Analysis of the ToothGrowth dataset

Luciano Lattes

June, 2015

For the second part of the Course Project I'm going to analyze the ToothGrowth data in the R datasets package.

Exploratory data analysis

This dataset describes the Effect of Vitamin C on Tooth Growth in Guinea Pigs. It consists in 60 observations of 3 variables. The vitamin C was delivered by two methods: Orange juice (OJ) or ascorbic acid (VC) in different dose levels.

Let's take a look at the data to get familiarized with it:

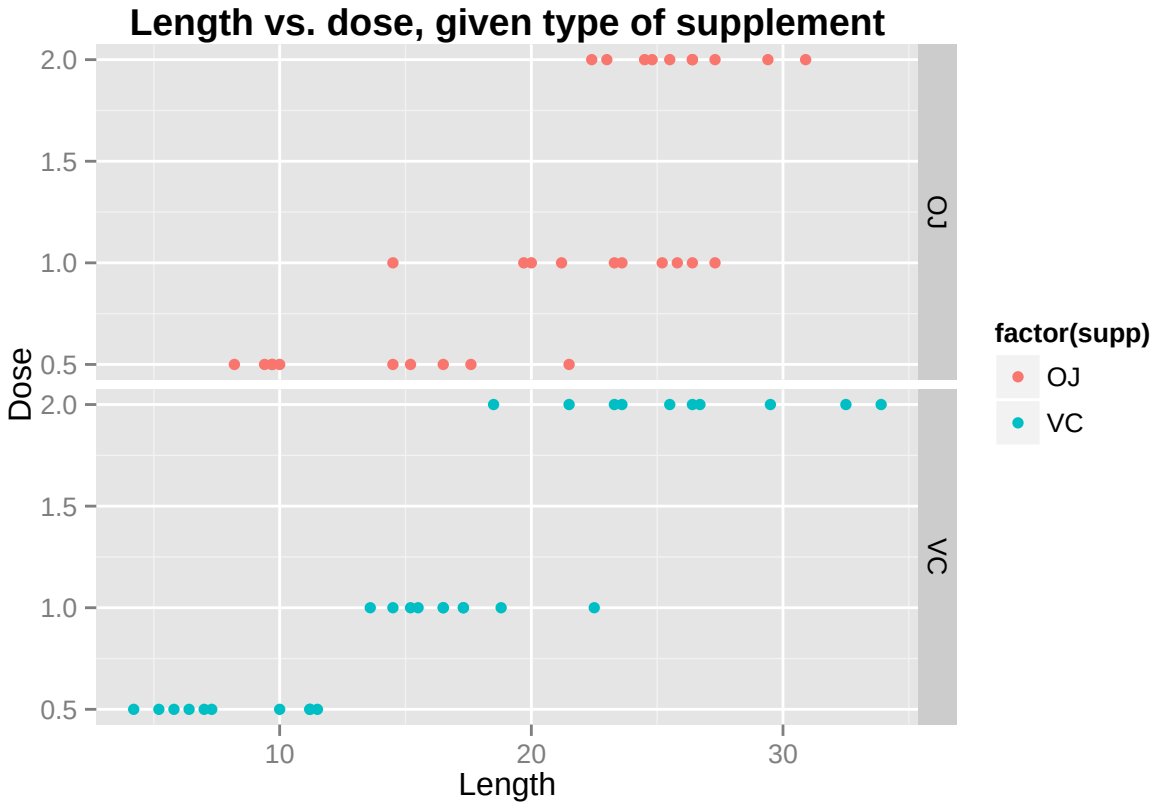
```
summary(ToothGrowth)
```

```
##       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
```

```
# Display a random sample of rows in the dataset.
ToothGrowth[sample(nrow(ToothGrowth), 10), ]
```

```
##     len supp dose
## 44 26.4   OJ  1.0
## 18 14.5   VC  1.0
## 25 26.4   VC  2.0
## 52 26.4   OJ  2.0
## 49 14.5   OJ  1.0
## 17 13.6   VC  1.0
## 50 27.3   OJ  1.0
## 59 29.4   OJ  2.0
## 37  8.2   OJ  0.5
## 6  10.0   VC  0.5
```

To have a better sense of what the data looks like, let's make a plot. I'll take advantage of the `supp` factor to make a faceted plot and try to see the relation between length and dose for each delivery method:



Seems that the **orange juice** method has a bigger impact in the length of the tooth and that the values of the length are more concentrated around the mean in the case of **ascorbic acid** except for the 2 mg. dose, where this facts seem to be exactly the opposite. Let's try to confirm this numerically:

- **Dose of 0.5 mg.:**

```
mean(ToothGrowth[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 0.5, ]$len)
```

```
## [1] 13.23
```

```
mean(ToothGrowth[ToothGrowth$supp == "VC" & ToothGrowth$dose == 0.5, ]$len)
```

```
## [1] 7.98
```

- **Dose of 1.0 mg.:**

```
mean(ToothGrowth[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 1.0, ]$len)
```

```
## [1] 22.7
```

```
mean(ToothGrowth[ToothGrowth$supp == "VC" & ToothGrowth$dose == 1.0, ]$len)
```

```
## [1] 16.77
```

- **Dose of 2.0 mg.:**

```
mean(ToothGrowth[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 2.0, ]$len)
```

```
## [1] 26.06
```

```
mean(ToothGrowth[ToothGrowth$supp == "VC" & ToothGrowth$dose == 2.0, ]$len)
```

```
## [1] 26.14
```

We can confirm that the **orange juice** method is more effective in terms of length for lower dose levels and **ascorbic acid** has a slightly bigger impact in length for the 2 mg. dose.

Confidence intervals

For calculating the confidence intervals to compare tooth growth by supp and dose let's draw some convenient groups beforehand...

```
oj_05 <- ToothGrowth[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 0.5, ]
vc_05 <- ToothGrowth[ToothGrowth$supp == "VC" & ToothGrowth$dose == 0.5, ]
oj_10 <- ToothGrowth[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 1.0, ]
vc_10 <- ToothGrowth[ToothGrowth$supp == "VC" & ToothGrowth$dose == 1.0, ]
oj_20 <- ToothGrowth[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 2.0, ]
vc_20 <- ToothGrowth[ToothGrowth$supp == "VC" & ToothGrowth$dose == 2.0, ]
```

... and print them out.

```
oj_05
```

```
##     len supp dose
## 31 15.2   OJ  0.5
## 32 21.5   OJ  0.5
## 33 17.6   OJ  0.5
## 34  9.7   OJ  0.5
## 35 14.5   OJ  0.5
## 36 10.0   OJ  0.5
## 37  8.2   OJ  0.5
## 38  9.4   OJ  0.5
## 39 16.5   OJ  0.5
## 40  9.7   OJ  0.5
```

```
vc_05
```

```
##     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
```

```
## 7  11.2   VC  0.5
## 8  11.2   VC  0.5
## 9   5.2   VC  0.5
## 10  7.0   VC  0.5
```

```
oj_10
```

```
##     len supp dose
## 41 19.7   OJ    1
## 42 23.3   OJ    1
## 43 23.6   OJ    1
## 44 26.4   OJ    1
## 45 20.0   OJ    1
## 46 25.2   OJ    1
## 47 25.8   OJ    1
## 48 21.2   OJ    1
## 49 14.5   OJ    1
## 50 27.3   OJ    1
```

```
vc_10
```

```
##     len supp dose
## 11 16.5   VC    1
## 12 16.5   VC    1
## 13 15.2   VC    1
## 14 17.3   VC    1
## 15 22.5   VC    1
## 16 17.3   VC    1
## 17 13.6   VC    1
## 18 14.5   VC    1
## 19 18.8   VC    1
## 20 15.5   VC    1
```

```
oj_20
```

```
##     len supp dose
## 51 25.5   OJ    2
## 52 26.4   OJ    2
## 53 22.4   OJ    2
## 54 24.5   OJ    2
## 55 24.8   OJ    2
## 56 30.9   OJ    2
## 57 26.4   OJ    2
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

```
vc_20
```

```
##     len supp dose
## 21 23.6   VC    2
```

```
## 22 18.5   VC    2
## 23 33.9   VC    2
## 24 25.5   VC    2
## 25 26.4   VC    2
## 26 32.5   VC    2
## 27 26.7   VC    2
## 28 21.5   VC    2
## 29 23.3   VC    2
## 30 29.5   VC    2
```

Let's find the confidence interval for each pair of comparable groups using the **t-test**. The observations between the groups are naturally assumed to be statistically independent, so an independent group Student's T interval will be used instead of a paired one.

```
tt05 <- t.test(oj_05$len, vc_05$len, paired = FALSE)
tt05
```

```
##
##  Welch Two Sample t-test
##
## data:  oj_05$len and vc_05$len
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean of x mean of y
##     13.23      7.98
```

```
tt10 <- t.test(oj_10$len, vc_10$len, paired = FALSE)
tt10
```

```
##
##  Welch Two Sample t-test
##
## data:  oj_10$len and vc_10$len
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean of x mean of y
##     22.70     16.77
```

```
tt20 <- t.test(oj_20$len, vc_20$len, paired = FALSE)
tt20
```

```
##
##  Welch Two Sample t-test
##
## data:  oj_20$len and vc_20$len
## t = -0.0461, df = 14.04, p-value = 0.9639
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

- We can conclude that for dose of 0.5 mg., as the interval is above zero, **orange juice** is better than **ascorbic acid** in terms of tooth length: 1.7190573, 8.7809427.
- Then, for dose of 1.0 mg., as the interval is above zero too, **orange juice** is better than **ascorbic acid** in terms of tooth length: 2.8021482, 9.0578518.
- Finally, for dose of 2.0 mg., as the interval is almost symetric around zero (contains zero), suggesting **no difference** in terms of tooth length: -3.7980705, 3.6380705.

Further analysis

Let's draw 2 bigger groups from the original `ToothGrowth` dataset, `OJ` and `VC`, and check the **t-test** results.

```
oj <- ToothGrowth[ToothGrowth$supp == "OJ", ]
vc <- ToothGrowth[ToothGrowth$supp == "VC", ]
tt_by_method <- t.test(oj$len, vc$len, paired = FALSE)
tt_by_method
```

```
##
##  Welch Two Sample t-test
##
## data:  oj$len and vc$len
## 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
```

The left side of the confidence interval -0.1710156, 7.5710156 is slightly under zero, showing the influence of the **ascorbic acid** method for 2.0 mg. dose in this more general analysis.