ToothGrowth dataset analysis

Federico Calore

24 Oct 2015

This document performs some exploratory data analysis on the ToothGrowth dataset provided with the base R package, and checks some relevant hypotheses based on the data.

From the description in the dataset package:

> **The Effect of Vitamin C on Tooth Growth in Guinea Pigs**
> The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice or ascorbic acid (a form of vitamin C and coded as VC).

Dataset columns:

variable	type	description
$ len	numeric	Tooth length
$ supp	factor	Supplement type (VC or OJ).
$ dose	numeric	Dose in milligrams/day

Initial data exploration

Load the ToothGrowth data from R datasets and perform a quick summary:

```r
data("ToothGrowth")
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 ...
```

```r
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  
```

The dataset contains results of 10 samples for each combination of *supplement type* and *dose*.

```
## Count of observations:

##
##       0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

We can quickly check the mean of the tooth growth subsetted by *supplement type* and *dose*:

```
# mean of length by supplement type
tapply(ToothGrowth$len, INDEX = ToothGrowth$supp, FUN = mean)
```

```
      OJ       VC
20.66333 16.96333
```

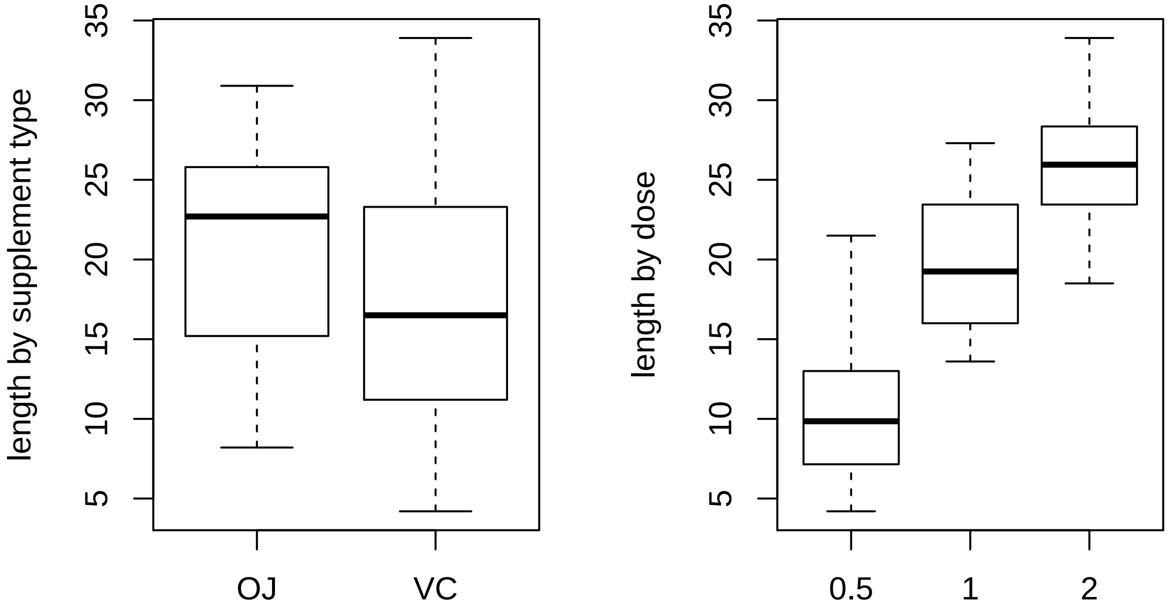
```
# mean of length by dose
tapply(ToothGrowth$len, INDEX = ToothGrowth$dose, FUN = mean)
```

```
   0.5      1      2
10.605 19.735 26.100
```

```
# mean of length by combinations of supplement and dose
xtabs(len/10 ~ supp + dose, data = ToothGrowth)
```

```
    dose
supp   0.5     1     2
  OJ 13.23 22.70 26.06
  VC  7.98 16.77 26.14
```

We can also explore the results with some boxplots to highlight differences in distribution of the tooth length variable by the *supp* and *dose* factors:



Hypothesis testing

Hypothesis testing on supplement type

Looking at the boxplot on the left, it looks like the ascorbic acid is less effective than the orange juice. We want to check if we have enough statistical evidence to support this. We can formally define the alternatives as follows:

$$H_0 : \mu_{OJ} = \mu_{VC}$$
$$H_1 : \mu_{OJ} \neq \mu_{VC}$$

We will use a two sided t-test to verify the interval of confidence in order to reject the *null* hypothesis H_0. We use unpaired observations, since the guinea pigs in one group didn't have anything to do with the ones in the other group, and assume unequal variances.

```
t <- t.test(len ~ supp, data = ToothGrowth)
t$conf[1:2]
## [1] -0.1710156  7.5710156
t$p.value
## [1] 0.06063451
```

The *p-value* is greater than 0.05 and the confidence interval contains zero, therefore we don't have enough statistical evidence to reject the *null* hypothesis with 95% confidence.

If we chose a relaxed *90% interval* for the test, or we used a *one-sided* hypothesis (*greater than*), the interval would not include zero, and we would reject the *null* hypothesis in favor of the alternative:

```
 # 90% confidence interval
t.test(len ~ supp, data = ToothGrowth, conf.level = 0.90)$conf[1:2]
```

```
## [1] 0.4682687 6.9317313
```

```
# one sided alternative hypothesis
t.test(len ~ supp, data = ToothGrowth, alternative = "greater")$conf[1:2]
```

```
## [1] 0.4682687       Inf
```

Hypothesis testing on doses

We have three levels for the doses, so we will have to compare them in pairs. The *null* hypothesis will assume an equal mean across levels, and the alternative will assume there is a difference.

As in the previous example, we will use two sided t-tests with unpaired observations and unequal variances. The result is formatted in a matrix to improve readability.

```
##   comparisons interval.1 interval.2 p.values
## 1     .5 vs 1    -11.984     -6.276    0e+00
## 2     .5 vs 2    -18.156    -12.834    0e+00
## 3      1 vs 2     -8.996     -3.734    2e-05
```

All three intervals are entirely below zero, so we would reject the *null* hypothesis in favor of the alternative in all cases. That implies *0.5 < 1 < 2*, or in other words, the more vitamin C, the stronger the effect.