Statistical Inference Course Project - Part 2: Simulations

Alejandro Osorio

March 28, 2018

We're going to analyze the ToothGrowth data in the R datasets package.

General Parameters

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data("ToothGrowth")
```

Basic exploratory data analysis

Starting with a quick view at the table structure:

```
head(ToothGrowth)
```

```
##    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
```

So it consists of 3 variables: Length, Supplement and Dose.

Now a quick glance at the table's size:

```
dim(ToothGrowth)
```

```
## [1] 60  3
```

So it's a small sample, therefore it'll require Student's test.

Just in case, let's check it there are any NAs in there:

```
table(is.na(ToothGrowth))
```

```
##
## FALSE
##   180
```

So there are no NAs. Now a quick glance at the variable's attributes:

```
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 ...
```

It shows length as a numeric variable, the supplements as 2 types, stored as a factor ("OJ" and "VC"), and the dose as a repeating number. Let's check how supplements are related to which dosages then:

```
with(ToothGrowth,table(dose, supp))
```

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

As we can see, the sample consists of 3 dosages, for which there are 10 records of each supplement.

Additionally, a summary of the data to check any further information of interest:

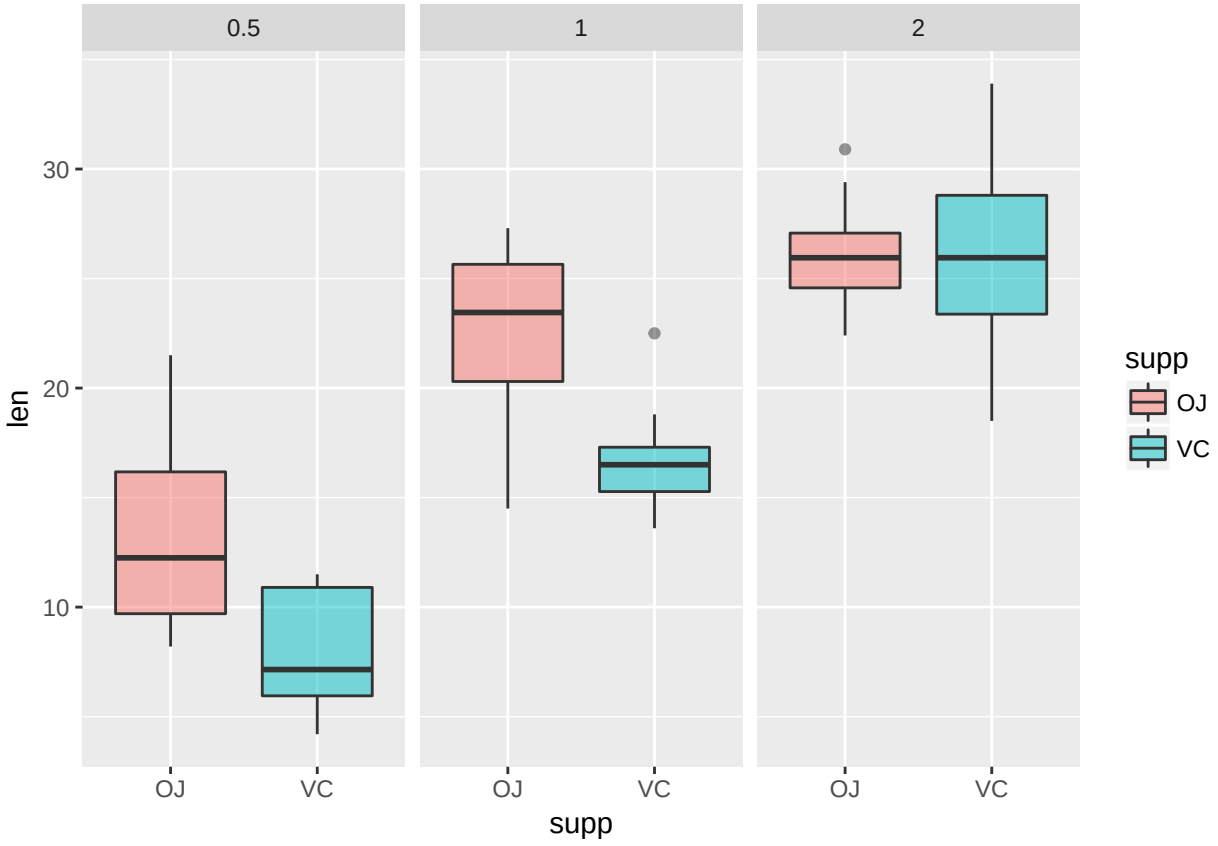
```
summary(ToothGrowth)
```

```
##       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
```

It shows a clear difference between the minimum and maximum tooth length, suggesting there's a good reason for some hypotheses to check.

Finally, a visual analysis of the data, using boxplots by supplement and dose, to check for evident patterns with which to work later.

```
ggplot(ToothGrowth, aes(x = supp, y = len)) +
        geom_boxplot(aes(fill = supp), alpha = 0.5) +
        facet_wrap(~ dose, nrow = 1)
```



It suggests a clear relation between both supplement and dosage, with length.

The questions that arise from the exploratory analysis are, then:

1.- Can it be infered that, with higher dosages, longer teeth will be attained?

2.- If point 1 results to be true, Which supplement is more effective (obtains longer teeth)?

Hypotheses testing

Both questions are tackled through hypothesis testing, based on the following assumptions:

a) Variables are iid
b) Samples are not paired from one dosage to the next (different subjects were monitored for each trial, with no relation whatsoever between them).
c) Underlying distributions for each sample, along dosages and supplements, have different variances.

Question 1: Higher dosages imply longer teeth?

The null hypothesis, for this case, is: Teeth growth is not affected by dosage.

The alternative hypothesis is that, the higher the dosage, the longer teeth growth obtained.

In order to cover the whole spectrum of dosages (0.5, 1 and 2), two hypothesis testings are carried on: a) one for the dosage interval of 0.5 - 1. b) the second one for the interval of 1 - 2, as follows:

Hypothesis testing for range 0.5 - 1

Selected samples are those associated to both dosages of 0.5 and dosages of 1, as follows:

```
dose05 <- filter(ToothGrowth, ToothGrowth$dose == 0.5) %>%
        .$len
dose1 <- filter(ToothGrowth, ToothGrowth$dose == 1) %>%
        .$len
```

The one sided Student's test, then, based on the previous assumptions (included by default in the t.test) plus the filtered samples, goes as follows:

```
t.test(dose1, dose05)
```

```
##
##  Welch Two Sample t-test
##
## data:  dose1 and dose05
## t = 6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.276219 11.983781
## sample estimates:
## mean of x mean of y
##    19.735    10.605
```

Which shows a greater teeth growth for dose1 (mean = 19.73) over dose05 (mean = 10.61), with a low enough p-value for the alternative hypothesis to replace the null one.

Hypothesis testing for range 1 - 2

Based on the same previous logic, the extra required sample, for dosage of 2 is obtained as follows:

```
dose2 <- filter(ToothGrowth, ToothGrowth$dose == 2) %>%
        .$len
```

The one sided Student's test, then, goes as follows:

```
t.test(dose2, dose1)
```

```
##
##  Welch Two Sample t-test
##
## data:  dose2 and dose1
## t = 4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.733519 8.996481
## sample estimates:
## mean of x mean of y
##    26.100    19.735
```

Which shows a greater teeth growth for dose2 (mean = 26.10) over dose1 (mean = 19.74), with a low enough p-value for the alternative hypothesis to, again, replace the null one.

Question 2: Which supplement is more effective (obtains longer teeth)?

The null hypothesis, for this case, is: Teeth growth is the same for both supplements.

The alternative hypothesis is that, OJ supplement is more effective than VC.

In this case, just one hypothesis testing is carried on, with the samples being those associated to OJ and VC supplements respectively, as follows:

```
suppoj <- filter(ToothGrowth, ToothGrowth$supp == "OJ") %>%
        .$len
suppvc <- filter(ToothGrowth, ToothGrowth$supp == "VC") %>%
        .$len
```

The one sided Student's test, then, based on the previous assumptions (included by default in the t.test) plus the filtered samples, goes as follows:

```
t.test(suppoj, suppvc)
```

```
##
##  Welch Two Sample t-test
##
## data:  suppoj and suppvc
## 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
```

Which shows higher effectiveness of supplement OJ (mean = 20.66), over supplement VC (mean = 16.96) with a low enough p-value for the alternative hypothesis to replace the null one.

Final Conclusions

Based on the previous preliminary data analysis, plus hypotheses testing, it can be concluded that:

1.- The higher the dosage, the longer the teeth growth obtained.

2.- Supplement OJ is more effective than supplement VC.