Linguistica Applicata

*Esercitazione di R per linguisti*

(A/A 2020-21)

## Appello autunnale del 7 settembre 2021

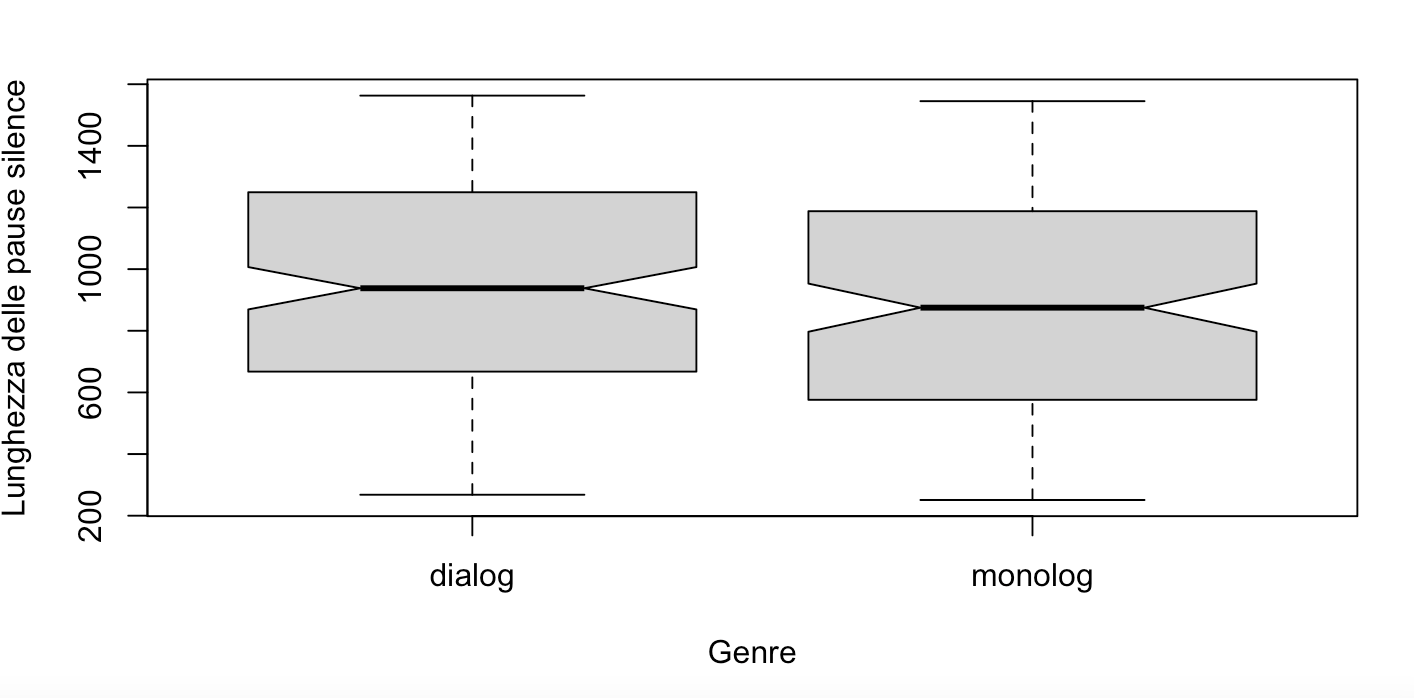
**Esercizio 1**

Caricate il file *esercizio\_1.txt*, che contiene vari dati riguardanti la lunghezza di vari tipi di pause (FILLER) in testi appartenenti a generi diversi di parlato (GENRE):

df <- read.table("esercizio\_1.txt", header = T, row.names = 1)

* rappresentate graficamente la distribuzione della lunghezza delle pause di tipo “silence” rispetto ai generi di parlato (GENRE)

boxplot(df$LENGTH[df$FILLER == "silence"] ~ df$GENRE[df$FILLER == "silence"], notch=T, xlab = "Genre", ylab = "Lunghezza delle pause silence")



* calcolate i valori di media, deviazione standard e range interquartile della lunghezza delle pause di tipo “silence” per i diversi generi di parlato

dialog <- df$LENGTH[df$GENRE == "dialog" & df$FILLER == "silence"]

monolog <- df$LENGTH[df$GENRE == "monolog" & df$FILLER == "silence"]

mean(dialog)

[1] 943.2849

sd(dialog)

[1] 353.511

IQR(dialog)

[1] 582

mean(monolog)

[1] 887.9281

sd(monolog)

[1] 380.1689

IQR(monolog)

[1] 612

* trasformate la lunghezza delle pause di tipo “silence” in z-score;

z.score <- (df[df$FILLER == "silence", "LENGTH"]/mean(df[df$FILLER == "silence", "LENGTH"]))/sd(df[df$FILLER == "silence", "LENGTH"])

* dimostrate se esiste una differenza statisticamente nella lunghezza delle pause di tipo “silence” nei diversi generi di parlato

Prima verifico se sono normalmente distribuite:

**H0:** La distribuzione del campione è normale.

**p-value > 0.05 Normale | p-value < 0.05 Non normale**

shapiro.test(monolog)

Shapiro-Wilk normality test

data: monolog

W = 0.95551, p-value = 8.176e-05

Non è normalmente distribuita

shapiro.test(dialog)

Shapiro-Wilk normality test

data: dialog

W = 0.96669, p-value = 0.0002825

Non è normalmente distribuita

Uso Test U di Wilcoxon

*H0: La differenza nelle lunghezze delle pause dei due campioni non è significativa.*

*p-value > 0.05 Non significativa | p-value < 0.05 Significativa*

wilcox.test(monolog,dialog)

Wilcoxon rank sum test with continuity correction

data: monolog and dialog

W = 12524, p-value = 0.1799

alternative hypothesis: true location shift is not equal to 0

Dal momento che il valore del p-value è maggiore di 0,05, la differenza nella lunghezza delle pause di tipo “silence” tra i due campioni non è significativa

**Esercizio 2**

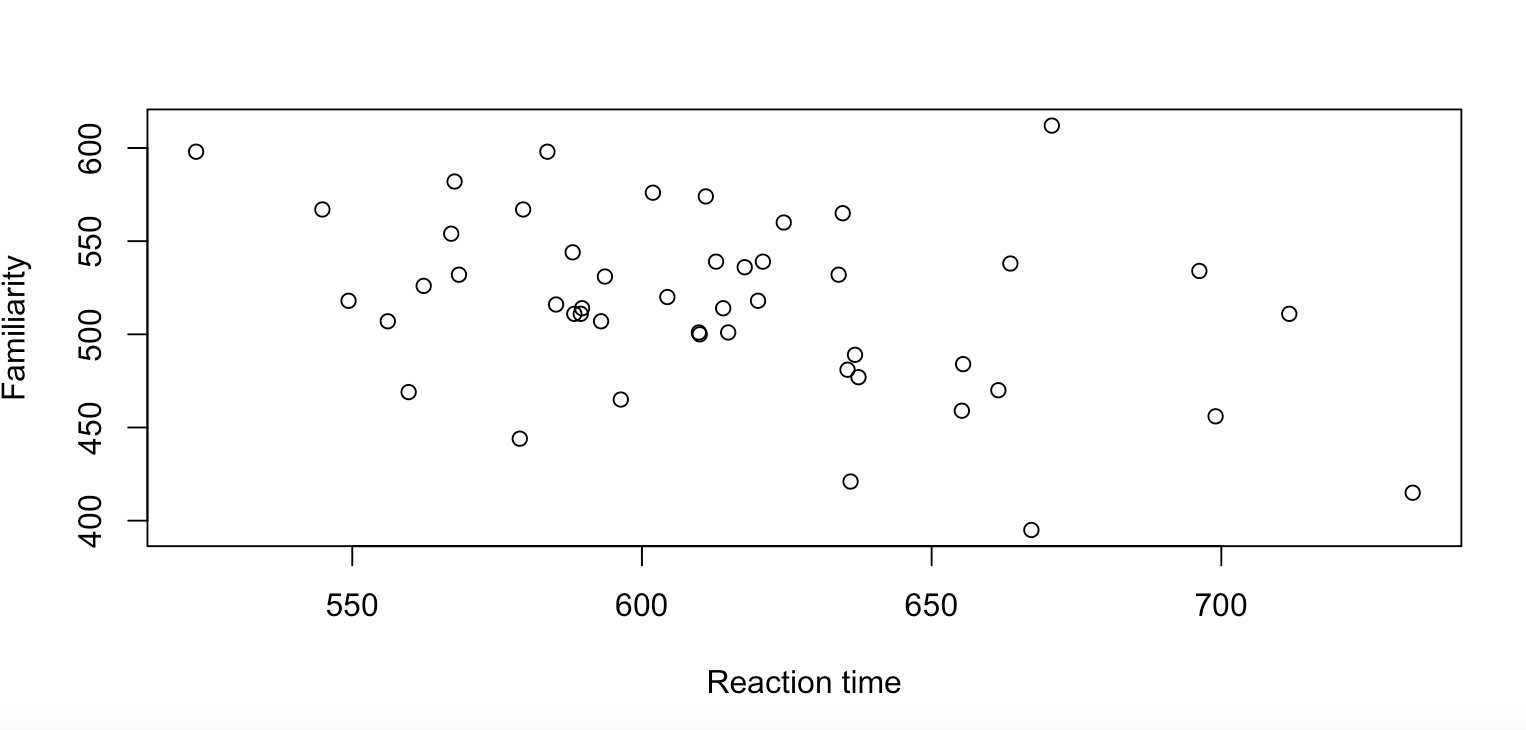
Caricate il file *esercizio\_2.txt*, che contiene i tempi di reazioni per un insieme di parole inglesi:

df1 <- read.table("esercizio\_2.txt", header = T)

* rappresentate graficamente la distribuzione delle due variabili REACTTIME e FAMILIARITY;

Sono entrambe numeriche, quindi:

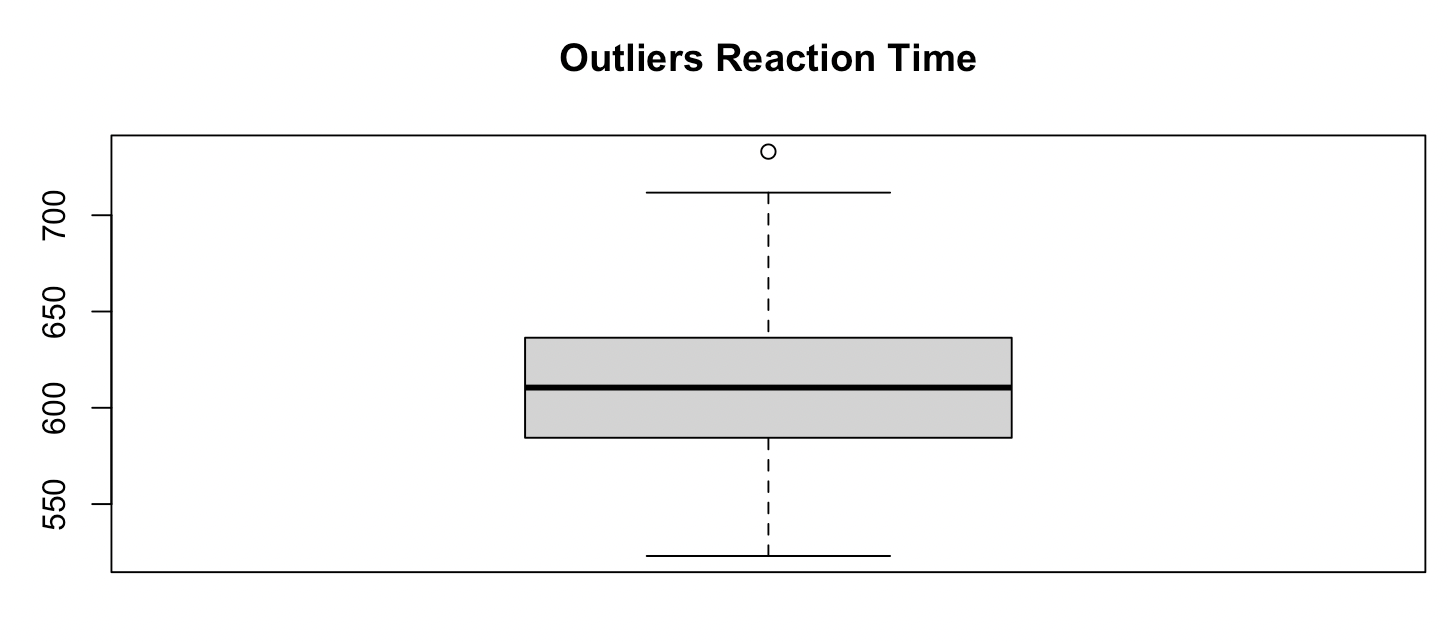
plot(df1$REACTTIME, df1$FAMILIARITY, xlab = "Reaction time", ylab = "Familiarity")



* indicate quali sono i valori outliers, se esistono, delle due variabili;

outliers reaction time:

boxplot(df1$REACTTIME, main = "Outliers Reaction Time")



boxplot.stats(df1$REACTTIME)

$stats

[1] 523.0493 584.4307 610.4993 636.3922 711.7317

$n

[1] 48

$conf

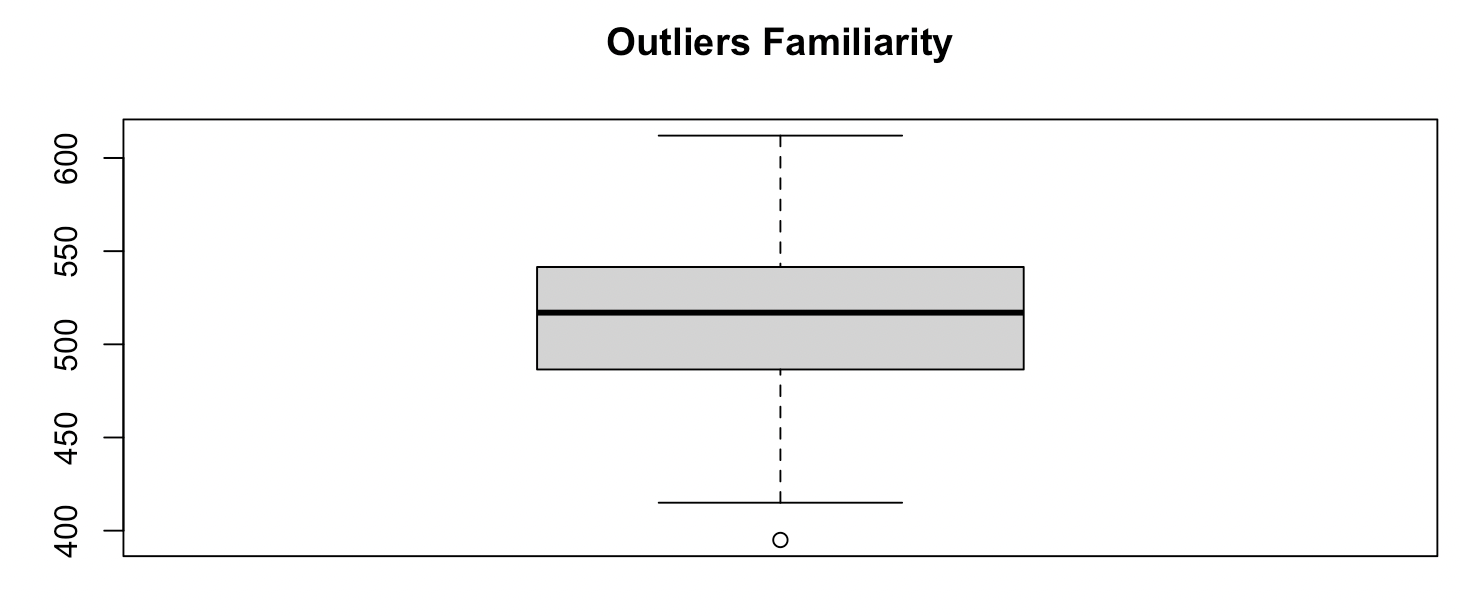
[1] 598.6492 622.3493

**$out**

**[1] 733.0323 -> è il valore dell’outlier**

Outliers familiarity:

boxplot(df1$FAMILIARITY, main = "Outliers Familiarity")



boxplot.stats(df1$FAMILIARITY)

$stats

[1] 415.0 486.5 517.0 541.5 612.0

$n

[1] 48

$conf

[1] 504.4571 529.5429

**$out**

**[1] 395 -> è il valore dell’outlier**

* dimostrate se le due variabili sono distribuite normalmente oppure no;

**H0:** La distribuzione del campione è normale.

**p-value > 0.05 Normale | p-value < 0.05 Non normale**

Distribuzione REACTTIME:

shapiro.test(df1$REACTTIME)

Shapiro-Wilk normality test

data: df1$REACTTIME

W = 0.97542, p-value = 0.4052

Valore di p-value maggiore di 0,05, quindi la distribuzione è normale

Distribuzione FAMILIARITY:

shapiro.test(df1$FAMILIARITY)

Shapiro-Wilk normality test

data: df1$FAMILIARITY

W = 0.98283, p-value = 0.6995

Valore di p-value maggior di 0,05, quindi la distribuzione è normale

* calcolate l’indice di correlazione e il coefficiente di determinazione per queste due variabili;

Dal momento che le due variabili sono normalmente distribuite, uso il metodo di Pearson per calcolare l’indice di correlazione:

cor(df1$REACTTIME, df1$FAMILIARITY, method = "pearson")

[1] -0.400589

Coefficiente di determinazione:

model<- lm(df1$REACTTIME ~ df1$FAMILIARITY)

summary(model)["r.squared"]

$r.squared

[1] 0.1604716

* costruite la retta di regressione.

Riprendendo la variabile model del punto precedente:

plot(df1$FAMILIARITY, df1$REACTTIME, xlab= "Familiarity", ylab = "Reaction Time")

abline(model, col = "red")

