4.2 Dose

Conclusions:

Here are the confidence intervals shown by the t.tests :

- groups 1 & 0.5 : from 6.39 to 11.87 (p-value = 1.225e-06)
- groups 2 & 1 : from 3.47 to 9.26 (p-value = 1.93e-04)
- groups 2 & 0.5 : from 12.62 to 18.37 (p-value = 7.19e-10)

We see that for each t.test, 0 is not included in the confidence interval, and each p-value is very small. This means that we can reject the null hypothesis (H0) and so we can say that the difference in the means of the different groups is significant. Furthermore, as we have got a positive interval each time, we can be confident that the mean of group dose = 1 may be superior to the mean of group dose = 0.5 and that the mean of group dose = 2 may be superior to the mean of the group dose = 1.

Appendix

Figure 1

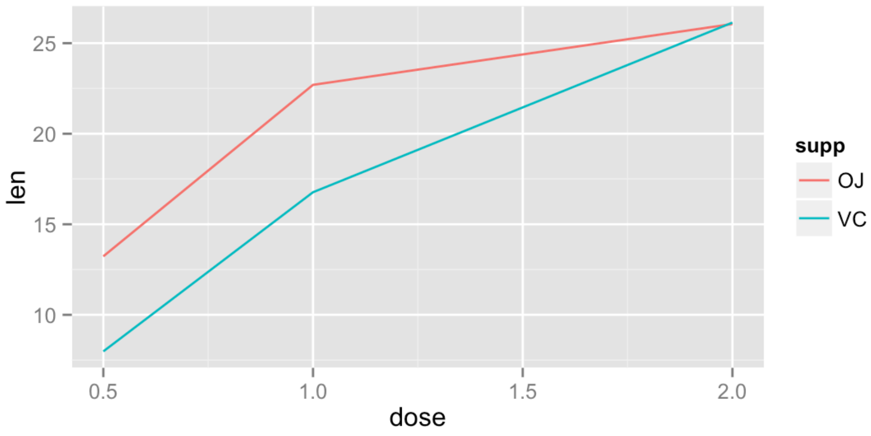

Figure 2

```
summary(ToothGrowth)

##       len        supp        dose      
##  Min.   : 4.20   OJ:30   Min.   :0.500  
##  1st Qu.:13.07   VC:30   1st Qu.:0.500  
##  Median :19.25           Median :1.000  
##  Mean   :18.81           Mean   :1.167  
##  3rd Qu.:25.27           3rd Qu.:2.000  
##  Max.   :33.90           Max.   :2.000  
```

Figure 3

```
g1 <- ToothGrowth[ToothGrowth$supp == "OJ", "len"]
g2 <- ToothGrowth[ToothGrowth$supp == "VC", "len"]

t.test(g2, g1)

## 
##  Welch Two Sample t-test
## 
## data:  g2 and g1
## 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:
##  -7.5710156  0.1710156
## sample estimates:
## mean of x mean of y 
##  16.96333  20.66333 
```

Figure 4

```
gd1 <- ToothGrowth[ToothGrowth$dose == 0.5, "len"]
gd2 <- ToothGrowth[ToothGrowth$dose == 1, "len"]
gd3 <- ToothGrowth[ToothGrowth$dose == 2, "len"]

t.test(gd2, gd1, paired = TRUE)

## 
##  Paired t-test
## 
## data:  gd2 and gd1
## t = 6.9669, df = 19, p-value = 1.225e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.387121 11.872879
## sample estimates:
## mean of the differences 
##                    9.13

t.test(gd3, gd2, paired = TRUE)

## 
##  Paired t-test
## 
## data:  gd3 and gd2
## t = 4.6046, df = 19, p-value = 0.0001934
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.471814 9.258186
## sample estimates:
## mean of the differences 
##                   6.365

t.test(gd3, gd1, paired = TRUE)

## 
##  Paired t-test
## 
## data:  gd3 and gd1
## t = 11.291, df = 19, p-value = 7.19e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  12.6228 18.3672
## sample estimates:
## mean of the differences 
##                  15.495
```

Figure 5

```
rbind(t.test(gd2, gd1, paired = TRUE)$conf.int, t.test(gd2, gd1, paired = FALSE)$conf.int)

##          [,1]     [,2]
## [1,] 6.387121 11.87288
## [2,] 6.276219 11.98378

rbind(t.test(gd3, gd2, paired = TRUE)$conf.int,t.test(gd3, gd2, paired = FALSE)$conf.int)

##          [,1]     [,2]
## [1,] 3.471814 9.258186
## [2,] 3.733519 8.996481

rbind(t.test(gd3, gd1, paired = TRUE)$conf.int, t.test(gd3, gd1, paired = FALSE)$conf.int)

##          [,1]     [,2]
## [1,] 12.62280 18.36720
## [2,] 12.83383 18.15617
```