Descriptive Statistical Analysis with R

Paolo Girardi and Livio Finos

15/10/2020

Contents

Ph.D. Course in Neuroscience
Calendar of the Basic Courses – Academic Year 2020-2021
Basic Concept of Statistics

Lesson 3 - Descriptive Statistical Analysis with R and tests

First phases

Descriptive analysis is used to describe the basic features of the data in the study. They provide simple summaries about the sample and the measures. Together with simple graphical analysis, they form the basic virtual of any quantitative analysis of data.

```
# remove all in the R environment
rm(list=ls())
```

Now we import a dataset in EXCEL format. Let's install a package to do that (package **readxl**).

```
# if not installed, digit install.packages("readxl")
library(readxl)
```

Now we import the dataset "cat_ex.xlsx" in EXCEL format.

```
setwd("/Users/Paolo/Dropbox/Dottorato_Neurosciences")
DATASET <- read_excel("cat_ex.xlsx")
```

Let's see what we have imported.

```
View(DATASET)
dim(DATASET)
```

```
## [1] 63  5
```

```
str(DATASET)
```

```
## tibble [63 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Id        : num [1:63] 1 1 1 2 2 2 3 3 3 4 ...
##  $ Gruppi    : chr [1:63] "HC" "HC" "HC" "HC" ...
##  $ condizioni: chr [1:63] "Volti" "Scene" "Parole" "Volti" ...
##  $ Y1        : num [1:63] 0.662 0.864 0.762 0.71 0.813 ...
##  $ Y2        : num [1:63] 0.996 0.87 1.271 1.483 0.825 ...
```

```
DATASET=as.data.frame(DATASET)

#
```

The data is formed by 21 subjects who took part in a study measuring the cognitive ability through a verbal fluency test. The study enrolled healthy controls (11) and subjects with the Parkison Disease (10).

The test consisted on:
- a phonological fluency test with the use of three letters (different at each condition);
- a semantic fluency test using three categories (the type of condition is reported in the variable "condizioni").

We have 5 variables:
- ID: subject ID
- Gruppi: HC= Healthy Control; LE=Parkison Disease
- Condizioni: type of "subject" on the fluency test
- Y1: Phonemic fluency index: Z-score on the fluency test - Phonemic
- Y2: Semantic fluency index: Z-score on the fluency test - Semantic

```r
DATASET$Gruppi=factor(DATASET$Gruppi)
DATASET$condizioni=factor(DATASET$condizioni)
str(DATASET)
```

```
## 'data.frame':    63 obs. of  5 variables:
##  $ Id        : num  1 1 1 2 2 2 3 3 3 4 ...
##  $ Gruppi    : Factor w/ 2 levels "HC","LE": 1 1 1 1 1 1 1 1 1 1 ...
##  $ condizioni: Factor w/ 3 levels "Parole","Scene",..: 3 2 1 3 2 1 3 2 1 3 ...
##  $ Y1        : num  0.662 0.864 0.762 0.71 0.813 ...
##  $ Y2        : num  0.996 0.87 1.271 1.483 0.825 ...
```

```r
# We can change the name
names(DATASET)
```

```
## [1] "Id"         "Gruppi"     "condizioni" "Y1"         "Y2"
```

```r
names(DATASET)[4:5]<-c("Phonemic","Semantic")
names(DATASET)
```

```
## [1] "Id"         "Gruppi"     "condizioni" "Phonemic"   "Semantic"
```

```r
# I can calculate the difference between the Z-score on phonological and semantic test.
DATASET$delta=DATASET$Phonemic-DATASET$Semantic
DATASET$delta
```

```
##  [1] -0.333604757 -0.005080331 -0.509661819 -0.772923006 -0.012313640
##  [6]  0.006906715 -0.516024693 -0.723058795 -0.182181532 -0.687452482
## [11] -0.140116687  0.216840937 -0.805774885 -0.967011564 -0.557855064
## [16] -0.188255956 -1.379416347 -0.431996308 -0.867955450 -0.596997921
## [21]  0.001557256 -0.541236789 -0.832196235 -0.401281630 -1.399747826
## [26]  1.553087875  1.729171409 -0.362111916 -0.791746169  0.114356303
## [31] -0.954005200 -0.777088182 -1.735565316 -1.550650883 -1.452093166
## [36] -1.870504585 -0.808755358 -1.066325518 -0.208962929 -0.484628239
## [41] -0.798624740 -0.331374774 -1.134931507 -5.841208263 -1.681766687
## [46] -1.979407799 -0.609277539 -0.525475889 -0.900821427 -0.532275204
## [51] -0.190593973  0.107637326  0.350684682  0.271671808 -0.784577710
## [56] -0.710914458 -0.466386930 -0.374177497 -0.570108028  0.223138391
## [61]  0.049615349 -0.119425555  0.341029698
```

Univariate analysis

A simple way is to perform a separate analysis for each variable.

```
table(DATASET$Id)
```

```
## 
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 
##  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3
```

```
# 3 tests for each ID
table(DATASET$Gruppi)
```

```
## 
## HC LE 
## 33 30
```

```
# 33 for HC, 30 for LE
table(DATASET$condizioni)
```

```
## 
## Parole  Scene  Volti 
##     21     21     21
```

```
# Condition is repeacted 21 times each ID

#Some indices for the quantitative variables
summary(DATASET$Phonemic)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## -3.9973  0.3246  0.5623  0.4169  0.6999  0.9374
```

```
summary(DATASET$Semantic)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## -1.2170  0.7278  1.0764  1.0121  1.3589  2.3298
```

```
summary(DATASET$delta)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## -5.8412 -0.8205 -0.5323 -0.5952 -0.1611  1.7292
```

```
# other indices
mean(DATASET$Phonemic)
```

```
## [1] 0.4168511
```

```
sd(DATASET$Phonemic)
```

```
## [1] 0.6393836
```

```
median(DATASET$Phonemic)
```

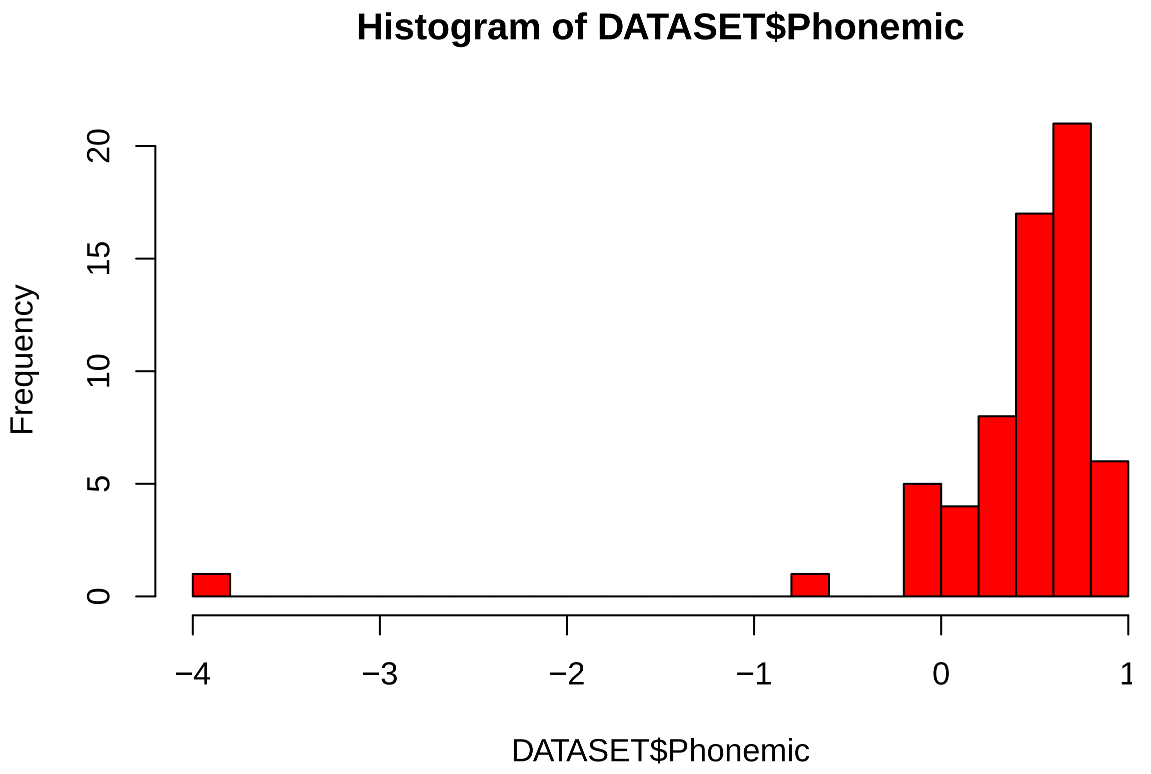
```
## [1] 0.5623233
```

```
IQR(DATASET$Phonemic)
```

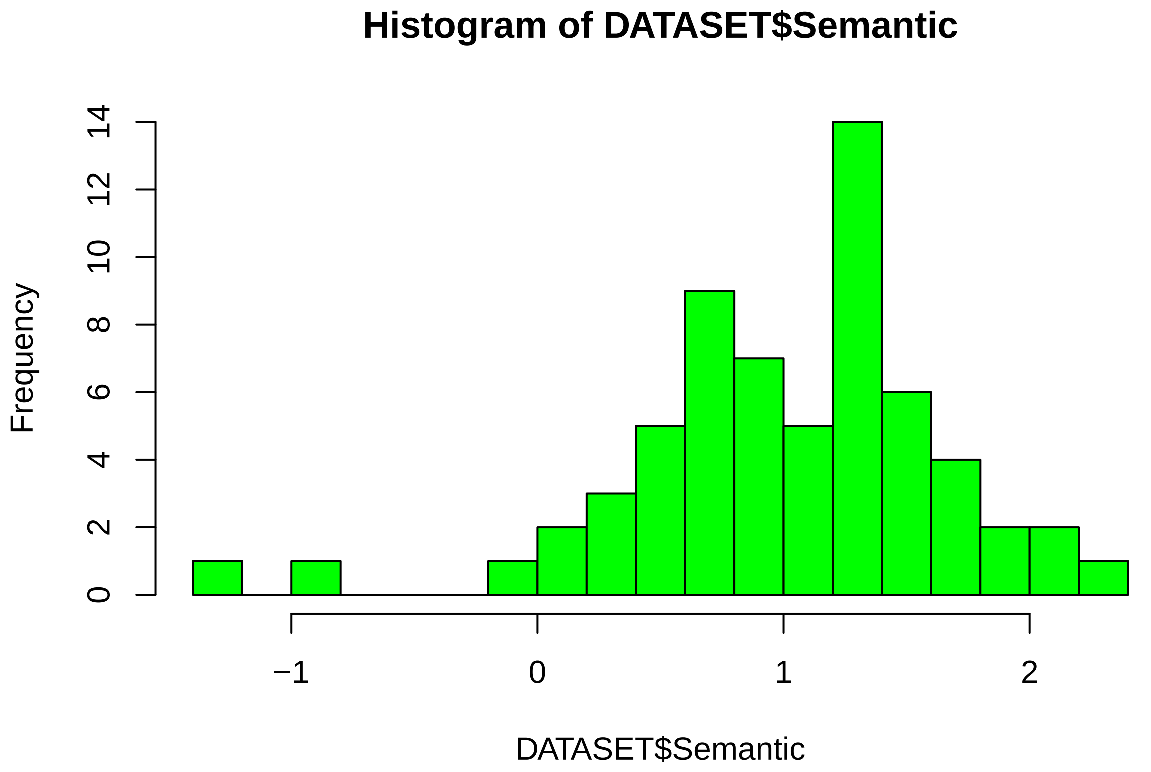
```
## [1] 0.3752947
```

Some Figures:

```
#Histogram
hist(DATASET$Phonemic, breaks = 20,col="red")
```

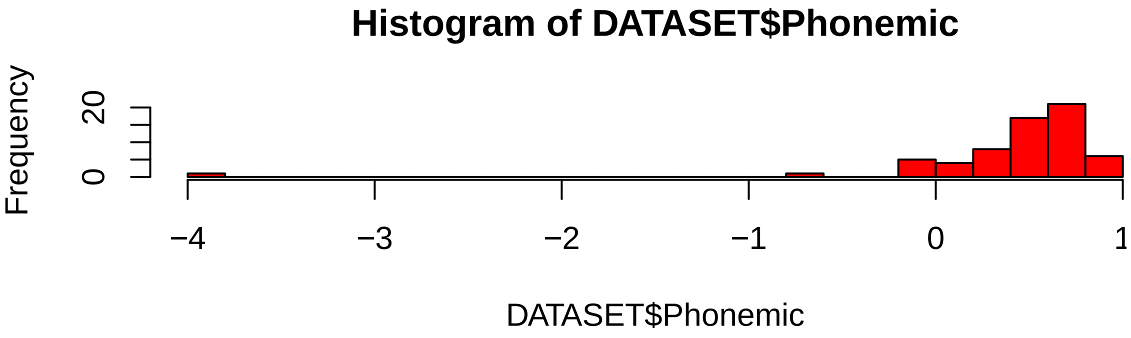


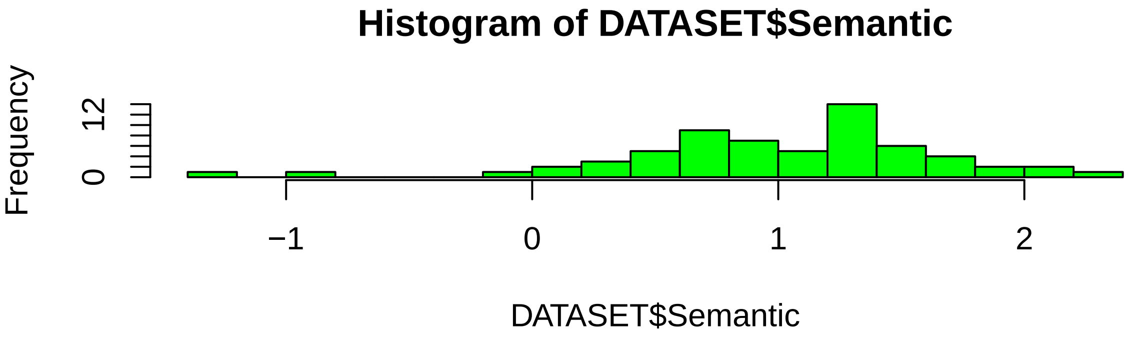
```
hist(DATASET$Semantic, breaks = 20,col="green")
```



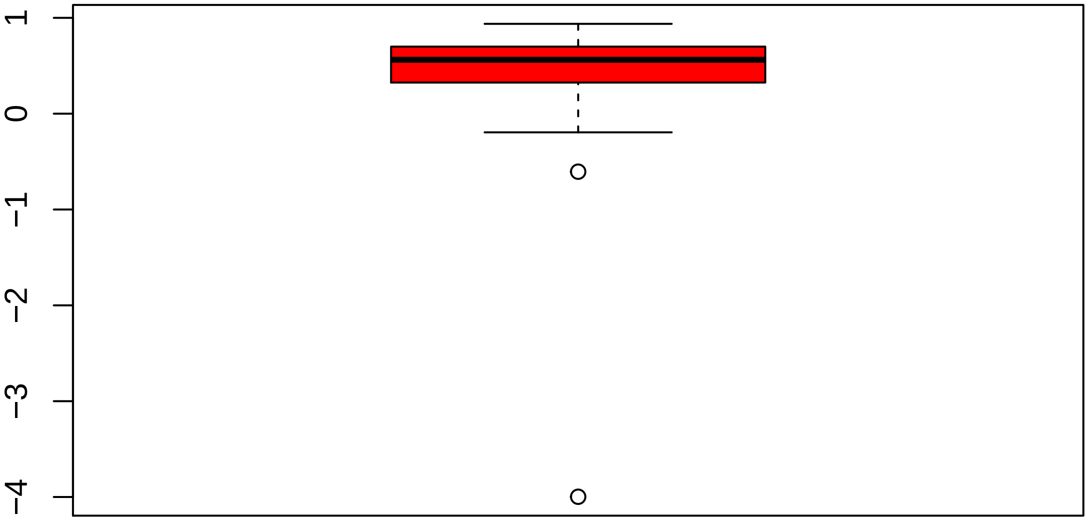
```
#Together
par(mfrow=c(2,1))
hist(DATASET$Phonemic, breaks = 20,col="red")
```

```r
hist(DATASET$Semantic, breaks = 20,col="green")
```

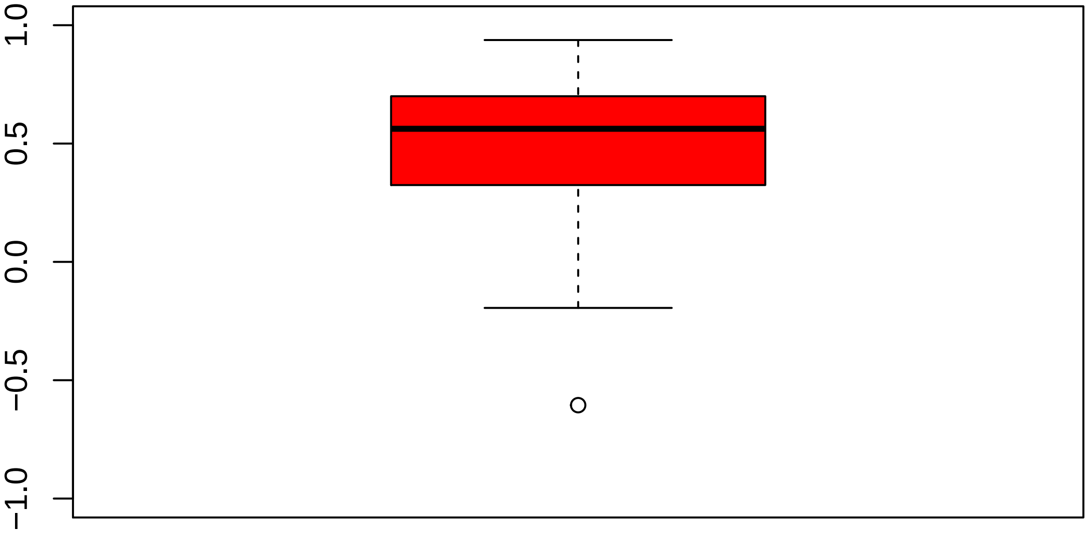




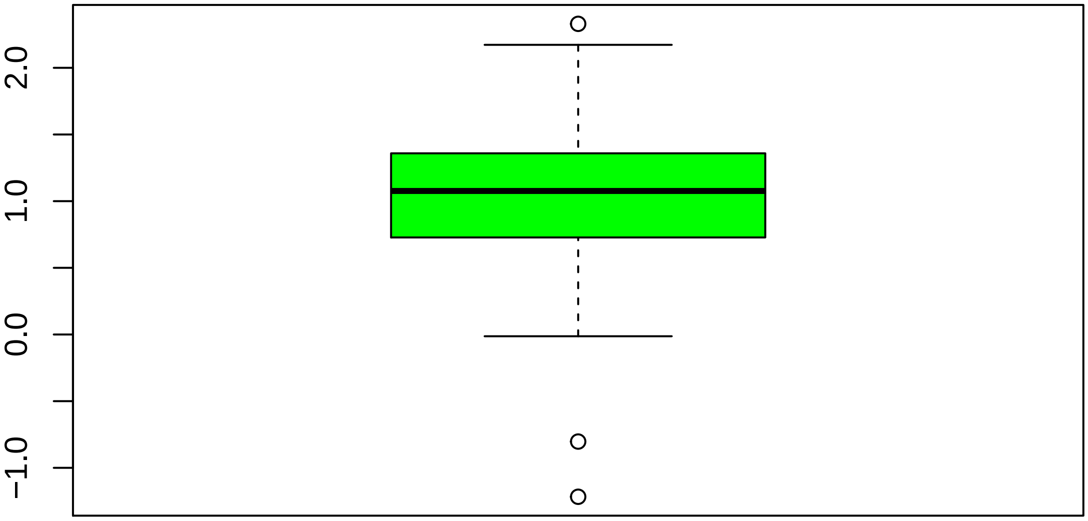
```r
par(mfrow=c(1,1))

#Boxplot
boxplot(DATASET$Phonemic, breaks = 20,col="red")
```



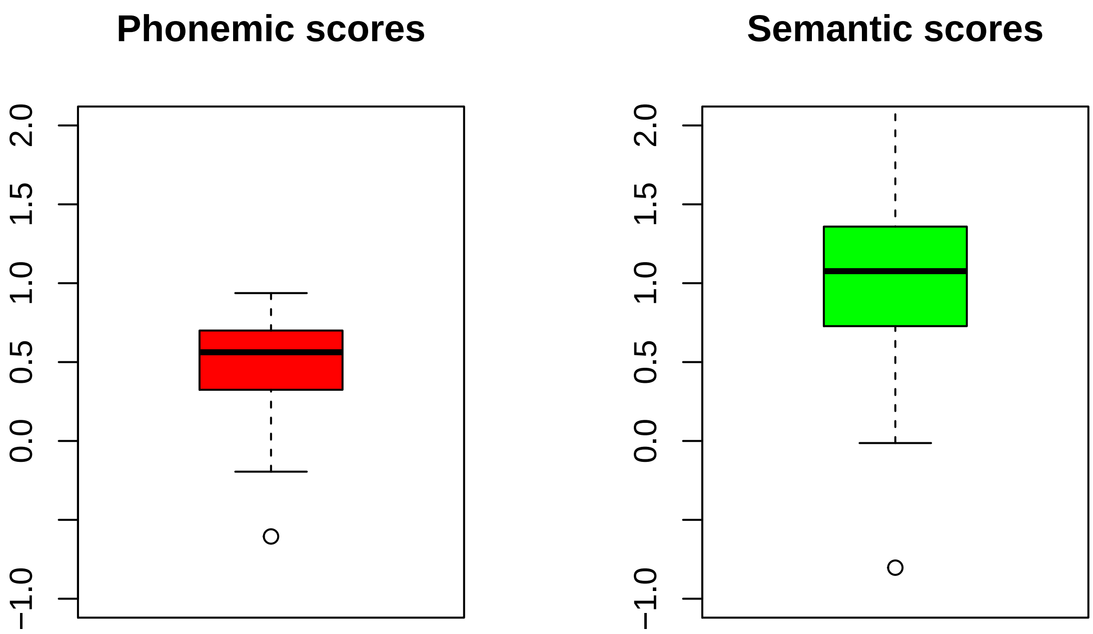
```r
#the presence of an outlier... I can limit the y axis extension from -1 to 1.
boxplot(DATASET$Phonemic, breaks = 20,col="red",ylim=c(-1,1))
```



```
boxplot(DATASET$Semantic, breaks = 20,col="green")
```



```
par(mfrow=c(1,2))
boxplot(DATASET$Phonemic, breaks = 20,col="red",ylim=c(-1,2),main="Phonemic scores")
boxplot(DATASET$Semantic, breaks = 20,col="green",ylim=c(-1,2),main="Semantic scores")
```



```
par(mfrow=c(1,1))
```

Bivariate analysis

We use a package ("tabs") to produce table (to export in Latex, Word, Html).

```
# if not installed, digit install.packages("tab")
library(tab)
```

```
## Loading required package: dplyr
```

```
## Warning: replacing previous import 'vctrs::data_frame' by 'tibble::data_frame'
## when loading '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
```

```
## Loading required package: knitr
```

```
# we use the function tabmulti, please see tabmulti help (?tabmulti)
#table by variable " Gruppi"
tab1<-tabmulti(data=DATASET, condizioni+Phonemic+Semantic+delta~Gruppi)
# The numeric variables are summarized with MEAN and SD ad a p-value with t.test. I can change to MEDIA
library(knitr)
kable(tab1)
```

Variable	HC	LE	P
condizioni, n (%)			1.00
Parole	11 (33.3)	10 (33.3)	
Scene	11 (33.3)	10 (33.3)	
Volti	11 (33.3)	10 (33.3)	
Phonemic, M (SD)	0.64 (0.16)	0.17 (0.85)	0.005
Semantic, M (SD)	1.06 (0.66)	0.96 (0.59)	0.50
delta, M (SD)	-0.42 (0.69)	-0.79 (1.15)	0.13

```
tab1b<-tabmulti(data=DATASET, condizioni+Phonemic+Semantic+delta~Gruppi,
                ymeasures = c("freq","median","median","median"))
```

```
## Warning in wilcox.test.default(x = c(0.66216038, 0.864458494, 0.761693083, :
## cannot compute exact p-value with ties
```

```
## Mann-Whitney U was used to test whether the distribution of Phonemic differs in the two groups.
```

```
## Mann-Whitney U was used to test whether the distribution of Semantic differs in the two groups.
```

```
## Mann-Whitney U was used to test whether the distribution of delta differs in the two groups.
```

```
#p.values are performed with a non parametric mann-withney test
kable(tab1b)
```

Variable	HC	LE	P
condizioni, n (%)			1.00

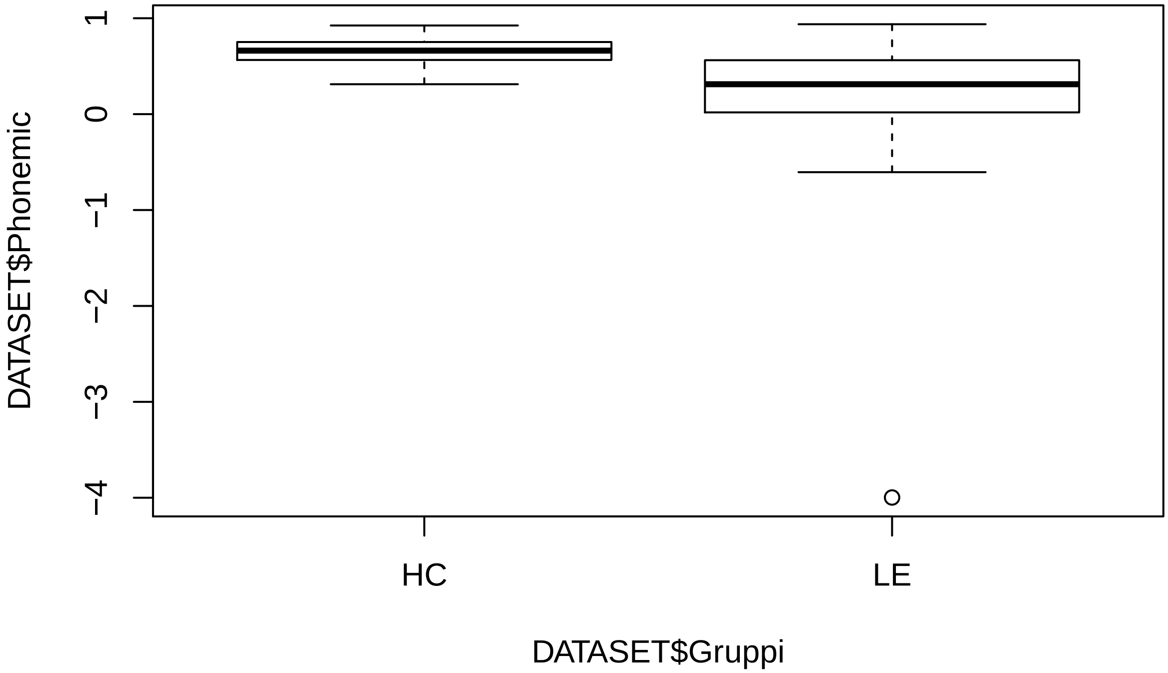
Variable	HC	LE	P
Parole	11 (33.3)	10 (33.3)	
Scene	11 (33.3)	10 (33.3)	
Volti	11 (33.3)	10 (33.3)	
Phonemic, Median (IQR)	0.66 (0.19)	0.31 (0.52)	<0.001
Semantic, Median (IQR)	1.23 (0.57)	0.79 (0.82)	0.15
delta, Median (IQR)	-0.52 (0.65)	-0.55 (0.83)	0.44

```
# I can export the tables in HTML format by means of print.html = TRUE, html.filename = "table1.html" p

# By condition
tab2<-tabmulti(data=DATASET, Phonemic+Semantic+delta~condizioni)
kable(tab2)
```

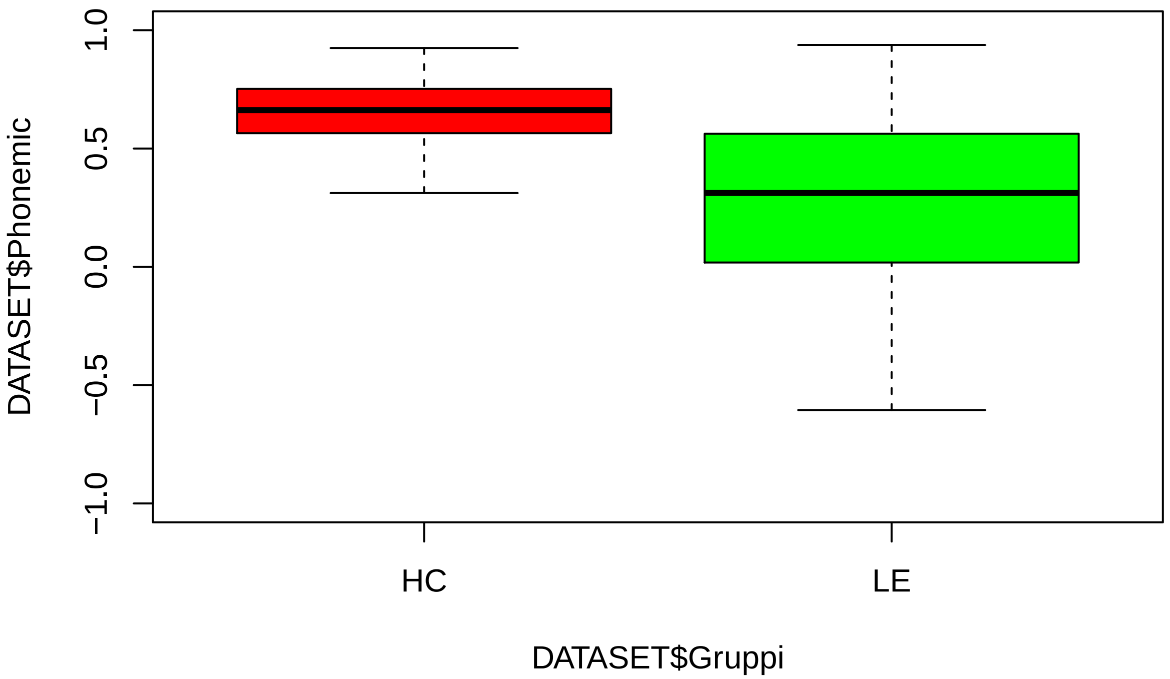
Variable	Parole	Scene	Volti	P
Phonemic, M (SD)	0.50 (0.38)	0.28 (1.02)	0.48 (0.24)	0.47
Semantic, M (SD)	0.79 (0.74)	1.04 (0.61)	1.21 (0.46)	0.09
delta, M (SD)	-0.29 (0.79)	-0.76 (1.33)	-0.73 (0.51)	0.21

And some bivariate graphs, Phonemic score.

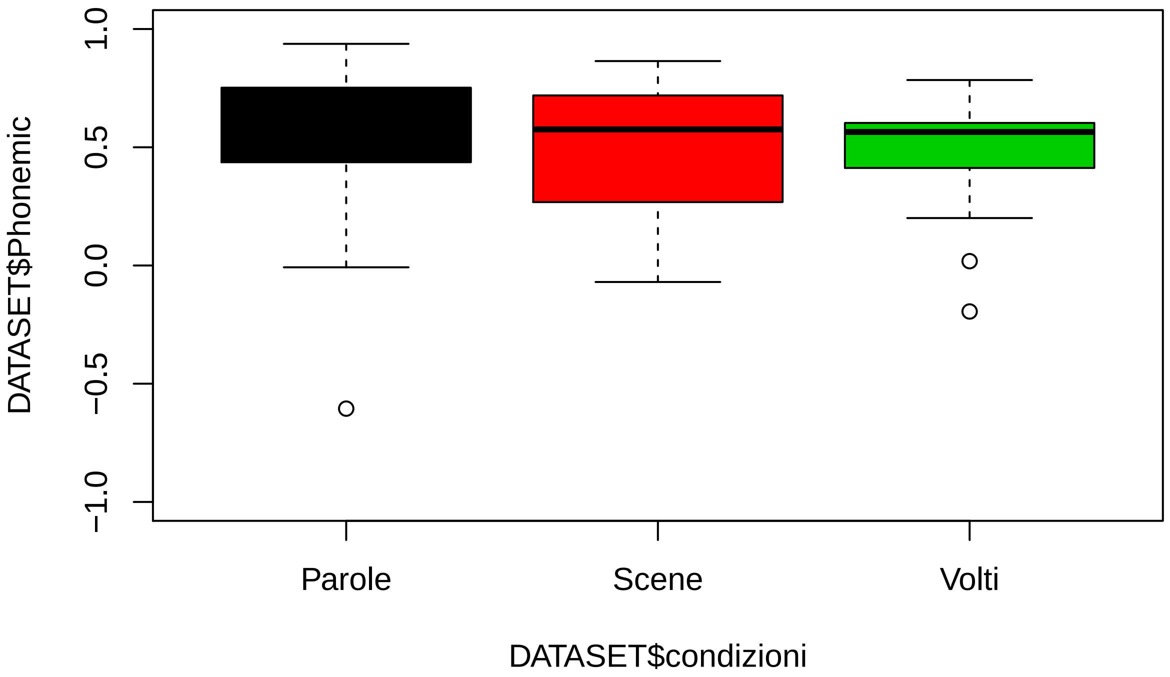
```
# by Gruppi
boxplot(DATASET$Phonemic~DATASET$Gruppi)
```



```
# add limits and colours
boxplot(DATASET$Phonemic~DATASET$Gruppi,ylim=c(-1,1),col=c("red","green"))
```

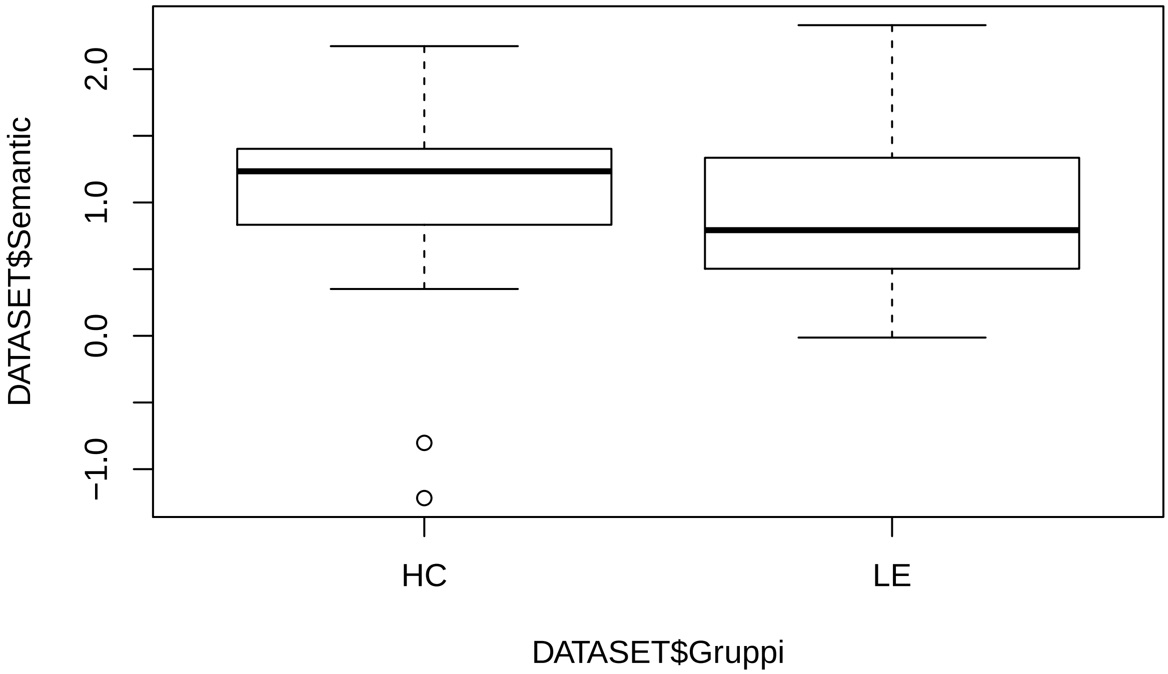


```
#by condizioni
boxplot(DATASET$Phonemic~DATASET$condizioni,ylim=c(-1,1),col=1:3)
```

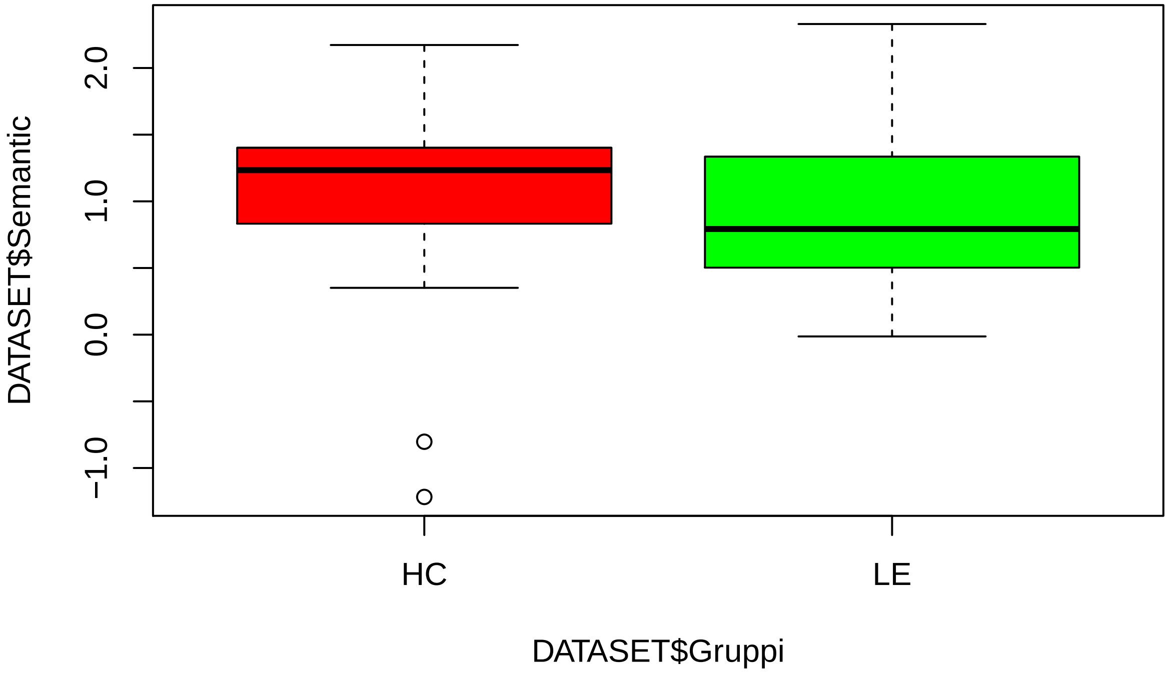


Semantic score

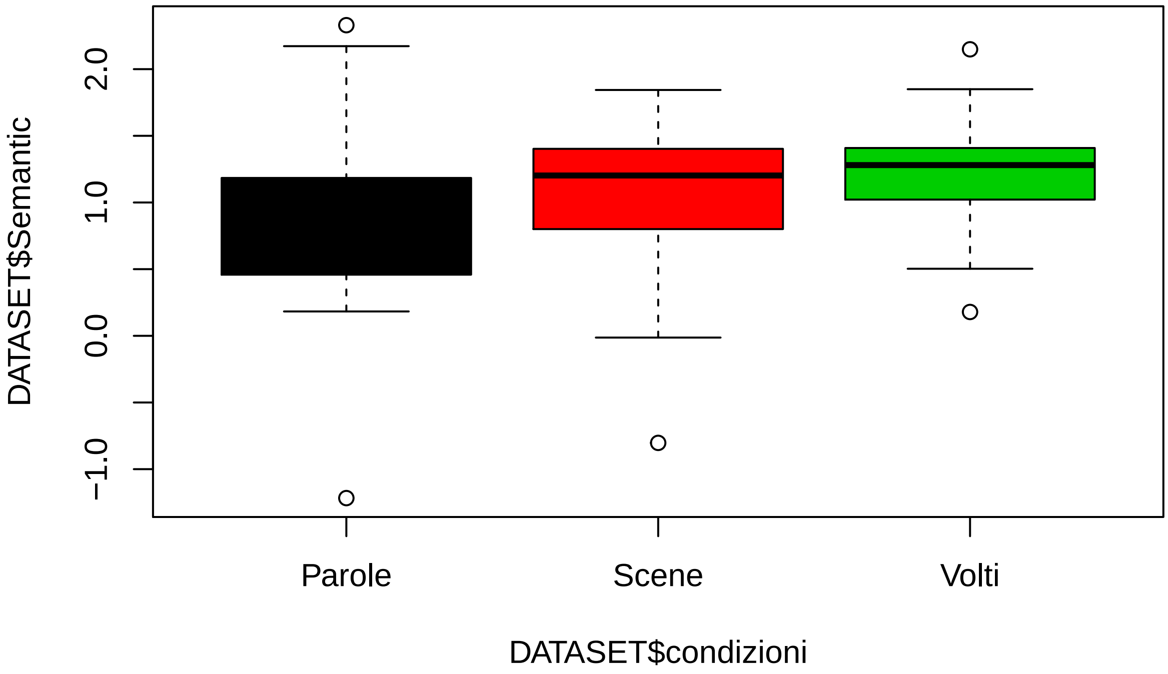
```
# by Gruppi
boxplot(DATASET$Semantic~DATASET$Gruppi)
```



```
# add limits and colours
boxplot(DATASET$Semantic~DATASET$Gruppi,col=c("red","green"))
```



```
#by condizioni
boxplot(DATASET$Semantic~DATASET$condizioni,col=1:3)
```



We can generate statistical analysis by means of functions done by ourselves. In particular we are going to use the library "doBy" that permits to perform a function by an other variable.

```
# if not installed, digit install.packages("doBy")
library(doBy)
```

```
##
## Attaching package: 'doBy'

## The following object is masked from 'package:dplyr':
##
##     order_by
```

```
#This function calculate for a numeric vector
# MEAN, MEDIAN, VARIANCE AND THE LENGTH
fun <- function(x){
  c(m=mean(x), me=median(x), v=var(x), n=length(x))
}
#I use the function summaryBy to apply the function "fun" by type of the variable "Gruppi" and "Condizi
summaryBy(Semantic ~ condizioni+Gruppi, data=DATASET,
          FUN=fun)
```

```
##   condizioni Gruppi Semantic.m Semantic.me Semantic.v Semantic.n
## 1     Parole     HC  0.8527763   0.8893248 0.73078861         11
## 2     Parole     LE  0.7220497   0.5448075 0.38738667         10
## 3      Scene     HC  1.0848226   1.2340554 0.50287548         11
## 4      Scene     LE  0.9891230   0.8357029 0.26934627         10
## 5      Volti     HC  1.2524907   1.2804987 0.08319085         11
## 6      Volti     LE  1.1559139   1.2867281 0.36407590         10
```

```
summaryBy(cbind(delta,Phonemic) ~ condizioni+Gruppi, data=DATASET,
          FUN=fun)
```

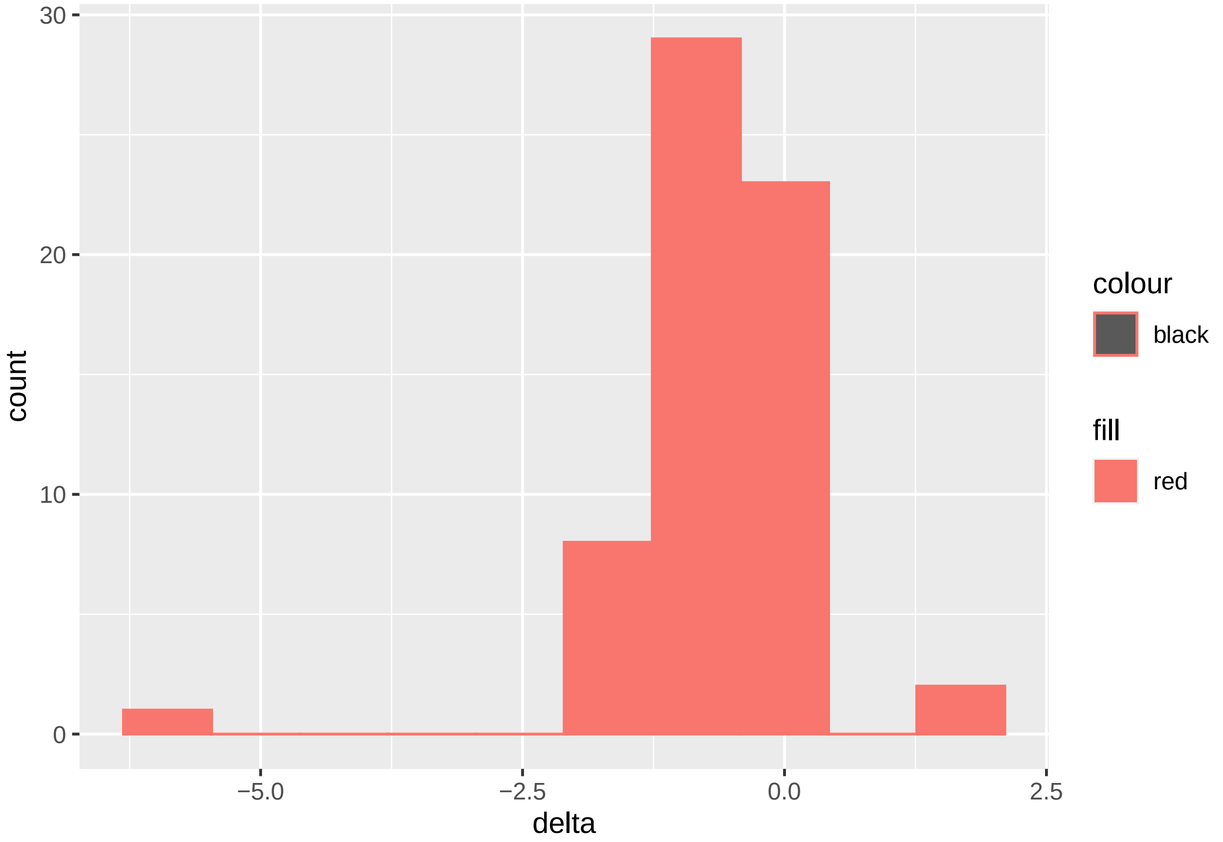
```
##   condizioni Gruppi    delta.m   delta.me   delta.v delta.n Phonemic.m
## 1     Parole     HC -0.1590645 -0.1821815 0.6735847      11  0.6937118
## 2     Parole     LE -0.4439226 -0.2701689 0.5866438      10  0.2781271
## 3      Scene     HC -0.4247216 -0.7230588 0.6088019      11  0.6601010
## 4      Scene     LE -1.1349568 -0.6600960 2.9715636      10 -0.1458338
## 5      Volti     HC -0.6753721 -0.6874525 0.1161715      11  0.5771187
## 6      Volti     LE -0.7860698 -0.7966665 0.4335148      10  0.3698441
##   Phonemic.me Phonemic.v Phonemic.n
## 1   0.7429972 0.02538157         11
## 2   0.3645354 0.18551385         10
## 3   0.6927294 0.02552888         11
## 4   0.2141080 1.89813504         10
## 5   0.6005389 0.01898461         11
## 6   0.4900615 0.08191612         10
```

More attractive graphs with GGplot2 package

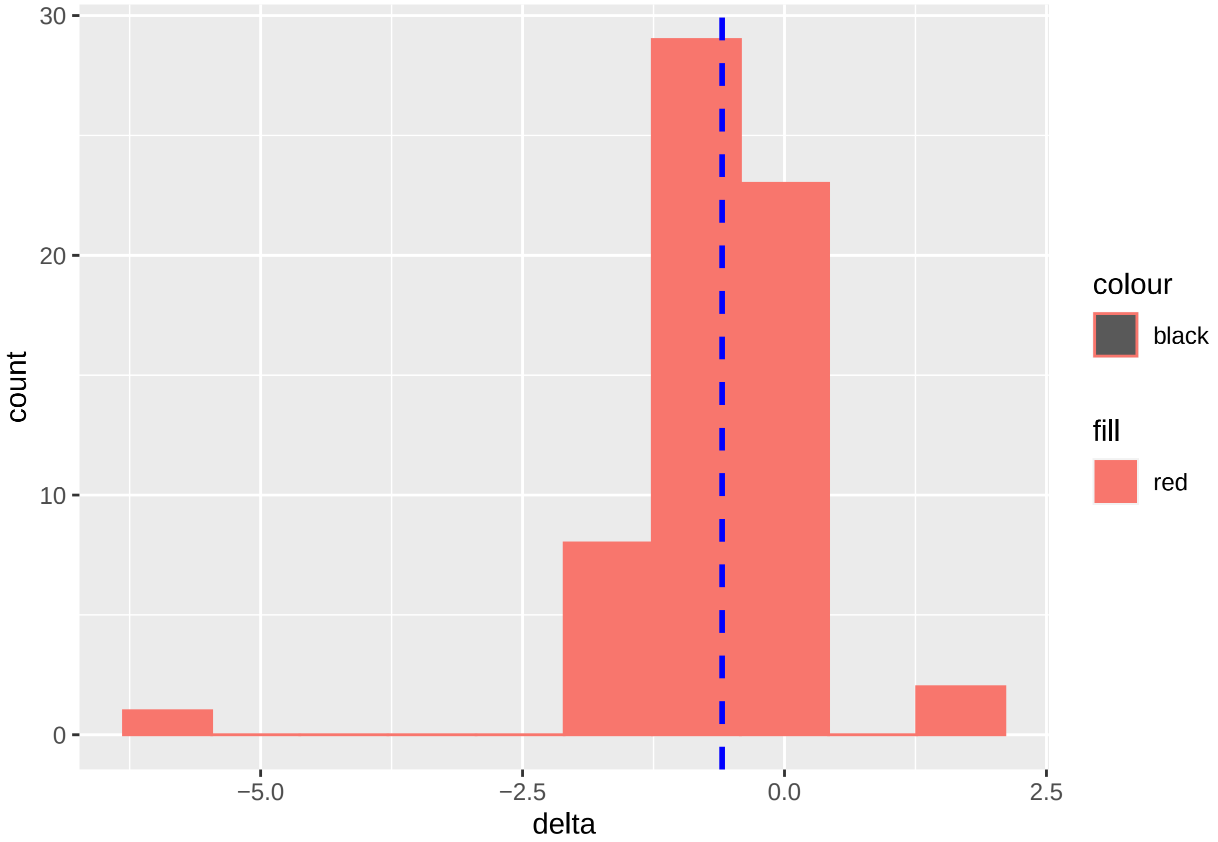
This package (GGplot2) offers to us the possibility to create elegant data visualisations. Please visit:

https://ggplot2.tidyverse.org/

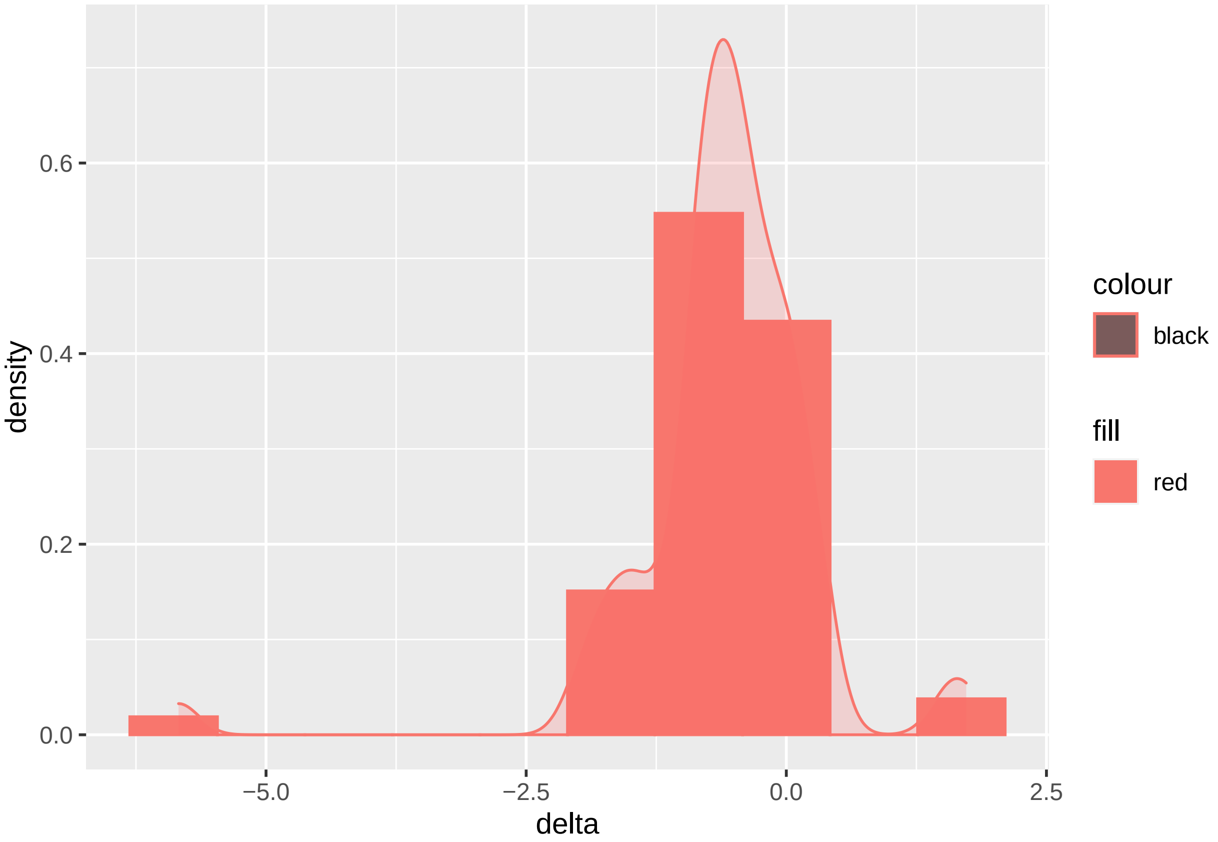
```r
# if not installed, digit install.packages("ggplot2")
library(ggplot2)
# an instogram
gg=ggplot(DATASET, aes(x=delta,color="black", fill="red")) +
  geom_histogram(bins=10)
gg
```



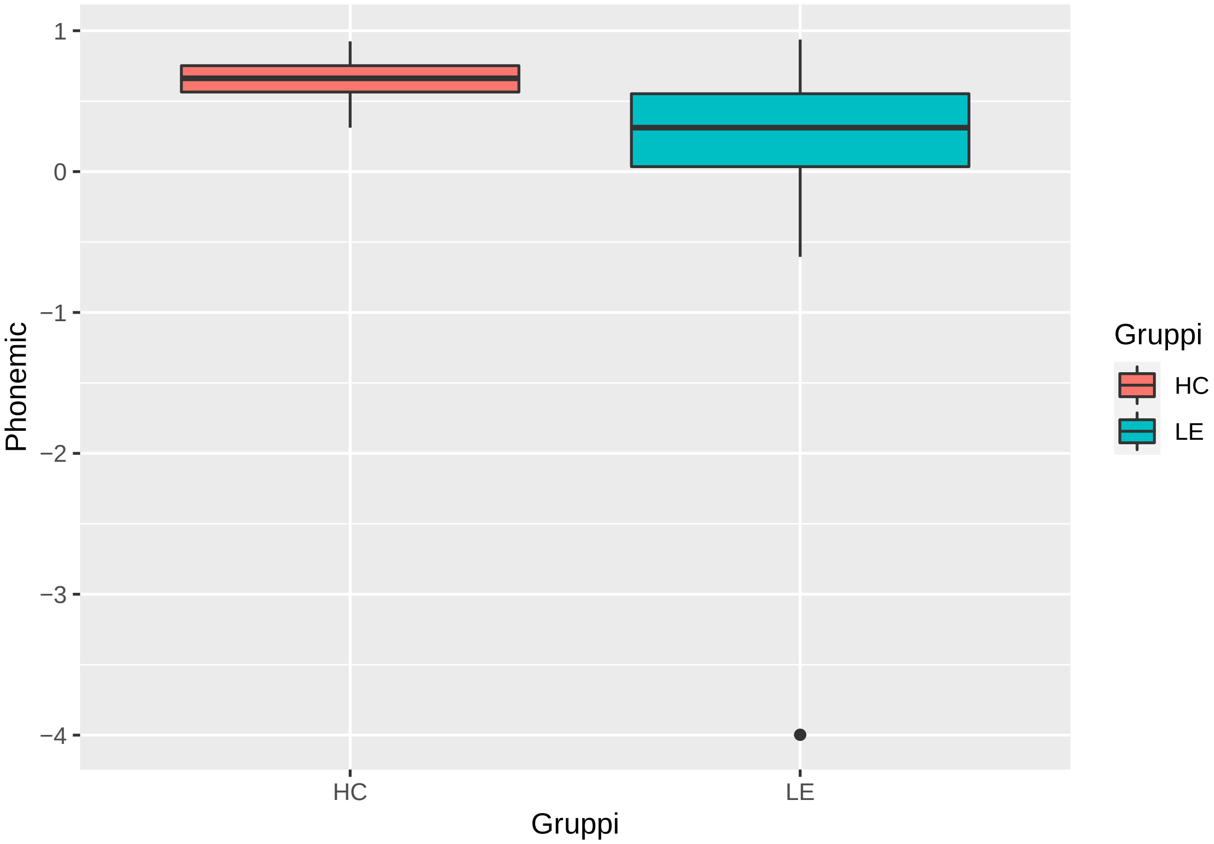
```r
# we add a mean line
gg=gg+ geom_vline(aes(xintercept=mean(delta)),
                  color="blue", linetype="dashed", size=1)
gg
```



```
# we add a smoothed density line
gg=ggplot(DATASET, aes(x=delta,y=..density..,color="black", fill="red")) +
  geom_histogram(bins=10)+
  geom_density(alpha=.2, fill="#FF6666")
gg
```

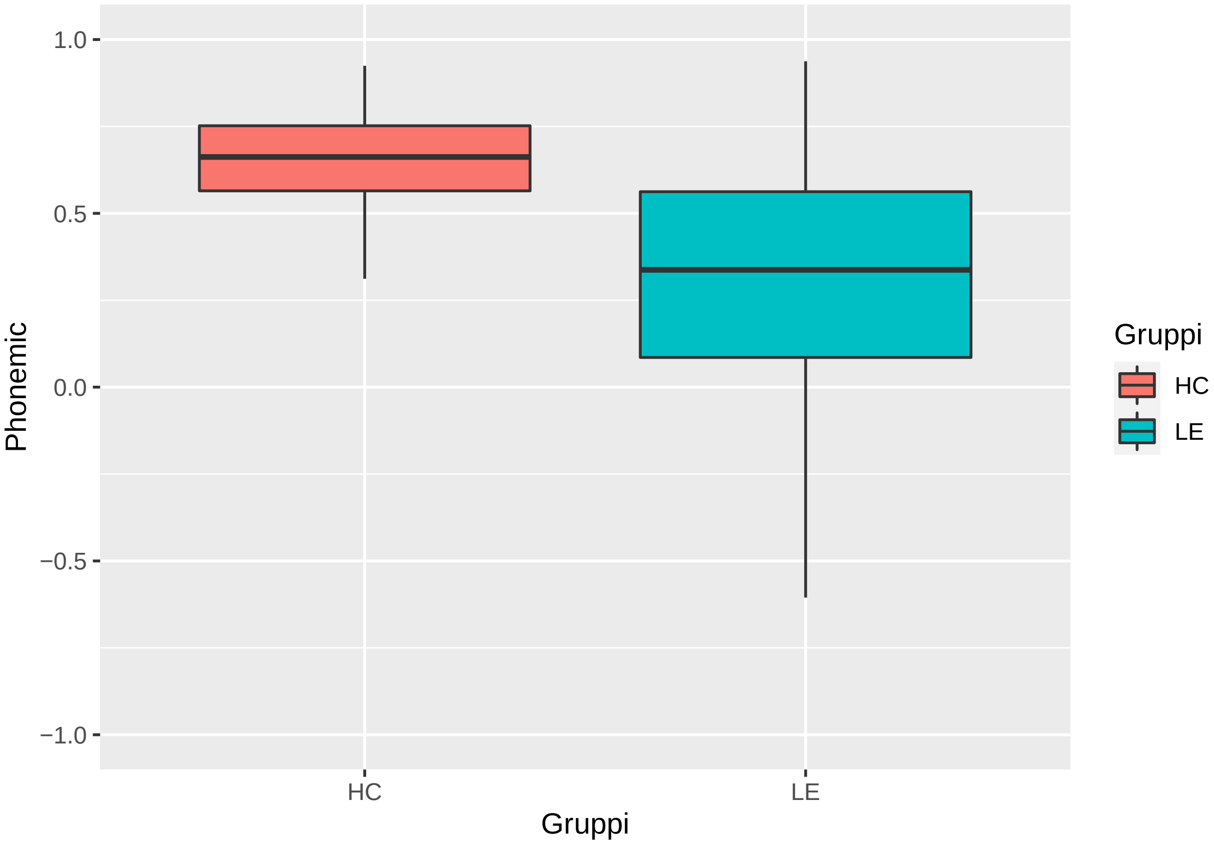


```
# a boxplot
p <- ggplot(DATASET, aes(x=Gruppi, y=Phonemic,fill=Gruppi)) +
  geom_boxplot()
p
```



```
# change y-axis limits
p <- ggplot(DATASET, aes(x=Gruppi, y=Phonemic,fill=Gruppi)) +
  geom_boxplot()+ylim(c(-1,1))
p
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```


1.0
0.5
0.0
−0.5
−1.0
Phonemic
HC
LE
Gruppi
Gruppi
HC
LE

How to perform a statistical test in R

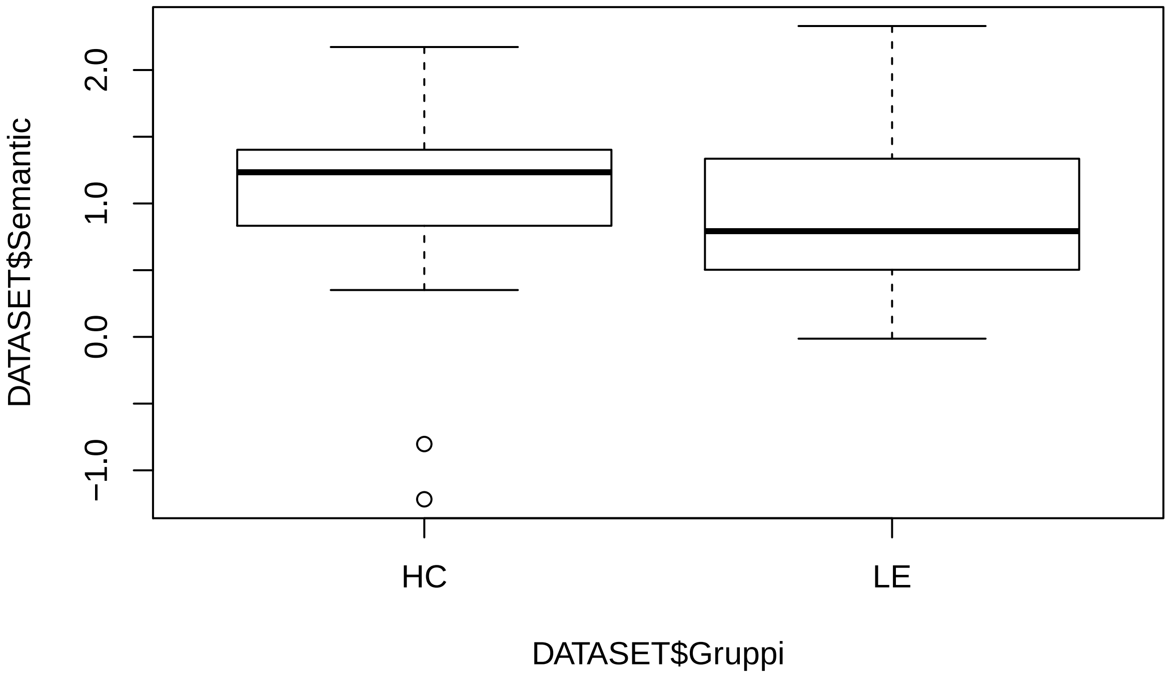
The type of the required test depends (mainly…) on:
- the type of selected variable
- the statistical assumptions made
- the distribution of the variabile
- the type of statistics chosen
- ….

In our dataset we can do several testing hyphotesis…
… express your research hypothesis!

A research hypothesis: "Is the distribution of the semantic score different among the two groups?"

With R, try to explain how a test can be performed to verify that research hypothesis.

```
boxplot(DATASET$Semantic~DATASET$Gruppi)
```



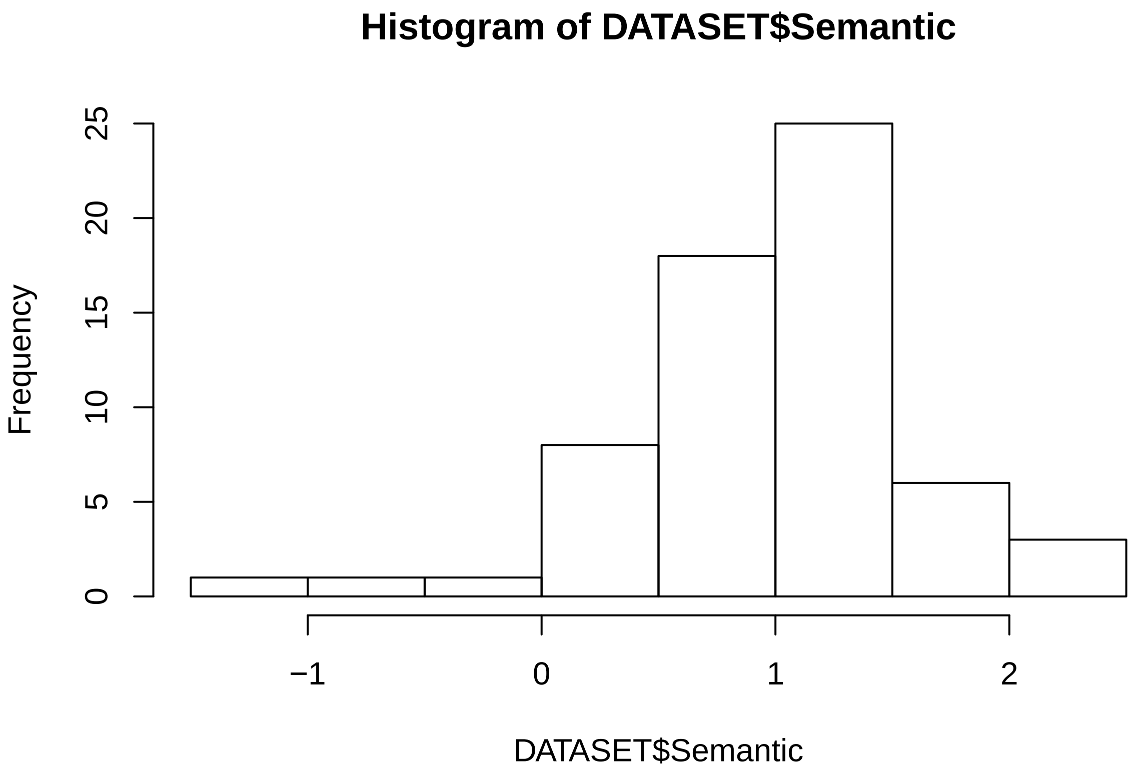
```
#There is a difference
summaryBy(Semantic~Gruppi,DATASET,FUN=summary)
```

```
##   Gruppi Semantic.Min. Semantic.1st Qu. Semantic.Median Semantic.Mean
## 1     HC   -1.21696596        0.8328461       1.2340554     1.0633632
## 2     LE   -0.01318654        0.5154333       0.7923201     0.9556955
##   Semantic.3rd Qu. Semantic.Max.
## 1         1.402468      2.172417
## 2         1.333809      2.329780
```

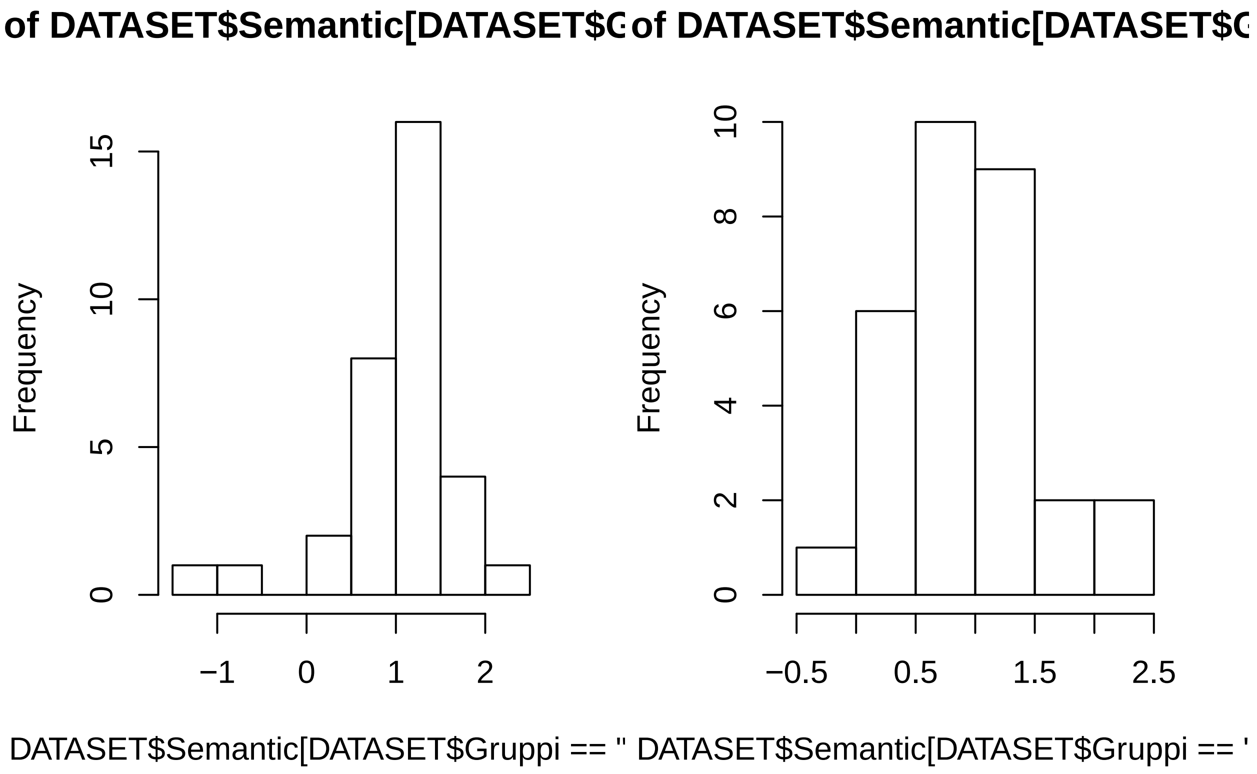
```
#a difference median of 0.44, mean 0.11
```

We are interested to verify if the mean of the semantic scores is equal or different in the two groups. First of all… is the distribution of the semantic score normally distributed?

```
hist(DATASET$Semantic)
```



```
par(mfrow=c(1,2))
hist(DATASET$Semantic[DATASET$Gruppi=="HC"])
hist(DATASET$Semantic[DATASET$Gruppi=="LE"])
```



```
par(mfrow=c(1,1))

#....to verify if the Semantic is normally distributed
################### we use a Shapiro Test.... shapiro.test()
# H0 Y~N(m,s2)
shapiro.test(DATASET$Semantic)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  DATASET$Semantic
## W = 0.9491, p-value = 0.01118
```

```
#p-value <0.05, but 0.01.... with alpha=0.05 I reject the Hyphotesis of normally distibution, but...
shapiro.test(DATASET$Semantic[DATASET$Gruppi=="HC"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  DATASET$Semantic[DATASET$Gruppi == "HC"]
## W = 0.83214, p-value = 0.0001411
```

```
shapiro.test(DATASET$Semantic[DATASET$Gruppi=="LE"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  DATASET$Semantic[DATASET$Gruppi == "LE"]
## W = 0.96101, p-value = 0.3286
```

```
#quite different in the two sub-samples
```

Parametric test - T of Student test

If the normal distribution can be assumed we choice to perform a T of Student test. Assumptions: - $Y_i \sim$ N(μ_{HC},σ^2_{HC}) for $i = 1, ..., 33$;
- $Y_i \sim$ N(μ_{LE},σ^2_{LE}) for $i = 34, ..., 63$;
- for each $i \neq j$, Y_i and Y_j are uncorrelated (hmmm this is may be a problem, the student can try to say why…):

H_0: $\mu_{HC} = \mu_{LE}$
H_A: $\mu_{HC} \neq \mu_{LE}$

```
#Is the variance of the two groups the same?
var(DATASET$Semantic[DATASET$Gruppi=="HC"])
```

```
## [1] 0.4392154
```

```
var(DATASET$Semantic[DATASET$Gruppi=="LE"])
```

```
## [1] 0.3498356
```

```
#similar, there is also a test for that
var.test(DATASET$Semantic~DATASET$Gruppi)
```

```
##
##  F test to compare two variances
##
## data:  DATASET$Semantic by DATASET$Gruppi
## F = 1.2555, num df = 32, denom df = 29, p-value = 0.5388
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6046372 2.5732847
## sample estimates:
## ratio of variances
##           1.255491
```

```r
# p>>>>0.05
t.test(DATASET$Semantic~DATASET$Gruppi,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  DATASET$Semantic by DATASET$Gruppi
## t = 0.67762, df = 61, p-value = 0.5006
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2100525  0.4253879
## sample estimates:
## mean in group HC mean in group LE
##        1.0633632        0.9556955
```

```r
###### T.TEST with equal variance between the two groups
# I accept that the mean of the semantic score is the same between the groups
```

Non parametric test - Mann-Whitney test

If the normal distribution can NOT be assumed we may choice to perform a Mann-Whitney test.

H_0: $Y_{HC} = Y_{LE}$
H_A: $Y_{HC} \neq Y_{LE}$

where Y_{HC} and Y_{LE} are the "unknown" distributions of the semantic score in HC and LE group.

```r
wilcox.test(DATASET$Semantic~DATASET$Gruppi)
```

```
##
##  Wilcoxon rank sum test
##
## data:  DATASET$Semantic by DATASET$Gruppi
## W = 601, p-value = 0.1474
## alternative hypothesis: true location shift is not equal to 0
```

```r
# As before... I do not reject H_0
```