Tooth Growth analysis

Emanuele Di Saverio

21 Mar 2015

Introduction

In this document, we're going to analyze the ToothGrowth dataset, which reports data about an experiment carried on in 1946 about the growth of tissues (specifically, odontoblast cells, which is the "growable" part of the teeth between dentine and pulp) in relation to two ways of vitamin C intake.

The three variables of the test are

- the Supplementary substance used either Vitamin C extract or a natural Orange Juice
- Dosage, either 0.5, 1 or 2.0 milligrams
- the resulting growth of the teeth after 42 days

60 different guinea pigs were tested in any combinations of the variables dose and supp, resulting in 10 data points for each configuration.

```
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 ...
```

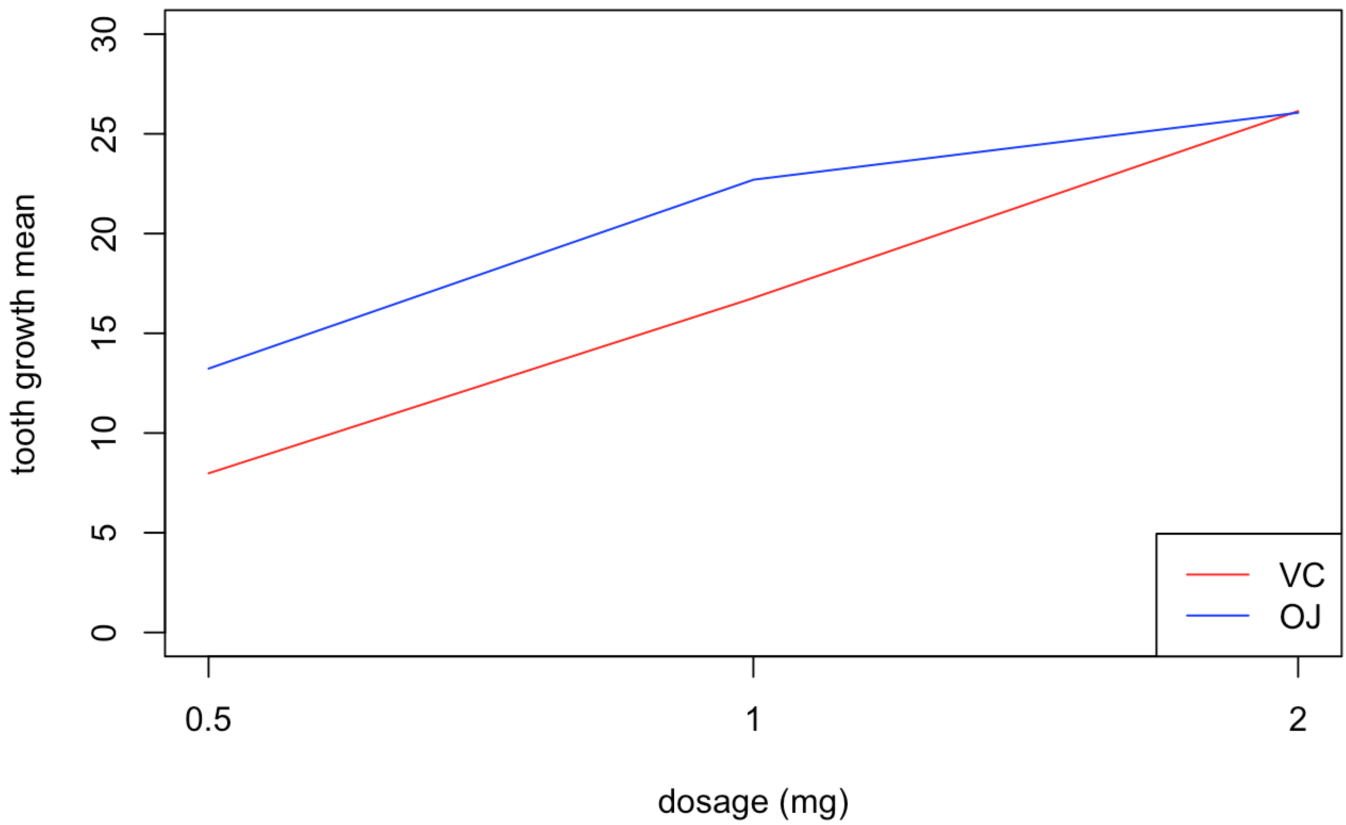
First, it seems reasonable to interpret dose as labels since continuity of the variable is not required

```
ToothGrowth$dose <- factor(ToothGrowth$dose)
```

Exploratory Analysis

A first analysis that we can do to visualize the data is averaging all observation for a given dosage and supplement and check their impact on the toot growth. This means averaging each of the group of 10 pigs and plotting.

```
onlyVC <- subset(ToothGrowth, supp == "VC")
onlyOJ <- subset(ToothGrowth, supp == "OJ")
growthVC <- cbind(levels(onlyVC$dose),tapply(onlyVC$len, onlyVC$dose, mean))
growthOJ <- cbind(levels(onlyVC$dose),tapply(onlyOJ$len, onlyOJ$dose, mean))
plot(growthVC[,2], t="l", col="red", ylim = c(0,30), xaxt="n", xlab ="dosage (mg)", ylab =
"tooth growth mean")
lines(growthOJ[,2], t="l",col="blue")
axis(1,at=1:3, labels=levels(onlyVC$dose))
legend("bottomright",legend=c("VC", "OJ"), col=c("red", "blue"), lty=c(1,1))
```

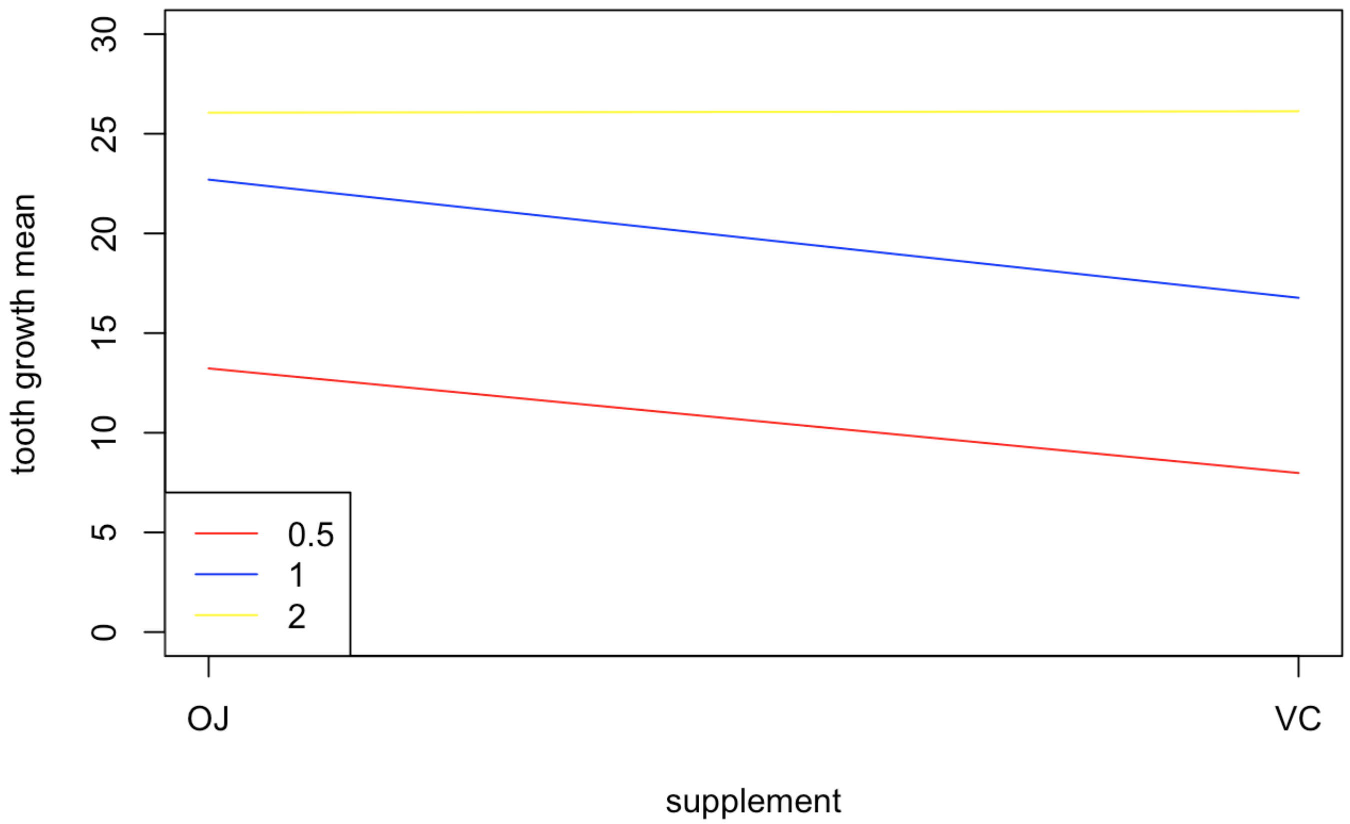


In the plot jsut shown, special care has been taken to not show the dosage on a linear scale (that is, intepreting those as numeric values) but as labels, since we cannot know and we don't want to assume that there is a linean relationship between the amount of vitamin C and the Tooth growth.

At this point we only want to show that one supplement relates to the other - looks that the simple orange juice is more effective at lower dosages, while at high dosage they seem equally effective.

This can be reinforced by the inverse chart, checking if, at a given dosage level, orange juice is more effective than Vitamin - C

```
only.5 <- subset(ToothGrowth, dose == "0.5")
only1 <- subset(ToothGrowth, dose == "1")
only2 <- subset(ToothGrowth, dose == "2")
growth.5 <- cbind(levels(only.5$supp),tapply(only.5$len, only.5$supp, mean))
growth1 <- cbind(levels(only1$supp),tapply(only1$len, only1$supp, mean))
growth2 <- cbind(levels(only2$supp),tapply(only2$len, only2$supp, mean))
plot(growth.5[,2], t="l", col="red", ylim = c(0,30), xaxt="n", xlab ="supplement", ylab =
"tooth growth mean")
lines(growth1[,2], t="l",col="blue")
lines(growth2[,2], t="l",col="yellow")
axis(1,at=1:2, labels=levels(only1$supp))
legend("bottomleft",legend=c("0.5", "1", "2"), col=c("red", "blue", "yellow"), lty=c(1,1))
```



The chart reinfoces our hypotheses.

Hypotheses Testing

Then the hypotheses we are going to test are three:

$H_0, .5 =$ There is no difference between means of tooth growth at .5 mg of dosage in case of juice or raw vitamin C

$H_a, .5 =$ Means of tooth growth at .5 mg of dosage is bigger in case of juice against raw vitamin C

$H_0, 1 =$ There is no difference between means of tooth growth at 1 mg of dosage in case of juice or raw vitamin C

$H_a, 1 =$ Means of tooth growth at 1 mg of dosage is bigger in case of juice against raw vitamin C

$H_0, 2 =$ There is no difference between means of tooth growth at 2 mg of dosage in case of juice or raw vitamin C

$H_a, 2 =$ Means of tooth growth at 2 mg of dosage is bigger in case of juice against raw vitamin C

Due to the really small sample size (10 sample for each mean) it is not the case to use standard confidence interval (which are based around the CLT, which may not hold at these low n values).

Since the pigs were different subjects, we cannot assume that the data is paired, and we cannot even assume that the two distributions compared (with vitamin C and Orange juice) have same variances (maybe the effect of the chemicals imply bigger or smaller variances).

A more proper tool in this case are the confdence intervals and tests based on the **t distribution**.

Lets first build a helper function to extract data vector the way we want it

```
getoothlength <- function(the_supp, the_dose) {
    subset(ToothGrowth, supp == the_supp & dose == the_dose)$len
}
```

Now heres the t-tests for the different hypotheses, We will be using the “greater” option of t.test since we want to explicitly test that the Orange Juice has greater effect (as our alternative hypotesis)

H0,.5

```
oj = getoothlength("OJ", "0.5")
vc = getoothlength("VC", "0.5")
t.test(oj,vc, paired= FALSE, var.equal=FALSE, alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  oj and vc
## t = 3.1697, df = 14.969, p-value = 0.003179
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.34604     Inf
## sample estimates:
## mean of x mean of y
##     13.23      7.98
```

As we can see that the 95% confidence interval for the means difference is $[2.34604; Inf]$, and **it does NOT include zero** we can reject the null hypothesis $H_0, .5 =$. We not also that the p value is 0.003, well under the 5% mark.

H0,1

```
oj = getoothlength("OJ", "1")
vc = getoothlength("VC", "1")
t.test(oj,vc, paired= FALSE, var.equal=FALSE, alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  oj and vc
## t = 4.0328, df = 15.358, p-value = 0.0005192
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.356158      Inf
## sample estimates:
## mean of x mean of y
##     22.70     16.77
```

As we can see that the 95% confidence interval is $[3.356158; Inf]$, and **it does NOT include zero** we can reject the null hypothesis $H_0, 1 =$. Here we the p-value is 0.0005, so we can be even more certain - as our exploratory charting activity hinted. #### H0,2

```
oj = getoothlength("OJ", "2")
vc = getoothlength("VC", "2")
t.test(oj,vc, paired= FALSE, var.equal=FALSE, alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  oj and vc
## t = -0.0461, df = 14.04, p-value = 0.5181
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -3.1335     Inf
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

The 95% confidence interval is $[3.1335; Inf]$ and **it does include zero**, so we cannot reject the null hypothesis $H_0, 2 =$.

Conclusion

For low and medium dosages of vitamin C, the orange juice is more effective in growing the dental tissue in guinea pigs. For high dosages, there is no difference.

Following this analysis we can assert **with statistical significance** what we intuitively showed graphically during our preliminary analysis.

Assumptions

An untested assumption that this data has underlying is that the vitamin C has any effect at all over the growth of the teeth of the pigs. I would have liked to test against a control group of pigs who get to drink plain water instead of ascorbic acid or orange juice.

The analysis has been carried without assumption of equal variance between OJ and VC pigs, but knowing a little bit more of medicine science can maybe bring to strengthen the assumption to equal variances. In the end, having more nutritive substance in the body may affect the growth in a linear (multiplicative) way.