Riassunto Statistica

# Interpretazione del p-value:

Ogni test ha un’ipotesi H0 e un’ipotesi H1.  
Se il p-value è MINORE di 0.05, H0 è RIFIUTATA, e quindi dobbiamo assumere che l’ipotesi H1 sia vera.  
Se il p-value è MAGGIORE di 0.05, H0 NON è RIFIUTATA, e quindi dobbiamo assumere che l’ipotesi H0 sia vera.

## Descrizione delle variabili:

**VARIABILE QUANTITATIVA--------------**

summary() → Riassunto in 5 numeri.

Str() → Struttura dei dati

Mean()

Dim() → Quanto sono grandi i due set di dati?

Sd() e Coefficiente di variazione (VCof, sd/mean) → deviazione standard, coefficiente di variazione.

Possiamo anche eventualmente calcolare il range con range(), la mediana con median().

Grafici: hist(), boxplot()

Verifica della normalità: shapiro.test(), ks.test(), qqplot(), qqline()

**VARIABILE QUALITATIVA----------------**

Str() → Struttura dei dati.

Dim() → Quante osservazioni abbiamo?

Table() → Crea una tabella delle frequenze

Grafici: barplot(), pie()

## TEST per le statistiche

* Test per la Normalità:
  + shapiro.test, ks.test
  + H0: La distribuzione è normale (o è uguale all’altra distribuzione per ks)
  + H1: La distribuzione non è normale (o non è uguale all’altra)
* Test sulle Medie:
  + t.test, wilcox.test, (z.test)
  + H0: Le medie sono uguali (o meglio, la sottrazione delle due medie è uguale a 0), oppure la media è 0
  + H1: Le medie non sono uguali, non è uguale a 0, oppure è maggiore o minore dell’altra o di 0.
* Test sulle Proporzioni:
  + chisq.test o z.test, fisher.test
  + H0: Le due proporzioni sono uguali, o la proporzione è uguale a 0.
  + H1: Le due medie sono diverse, diverse da 0, maggiore / minore tra loro, o maggiore / minore di 0
* Test sulle SD:
  + nessuno (chisq.test?)
  + H1: ??
  + H2: ??
* Test sulle varianze:
  + Bartlett.test(), vuole la tilde!
  + H0: Le varianze sono uguali.
  + H1: Le varianze sono diverse.

## Correlazione Lineare

* cor.test(), fa il test di correlazione lineare. Più il valore di R si avvicina a 1, meglio è il modello (<50 è brutto, da 50 a 70 è normale, più di 70 è bello, più di 90 è ottimo.)
  + H1: Non c’è una correlazione tra le due variabili.
  + H0: C’è correlazione tra variabili.
* lm(), fà il modello lineare. vuole la tilde! Calcola la retta migliore, il valore di R, e altre cose. Si mette lm() dentro una variabile di solito, e lo si legge con summary(lm())
  + Possiamo aggiungere -1 dopo le variabili, tipo lm(var.1 ~ var.2 -1) per forzare l’intercetta a 0. Il modello potrebbe migliorare o anche no.
* resid(), prendo un modello lineare e calcola i residui. Qqnorm(resid()) dovrebbe far vedere residui distribuiti normalmente.
  + Si può fare un shapiro.test dei residui, ma se i residui sono piccoli bisogna fare un po' di attenzione.
* plot(fitted(), resid()) fà un grafico di dispersione tra i valori aspettati dal modello e i residui. I punto dovrebbero sembrare completamente casuali.
* Se serve, abline(MODELLO\_LINEARE) mette la linea del modello dentro altri plots.

## ANOVA

* bartlett.test, Per poter usare l’anova, le varianze devono essere uguali (vedi sopra)
* shapiro.test, Per poter utilizzare l’anova, tutti i campioni devono essere normali
* anova(), fà il test anova. vuole la tilde!
  + H0: Tutte le medie sono uguali.
  + H1: Almeno una media è diversa.
* pairwise.t.test(adj="bonferroni"),
  + Quale media è diversa? Dobbiamo fare tanti t.test, ma dobbiamo aggiustare il p-value con Bonferroni.
* kruskal.test, vuole la tilde!
  + Se non possiamo usare l’anova, possiamo usare il test di Kruskal-Wallis.
  + H0: Le MEDIANE sono uguali.
  + H1: Almeno una mediana non è uguale.
* tapply(x, y, median, na.rm=T)
  + Per fare una cosa simile al pairwise.t.test possiamo usare questo comando.