Statistical inference project - PART 2

Stephanie

February 2016

Contents

Overview

In this report, we will perform a basic inferential analysis of the *ToothGrowth dataset* from the R datasets package.

As explained in the help file of this dataset, *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).*

Load the ToothGrowth data and perform some basic exploratory data analyses

Loading the ToothGrowth data

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

Exploratory data analysis

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
```

Let's have a look at 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 ...
```

The dataset is made of 60 observations and 3 variables.

```
table(ToothGrowth$dose)
```

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

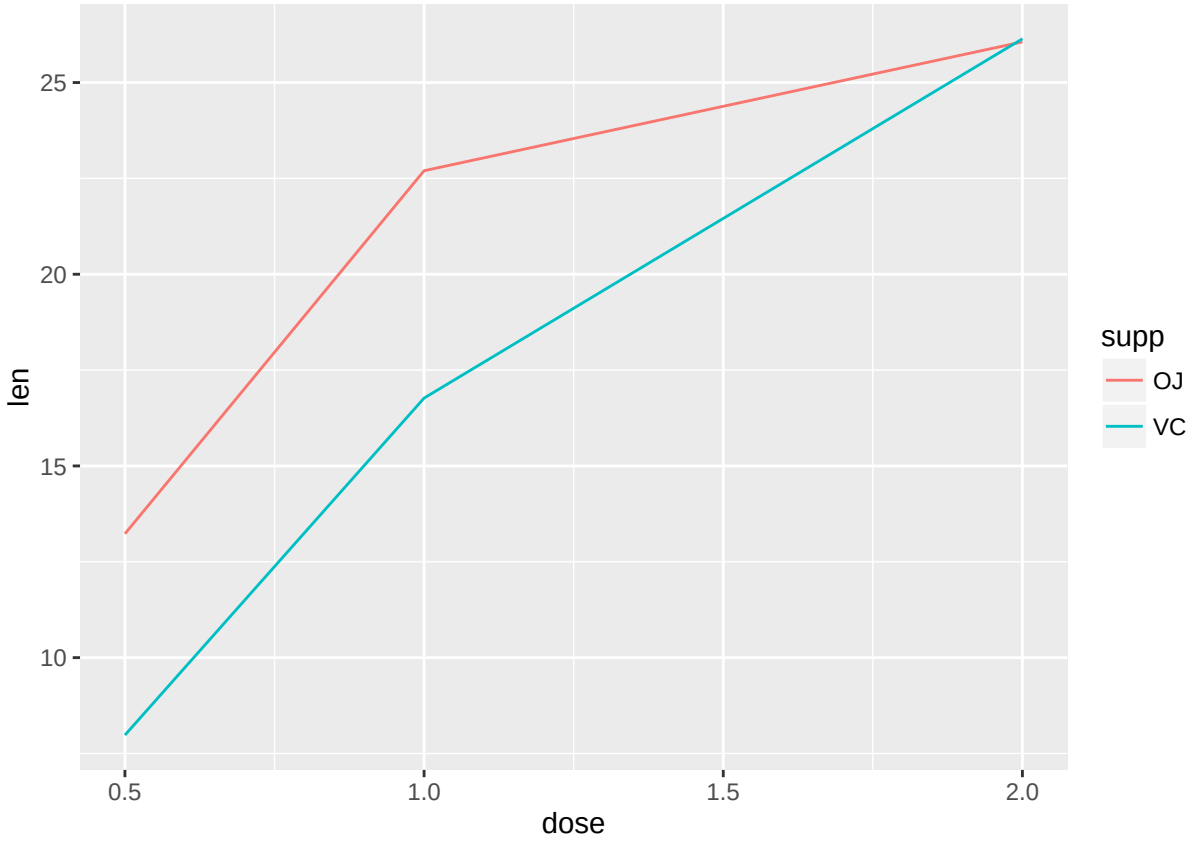
The numerical variable "dose" (= dose levels of Vitamin C) has only 3 values : 0.5, 1, 2 (in mg).

To explore the data, we can plot the evolution of the average length teeth depending on the method and the dose of vitamin C.

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

library(ggplot2)

plot1 <- ggplot(data = average, aes(x = dose, y = len, col = supp)) + geom_line()
plot1
```



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

Basic summary of the data

```
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
```

Inferential analysis to compare tooth growth by supp and dose

Comparing supp

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

```r
oj <- ToothGrowth[ToothGrowth$supp == "OJ", "len"]
vc <- ToothGrowth[ToothGrowth$supp == "VC", "len"]

t.test(oj, vc)
```

```
##
##  Welch Two Sample t-test
##
## data:  oj and vc
## 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 of x mean of y
##  20.66333  16.96333
```

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. We can't say that the difference in the mean of the 2 groups is significant.

Comparing 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.

```r
dose1 <- ToothGrowth[ToothGrowth$dose == 0.5, "len"]
dose2 <- ToothGrowth[ToothGrowth$dose == 1, "len"]
dose3 <- ToothGrowth[ToothGrowth$dose == 2, "len"]

t.test(dose2, dose1, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  dose2 and dose1
## 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
```

```r
t.test(dose3, dose2, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  dose3 and dose2
```

```
## 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(dose3, dose1, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  dose3 and dose1
## 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
```

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

```
rbind(t.test(dose2, dose1, paired = TRUE)$conf.int, t.test(dose2, dose1, paired = FALSE)$conf.int)
```

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

```
rbind(t.test(dose3, dose2, paired = TRUE)$conf.int,t.test(dose3, dose2, paired = FALSE)$conf.int)
```

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

```
rbind(t.test(dose3, dose1, paired = TRUE)$conf.int, t.test(dose3, dose1, paired = FALSE)$conf.int)
```

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

We can 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). So we can say that the difference in the means of the different groups is significant.

Assumptions and conclusions

Supp

Assumptions

- To perform those t.tests, we have assumed that the groups were unpaired. 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.
- We also 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.

Conclusions We have a confidence interval from -7.57 to 0.17, so 0 is inside the confidence interval. p-value = 0.06. 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. We can't say that the difference in the mean of the 2 groups is significant.

Dose

Assumptions

- To perform those t.tests, we have assumed that the groups were paired. 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.
- We have also 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.

Conclusions

Below 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. We can reject the null hypothesis (H0) and say that the difference in the means of the different groups is significant. As we have got a positive interval each time, we can be confident that the mean of the group dose = 1 may be superior to the mean of the group dose = 0.5 and that the mean of the group dose = 2 may be superior to the mean of the the group dose = 1.