Analysis of the ToothGrowth dataset

Cristian Bidea

10 April 2016

We set ourselves to analize the ToothGrowth dataset and see if we can come up with some insights.

Exploratory analysis

```
##       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  
```

First we do a summary of the dataset and we can see that the dataset is quite simple: just three columns. From the help file we find out more information about this particular dataset.

A data frame with 60 observations on 3 variables.

[,1] len numeric Tooth length [,2] supp factor Supplement type (VC or OJ). [,3] dose numeric Dose in milligrams/day

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).

We are now looking at two aspects:

1. First we want to know if the doses administered have different effects in each group (VC, OJ)
2. Second if one method is better than the other on the same administered dose.

As you can see in Fig. 1, in both methods increasing the dose increases the result and it seems that administering OJ is better for doses of 0.5 and 1 but then it becomes a little bit better for a dose of 2.

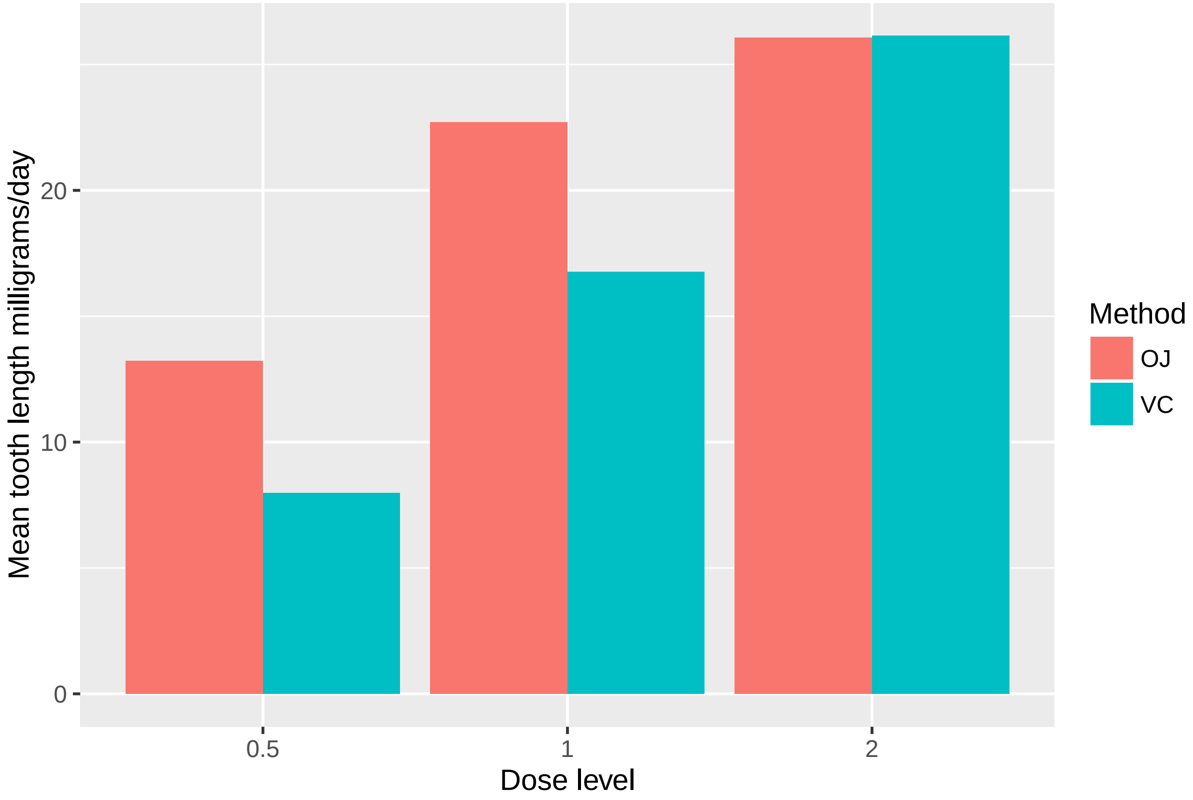
What I'm saying now is not rigorously tested. These are just intuition based conclusions formed looking at a graphic. We'll move on in the next section and we'll try to asses if these differences are in fact meaningful or not.

Confidence intervals and hypothesis testing

Why is computing confidence intervals useful? We plotted the mean of the samples depending on the method and dose but we must be confident that we actually have an increase. So we are computing confidence interval comparing the means of two independent populations: first comparing the 0.5 dose group with the 1 dose group and then the 1 dose group with the 2 dose group, in each of the two methods.

```
## Loading required package: ggplot2
##   len supp dose sd ci1 ci2 power
## 1   0   VC  0.5  0  NA  NA    NA
## 2   0   VC    1  0  NA  NA    NA
## 3   0   VC    2  0  NA  NA    NA
## 4   0   OJ  0.5  0  NA  NA    NA
## 5   0   OJ    1  0  NA  NA    NA
## 6   0   OJ    2  0  NA  NA    NA
```

Fig 1. Comparison of mean tooth lengths depending on dose and method



We need the standard deviation but we'll infer that from the CLT, knowing that

```
# computing the standard deviations
lvl <- "VC"
mean_tg[mean_tg$supp == lvl & mean_tg$dose == .5,]$sd <- sd(tg[tg$supp == lvl & tg$dose ==.5,]$len)
mean_tg[mean_tg$supp == lvl & mean_tg$dose == 1,]$sd <- sd(tg[tg$supp == lvl & tg$dose ==1,]$len)
mean_tg[mean_tg$supp == lvl & mean_tg$dose == 2,]$sd <- sd(tg[tg$supp == lvl & tg$dose ==2,]$len)
lvl <- "OJ"
mean_tg[mean_tg$supp == lvl & mean_tg$dose == .5,]$sd <- sd(tg[tg$supp == lvl & tg$dose ==.5,]$len)
mean_tg[mean_tg$supp == lvl & mean_tg$dose == 1,]$sd <- sd(tg[tg$supp == lvl & tg$dose ==1,]$len)
mean_tg[mean_tg$supp == lvl & mean_tg$dose == 2,]$sd <- sd(tg[tg$supp == lvl & tg$dose ==2,]$len)

for (x in c(1:2, 4:5))
{
    s1 <- mean_tg[x,]$sd
    s2 <- mean_tg[(x+1),]$sd
    mu1 <- mean_tg[x,]$len
    mu2 <- mean_tg[(x+1),]$len
    sd_p <- (9 * s1^2 + 9 * s2^2) / 18
    se <- sqrt(2*sd_p/10)
    mean_tg$ci1[x] <- mu2 - mu1 - qt(.95, 18) * se
    mean_tg$ci2[x] <- mu2 - mu1 + qt(.95, 18) * se
    mean_tg$power[x] <-power.t.test(n=20, delta = mu2 - mu1, sd = sd_p, type="one.sample", alt="one.sid
}

ci <- data.frame(
    ci1 = mean_tg$ci1[!is.na(mean_tg$ci1)],
    ci2 = mean_tg$ci2[!is.na(mean_tg$ci2)],
    method = mean_tg$supp[!is.na(mean_tg$ci1)],
    power = mean_tg$power[!is.na(mean_tg$ci1)]
)

ci
```

```
##         ci1       ci2 method     power
## 1 6.7477193 10.832281     VC 0.9999302
## 2 6.3994827 12.340517     VC 0.8658784
## 3 6.2173221 12.722678     OJ 0.7502929
## 4 0.7678858  5.952114     OJ 0.3639765
```

Looking at the confidence intervals we can see that there is an increase in effects that goes with the increase in dose because none of the four CIs don't touch 0.

Looking at the graphic in Fig. 1 and then observing that between 1 and 2 the increase in effects for the OJ method isn't that big, we ask ourselves if it wouldn't be better to take a lower dose (1) of OJ instead of taking a bigger dose (2) of VC.

So we compute one more confidence level comparing the means of these two groups to see if there is really an increase for a dose 2 of VC over a dose 1 of OJ.

```
##          ci1     ci2     power
## 1 0.04576028 6.83424 0.1920949
```

Conclusions

We've observed that there is a real increase of the effects that goes with an increased dose, especially for the VC group. For the OJ group the increase from dose 1 to dose 2 is questionable considering the low power.

A surprising result is conveyed by the confidence interval between the means of OJ(1) and VC(2) because this shows that the increase isn't that big and the low power showed us that a bigger sample is given to confirm that.

So it looks like the best bet is on OJ with a dose of 1, because the effects are bigger than in the case of VC with a dose of 1, and without a powerful increase in both VC and OJ with a dose of 2.