Analyzing ToothGrowth data

Sergio Martínez Tornell

22 November 2015

Summary

In this report we analyze the ToothGrowth dataset from package *datasets* in *R* we found that the outcome len is strongly related with dose and supp. Moreover, for certain *doses*, there is a significant difference for different *supp*.

Exploratory analysis

First, get a summary of the data.

```
summary(data)
```

```
##       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
```

```
str(data)
```

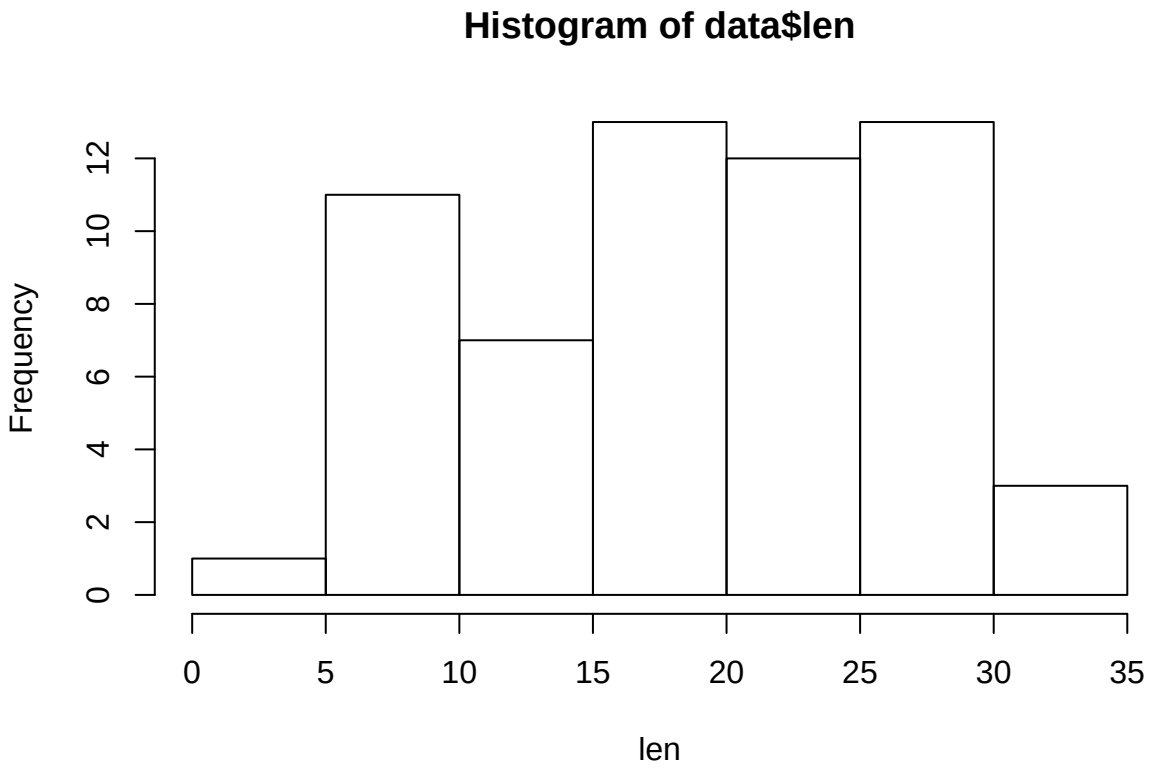
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

The data has 3 columns:

- **len:** Which is the outcome value.
- **supp:** Which seems to be a factor of the experiment.
- **dose:** Which is a numerical factor which takes only three values: 0.5, 1, 2.

Summarize and plot a histogram of the **len**.

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

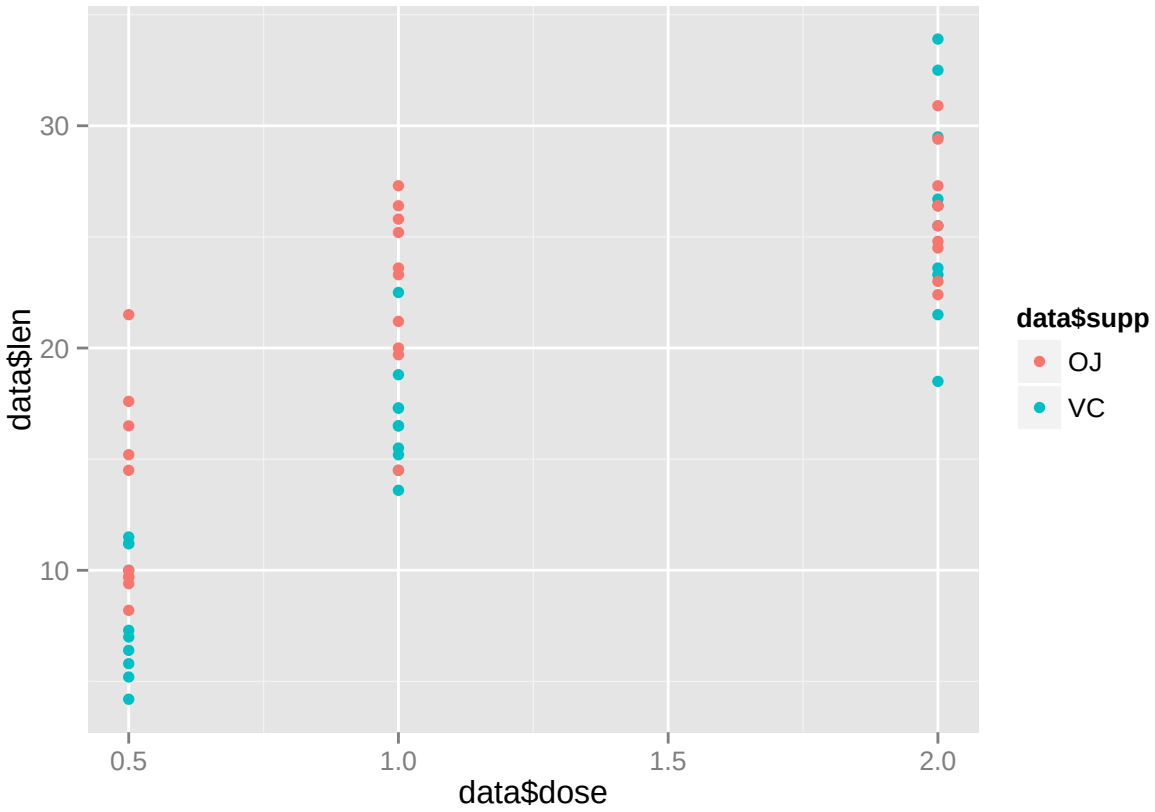


Check whether supp and dose are related

	0.5	1	2
OJ	10	10	10
VC	10	10	10

It seems that both were factors of the experiment.

Do some qplot:



It seems at *doses* of 0.5 and 1.0 each *supp* has a slightly different effect on the outcome. However at *dose* 2.0 it seems there is no difference between *supps*. Moreover, it seems to be a correlation between *dose* and *len* independently of the *supp*.

Analysis

Objective of the analysis

We want to validate our hypotheses stated at the end of the previous section: 1. There is a difference in the outcome *len* between different *doses* 2. There is a difference in the outcome *len* at *doses* 0.5 and 1.0 between the different *supps*. 3. There is **NO** difference in the outcome *len* at *dose* 2.0 between the different *supps*.

Validate the impact of *dose* in *len*

To simplify, we test if there is a significant difference between 0.5 and 2.0 *doses*. Since we can not establish a relationship between the measurements, we have performed a *t.test* for not paired populations assuming equal variance for both populations.

```
results <- t.test(data$len[data$dose == 0.5], data$len[data$dose == 2.0],
                  var.equal = TRUE, paired = FALSE)
results
```

```
##
##  Two Sample t-test
##
## data:  data$len[data$dose == 0.5] and data$len[data$dose == 2]
```

```
## t = -11.799, df = 38, p-value = 2.838e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15352 -12.83648
## sample estimates:
## mean of x mean of y
##    10.605    26.100
```

We have found that there is a mean difference of -15.495 with a p value of $2.8375532 \times 10^{-14}$. A very small p value suggest that this relationship is strong and it is rarely related to chance alone. We can confirm our hypotheses: **A bigger dose produce a bigger *len***.

Validate the impact of *supp* in *len* at different *doses*

```
results <- t.test(data$len[data$dose == 0.5 & data$supp == "VC" ],
                  data$len[data$dose == 0.5 & data$supp == "OJ" ],
                  var.equal = TRUE, paired = FALSE)
results
```

Testing the impact at *dose* 0.5

```
##
##  Two Sample t-test
##
## data:  data$len[data$dose == 0.5 & data$supp == "VC"] and data$len[data$dose == 0.5 & data$supp == "
## t = -3.1697, df = 18, p-value = 0.005304
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.729738 -1.770262
## sample estimates:
## mean of x mean of y
##      7.98     13.23
```

We have found that there is a mean difference $(VJ - OJ)$ of -5.25 with a p value of 0.0053037. A very small p value suggest that this relationship is strong and it is rarely related to chance alone. We can confirm our hypotheses: **There is different between *supps* at** 0.5 ***dose***.

```
results <- t.test(data$len[data$dose == 1 & data$supp == "VC" ],
                  data$len[data$dose == 1 & data$supp == "OJ" ],
                  var.equal = TRUE, paired = FALSE)
results
```

Testing the impact at *dose* 1.0

```
##
##  Two Sample t-test
##
```

```
## data:  data$len[data$dose == 1 & data$supp == "VC"] and data$len[data$dose == 1 & data$supp == "OJ"]
## t = -4.0328, df = 18, p-value = 0.0007807
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.019308 -2.840692
## sample estimates:
## mean of x mean of y
##     16.77     22.70
```

We have found that there is a mean difference $(VJ - OJ)$ of -5.93 with a p value of 7.8072617×10^{-4}. A very small p value suggest that this relationship is strong and it is rarely related to chance alone. We can confirm our hypotheses: **There is different between *supps* at** 1 ***dose***.

```
results <- t.test(data$len[data$dose == 2 & data$supp == "VC" ],
                  data$len[data$dose == 2 & data$supp == "OJ" ],
                  var.equal = TRUE, paired = FALSE)
results
```

Testing the impact at *dose* 2.0

```
##
##  Two Sample t-test
##
## data:  data$len[data$dose == 2 & data$supp == "VC"] and data$len[data$dose == 2 & data$supp == "OJ"]
## t = 0.046136, df = 18, p-value = 0.9637
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.562999  3.722999
## sample estimates:
## mean of x mean of y
##     26.14     26.06
```

We have found that there is a mean difference $(VJ - OJ)$ of 0.08 with a p value of 0.9637098. A big p value suggest that this relationship is weak and it is probably related to chance alone. We can confirm our stated hypotheses: **There is no different between *supps* at** 2.0 ***dose***.

Conclusion

In our test we confirmed that:

1. There is a difference in the outcome *len* between different *doses*
2. There is a difference in the outcome *len* at *doses* 0.5 and 1.0 between the different *supps*.
3. There is **NO** difference in the outcome *len* at *dose* 2.0 between the different *supps*.

The discovered relationships suggest that a better analysis considereng co-related variables may be desirable.