# Statistical Inference Course Project - Part 2: Basic Inferential Data Analysis

geotsa

September 16, 2019

## Task

In this report we're analyzing the ToothGrowth data in the R datasets package

### Loading the dataset

We are loading the necessary libraries:

- The "datasets" that contains the ToothGrowth dataset and

```
library(datasets)
```

- the graphics grammar"ggplot2" for its faceting capacity

```
library(ggplot2)
```

We can now load the ToothGrowth dataset

```
data(ToothGrowth)
?(ToothGrowth)
```

### Overview of the dataset

The dataset ToothGrowth deals with 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).

---

## Exploratory Data Analysis

Inspect dataset's structure

```r
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 ...
```

its first 5 rows

```r
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
```

and the summaries (five-number summaries and the means) of its variables (columns)

```r
summary(ToothGrowth)
```

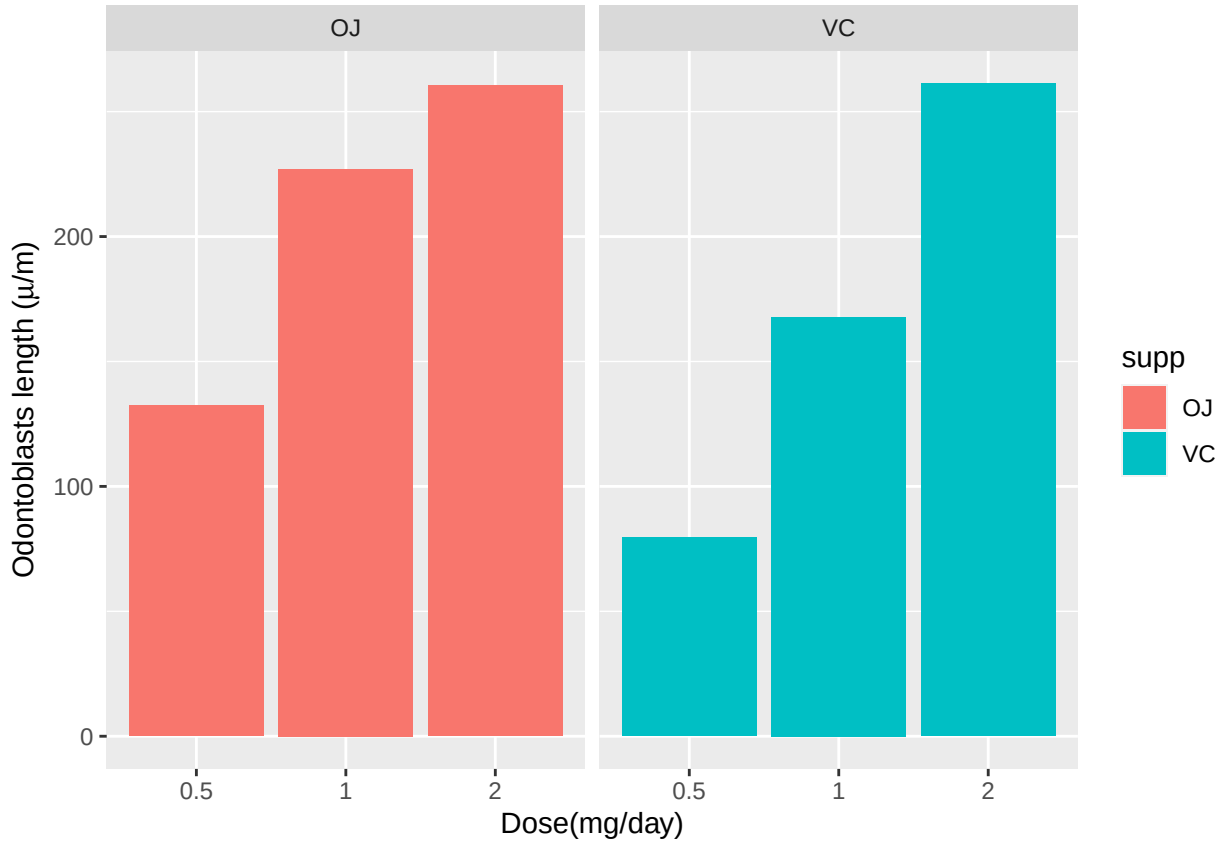
```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

We also correct the data, transfmrming the dose variable from numeric to factor as it can get only three values (0.5, 1.0, 2.0)

```r
ToothGrowth$dose<-as.factor(ToothGrowth$dose)
```

Finally we try to understand graphically the basic patterns in odontoblasts growing relating to the delivery method (OJ, VC) and dose levels of vitamin C(0.5, 1.0, 2.0 mg/day)

```r
ggplot(data=ToothGrowth, aes(x=dose, y=len, fill=supp)) +
    geom_bar(stat="identity") +
    facet_grid(. ~ supp) +
    xlab("Dose(mg/day)") +
    ylab(expression(paste("Odontoblasts length (",mu,"/m)")))
```




## Basic summary of the data. The data consist of 60 observations of the length of the odontoblasts. In 30 of them, vitamin C is provided by orange juice (OJ) and in the other 30 by ascorbic acid (VC). In each of these two cases, vitamin C is administered it doses of 0.5, 1.0 or 2.0 mg/day (10 obs for each sub-case, since Mean(dose) = 1.167 = (0.5+1.0 + 2.0) / 3). The exploratory analysis of the data shows that, between the administration of vitamin C by OJ or VC, there is some difference in the length of odontoblasts (especially in the 0.5 and 1.0 mg / day sub-cases). The OJ is appearing a stronger positive relationship with this length.

---

# Inferential Data Analysis

The foregoing leads to the formulation of more or less reasonable hypotheses. In what follows, we will attempt to test these hypotheses and draw conclusions about their confidence intervals and the statistical significance of observed trends.

## Hypotheses

### 1. Growth by supp (dose 0.5 mg/day case)

```
# We are selecting the obs corresponding to OJ AND dose of 0.5 mg/day
OJ05 <- ToothGrowth[which(ToothGrowth$supp=="OJ"&ToothGrowth$dose==0.5),1]
# and calculating their mean
mean(OJ05)
```

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

```
# We are selecting the obs corresponding to VC AND dose of 0.5 mg/day
VC05 <- ToothGrowth[which(ToothGrowth$supp=="VC"&ToothGrowth$dose==0.5),1]
# and calculating their mean
mean(VC05)
```

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

```
# Difference of the means of OJ and VC for a dose of 0.5mg/day
mean(OJ05)-mean(VC05)
```

```
## [1] 5.25
```

This last number (5.25) makes us want to test the hypothesis that the difference between the two mean values is statistically nsignificant against the alternativy hypothtsis that, for a dose of 0.5 mg/day, the odontoblasts length is significantly bigger in the case of vitamin C delivery by orange juice rather than by ascorbic acid.

```
t.test(OJ05, VC05, alternative = "g", paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ05 and VC05
## 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
```

Let's recap:

Higher doses of vitamin C cause more tooth growth.

H0: Difference in means (for 0.5 mg/day) for OJ and for VC is around 0

HA: Difference in means is greater than 0

Since the p-value (0.0031793) is less than the significance level (a=0.05) and the confidence interval is always bigger than 0, there is enough evidence to reject the null hypothesis. Therefore, as for 5 mg/day dose, orange juice has a significantly positive impact in greater tooth growth than ascorbic acid.

**2. Growth by supp (dose 1.0 mg/day case)**

```
# We are selecting the obs corresponding to OJ AND dose of 1.0 mg/day
OJ10 <- ToothGrowth[which(ToothGrowth$supp=="OJ"&ToothGrowth$dose==1.0),1]
# and calculating their mean
mean(OJ10)
```

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

```r
# We are selecting the obs corresponding to VC AND dose of 1.0 mg/day
VC10 <- ToothGrowth[which(ToothGrowth$supp=="VC"&ToothGrowth$dose==1.0),1]
# and calculating their mean
mean(VC10)
```

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

```r
# Difference of the means of OJ and VC for a dose of 1.0mg/day
mean(OJ10)-mean(VC10)
```

```
## [1] 5.93
```

This last number (5.93) makes us again want to test the hypothesis that the difference between the two mean values is statistically insignificant against the alternativy hypothtsis that, for a dose of 1.0 mg/day, the odontoblasts length is significantly bigger in the case of vitamin C delivery by orange juice rather than by ascorbic acid.

```r
t.test(OJ10, VC10, alternative = "g", paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ10 and VC10
## 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
```

Let's recap:

Higher doses of vitamin C cause more tooth growth.

H0: Difference in means (for 1.0 mg/day) for OJ and for VC is around 0

HA: Difference in means is greater than 0

Since the p-value (0.0005192) is less than the significance level (a=0.05) and the confidence interval is always bigger than 0, there is enough evidence to reject the null hypothesis. Therefore, as for 1.0 mg/day dose, orange juice has a significantly positive impact in greater tooth growth than ascorbic acid.

**3. Growth by supp (dose 2.0 mg/day case)**

```r
# We are selecting the obs corresponding to OJ AND dose of 2.0 mg/day
OJ20 <- ToothGrowth[which(ToothGrowth$supp=="OJ"&ToothGrowth$dose==2.0),1]
# and calculating their mean
mean(OJ20)
```

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

```r
# We are selecting the obs corresponding to VC AND dose of 2.0 mg/day
VC20 <- ToothGrowth[which(ToothGrowth$supp=="VC"&ToothGrowth$dose==2.0),1]
# and calculating their mean
mean(VC20)
```

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

```r
# Difference of the means of OJ and VC for a dose of 2.0 mg/day
mean(OJ20)-mean(VC20)
```

```
## [1] -0.08
```

Contrary to the previous cases this last number (-0.08) doesn't seem significant but we will test the hypothesis that the difference between the two mean values is statistically insignificant agaist the alternative hypothesis that (always for a dose of 2.0 mg/day) the odontoblasts length difference is significantly different in the case of vitamin C delivery by orange juice rather than by ascorbic acid.

```r
t.test(OJ20, VC20, alternative = "t", paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ20 and VC20
## t = -0.046136, 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
```

Let's recap:

Higher doses of vitamin C cause more tooth growth.

H0: Difference in means (for 2.0 mg/day) for OJ and for VC is around 0

HA: Difference in means is *not equal to 0

The p-value (0.9638516) is greater than the significance level (a=0.05) and the confidence interval contains zero so that we can say that there is not enough evidence to reject the null hypothesis. Therefore, application methods have no impact on tooth growth in the case of doses of 2.0 mg/day.

**4. Growth by supp (general case)**

```r
# We are selecting the obs corresponding to OJ
OJgen <- ToothGrowth[which(ToothGrowth$supp=="OJ"),1]
# and calculating their mean
mean(OJgen)
```

```
## [1] 20.66333
```

```r
# We are selecting the obs corresponding to VC
VCgen <- ToothGrowth[which(ToothGrowth$supp=="VC"),1]
# and calculating their mean
mean(VCgen)
```

```
## [1] 16.96333
```

```r
# Difference of the means of OJ and VC
mean(OJgen)-mean(VCgen)
```

```
## [1] 3.7
```

This last number (3.7) makes us again want to test the hypothesis that the difference between the two mean values is statistically insignificant or not. The alternative hypothesis is that, independently from the dose, the odontoblasts length is significantly bigger in the case of vitamin C delivery by orange juice rather than by ascorbic acid.

```r
t.test(OJgen, VCgen, alternative = "g", paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJgen and VCgen
## t = 1.9153, df = 55.309, p-value = 0.03032
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4682687       Inf
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

Let's recap:

Higher doses of vitamin C cause more tooth growth.

H0: Difference in means for OJ and for VC is around 0

HA: Difference in means is greater than 0

The p-value (0.0303173) is less than the significance level (a=0.05) and the confidence interval contains zero so that we can say that there is enough evidence to reject the null hypothesis. Therefore, even barely, the t.test shows that the application methods, irrespective of dose quantity, have a significant impact on tooth growth.

## Assumptions

The following conclusions are made on the basis of the assumptions that:

- the observations in the sample are effectively independent and identically distributed random variables (i.i.d.)
- the sample is representative of the population

## Conclusions

OJ ensures more tooth growth than VC for dosages 0.5 & 1.0. OJ and VC gives the same amount of tooth growth for dose amount 2.0 mg/day. For the entire trail we can conclude OJ is more effective than VC for the general case scenario.