Inference Analysis: the Tooth Growth case

Leandro GARCIA

OVERVIEW

In this project we will look into the Tooth Growth data set available in R package. This dataset contains the response of 2 groups of guinea pigs in the administration of Vitamin C dosis by 2 different ways. The response is the length of odontoblasts (cells responsible for tooth growth). 60 guinea pigs were tested. 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 (coded as *OJ*) or ascorbic acid (a form of vitamin C used as a preservative in food items and coded as *VC*).

About the Dataset

The Data set contains 60 observations and 3 features.An observation represent a subject for the a specific suplement. The features are :

- Len : Length of odontoblasts. This feature will be named as 'Length'

- dose : Vitamin C dose. This feature will be named as 'Dosage'

- Supp: The administrated supplement (Vitamin C). This feature will be named as 'Supplement'

    - OC: for Orange Juice

    - VC: Vitamin C in Ascorbic Acid form

-

Basic descriptive statics

```
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 ...
```

```
unique(ToothGrowth$dose)
```

```
## [1] 0.5 1.0 2.0
```

We have then, 2 main group (OC, VC), and 3 sub-groups by dosage (0.5, 1.0, 2.0). We will then proceed to group them and get the basic statistics:

```
data_df <- ToothGrowth

library(dplyr)
```

```
##
```

```
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
names(data_df) <- c("Length", "Supplement", "Dosage")

summary_df <- group_by(.data = data_df, Supplement, Dosage) %>% summarise_all(.funs = c("Mean" = mean,

summary_df
```

```
## # A tibble: 6 x 5
## # Groups:   Supplement [2]
##   Supplement Dosage  Mean    SD   Var
##   <fct>       <dbl> <dbl> <dbl> <dbl>
## 1 OJ            0.5 13.2   4.46 19.9
## 2 OJ            1   22.7   3.91 15.3
## 3 OJ            2   26.1   2.66  7.05
## 4 VC            0.5  7.98  2.75  7.54
## 5 VC            1   16.8   2.52  6.33
## 6 VC            2   26.1   4.80 23.0
```

CONFIDENCE INTERVALS

```r
auxiliary_df <- group_by(.data = data_df, Supplement, Dosage) %>% tally()
auxiliary_df$df <-auxiliary_df$n - 1

merged_df <- merge(x = summary_df, y = auxiliary_df, by = c("Supplement", "Dosage"))
merged_df
```

```
##   Supplement Dosage  Mean       SD       Var  n df
## 1         OJ    0.5 13.23 4.459709 19.889000 10  9
## 2         OJ    1.0 22.70 3.910953 15.295556 10  9
## 3         OJ    2.0 26.06 2.655058  7.049333 10  9
## 4         VC    0.5  7.98 2.746634  7.544000 10  9
## 5         VC    1.0 16.77 2.515309  6.326778 10  9
## 6         VC    2.0 26.14 4.797731 23.018222 10  9
```

```r
t <- qt(p = 0.95, df=9, lower.tail = TRUE)

merged_df$CIinf <- merged_df$Mean -  t * merged_df$SD / sqrt(merged_df$n)
merged_df$CIsup <- merged_df$Mean +  t * merged_df$SD / sqrt(merged_df$n)
merged_df$Confidence <- 0.90

select(.data=merged_df, Supplement, Dosage, df, Confidence, CIinf, Mean, CIsup)
```

```
##   Supplement Dosage df Confidence      CIinf  Mean     CIsup
## 1         OJ    0.5  9        0.9 10.644791 13.23 15.815209
## 2         OJ    1.0  9        0.9 20.432894 22.70 24.967106
## 3         OJ    2.0  9        0.9 24.520913 26.06 27.599087
## 4         VC    0.5  9        0.9  6.387828  7.98  9.572172
```

```
## 5          VC    1.0  9       0.9 15.311923 16.77 18.228077
## 6          VC    2.0  9       0.9 23.358846 26.14 28.921154
```

HYPOTHESIS TESTING

Our hypothesis is that for doses of 2.0 mg, the supplement type doesn't have a great effect on the tooth growth rate. In other words. $H_0 : \mu_{OJ,2.0} = \mu_{VC,2.0}$

So, our alternative Hipothesis is : $H_0 : \mu_{OJ,2.0} \neq \mu_{VC,2.0}$

ASSUMPTIONS

- All the numerical variables presented here are IID (Independent and Identical Distribuited)
- The observations may not come from the same population, so we are assuming:
  - A difference variance for each group
  - A subject only belong to a group.
- A t-test is the most suitable as we couldn't confirm that the distribution for each group is normal (see annexes)

t.test() in R

Parameters: - alternative = "two.sided"

- paired = FALSE
- var.equal = FALSE
- conf.level = 0.95

```
data_df <- mutate(.data = data_df, "Group" = paste(Supplement, as.character(Dosage), sep="_"))
t.test(x = data_df$Length[data_df$Group == "OJ_2"], y = data_df$Length[data_df$Group == "VC_2"], alterna
```

```
##
##  Welch Two Sample t-test
##
## data:  data_df$Length[data_df$Group == "OJ_2"] and data_df$Length[data_df$Group == "VC_2"]
## 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
```

CONCLUSIONS

Even when our exploratory analysis (see ANNEXES) gives the sensation that because that the group OJ_2 and VC_2 overlaps, their means are the same, in reality they are not and there is a difference between the 2 means.

ANNEXES

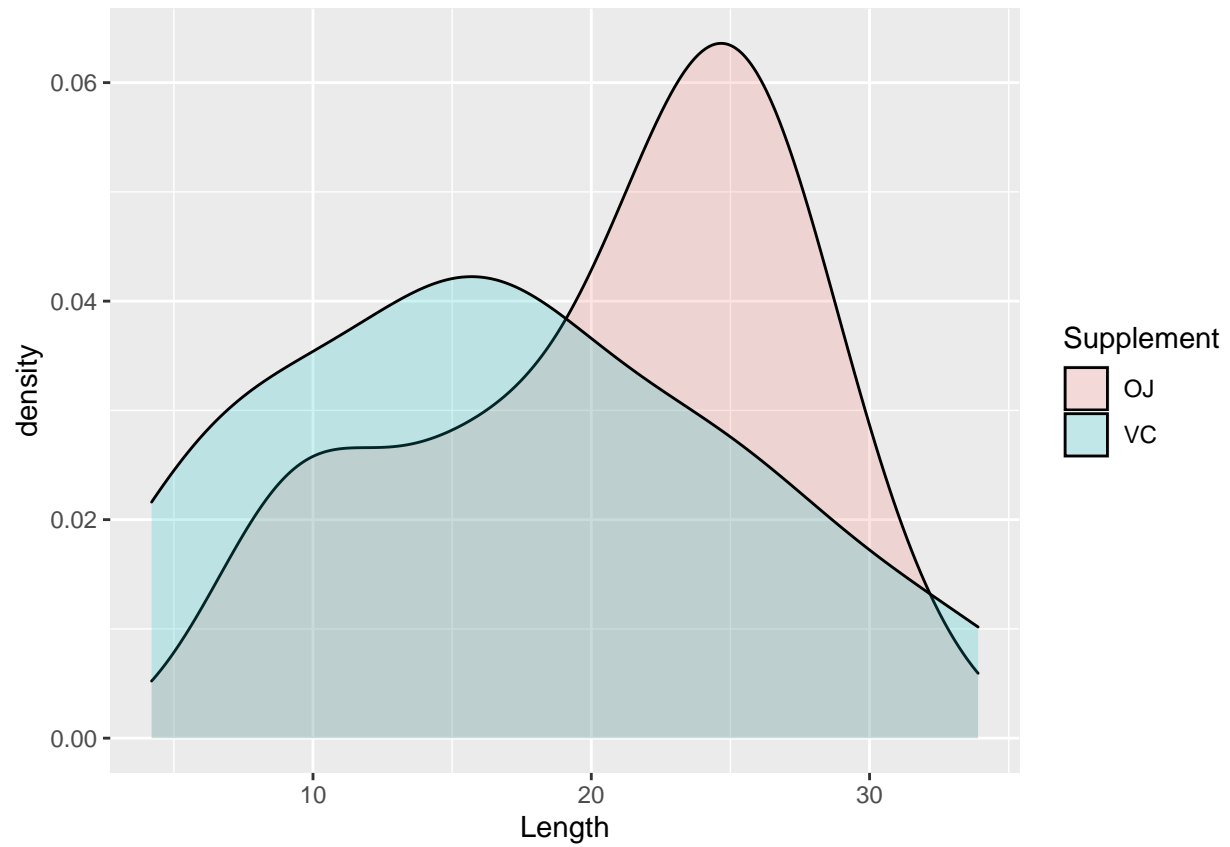## Basic Exploratory Analysis - Checking the distribution on the 6 groups

```
library(ggplot2)
data_df <- mutate(.data = data_df, "Group" = paste(Supplement, as.character(Dosage), sep="_"))
ggplot(data = data_df, aes(x = Length, fill = Group)) + geom_density(alpha = 0.2)
```
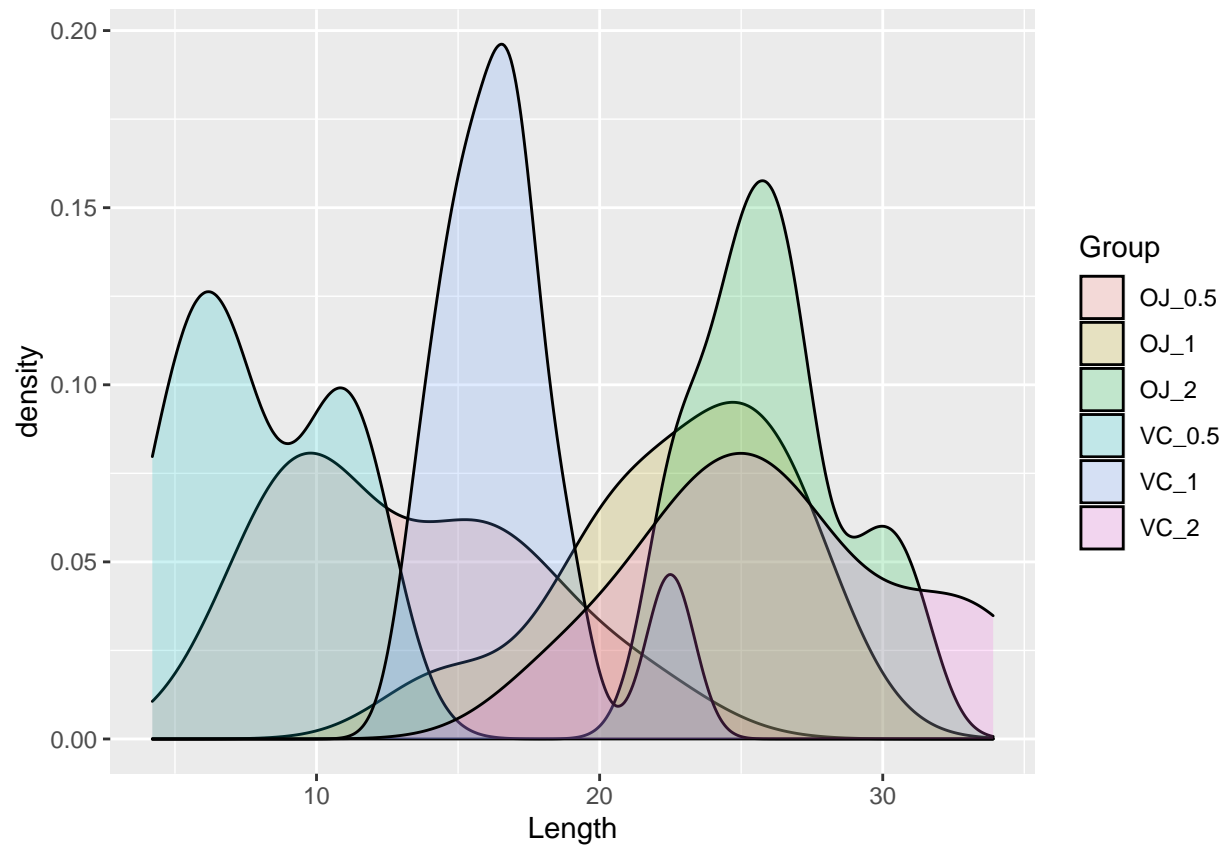

With the quantity of subjects, it is not clear if they follow a Normal distribution for the 6 groups, therefore we cannot easily kept the idea that the variables are 'normal distributed'.

Basic Exploratory Analysis - Checking the distribution on the 2 groups

```
ggplot(data = data_df, aes(x = Length, fill = Supplement)) + geom_density(alpha = 0.2)
```

We can neigher say that the 2 groups are normal distributed.

REFERENCE

- https://stat.ethz.ch/R-manual/R-patched/library/datasets/html/ToothGrowth.html