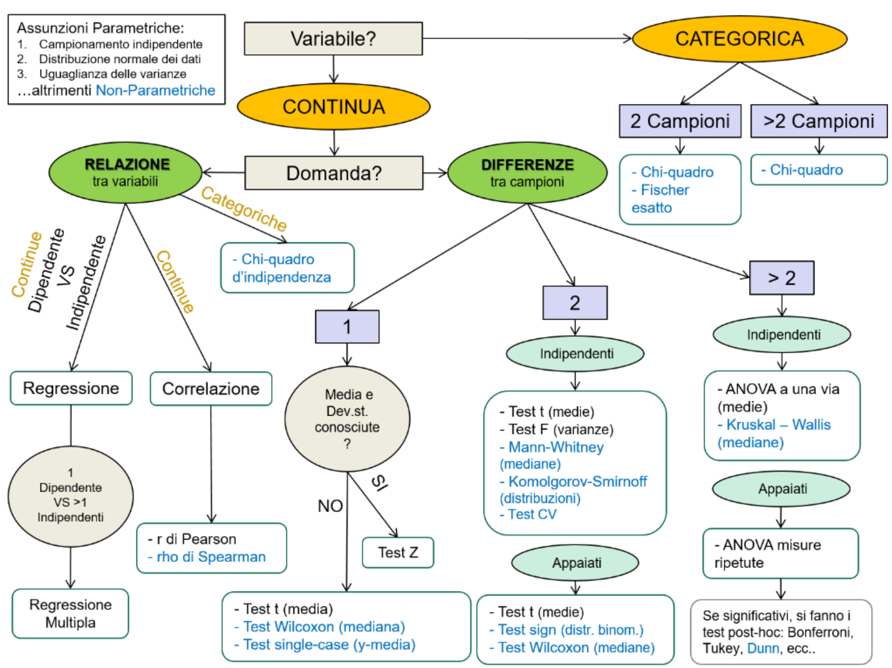
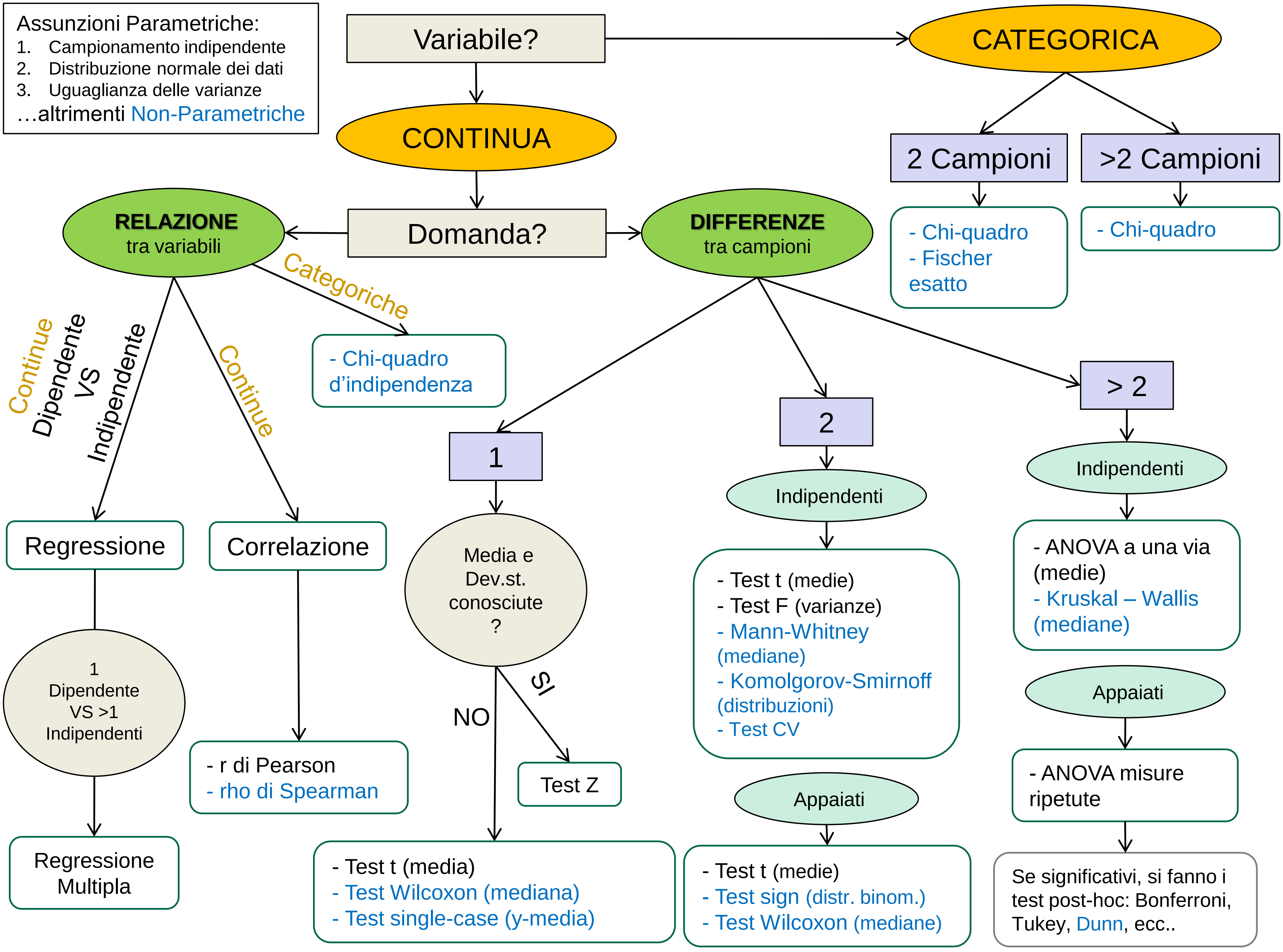
Corso di Laurea di Scienze Forestali ed Ambientali

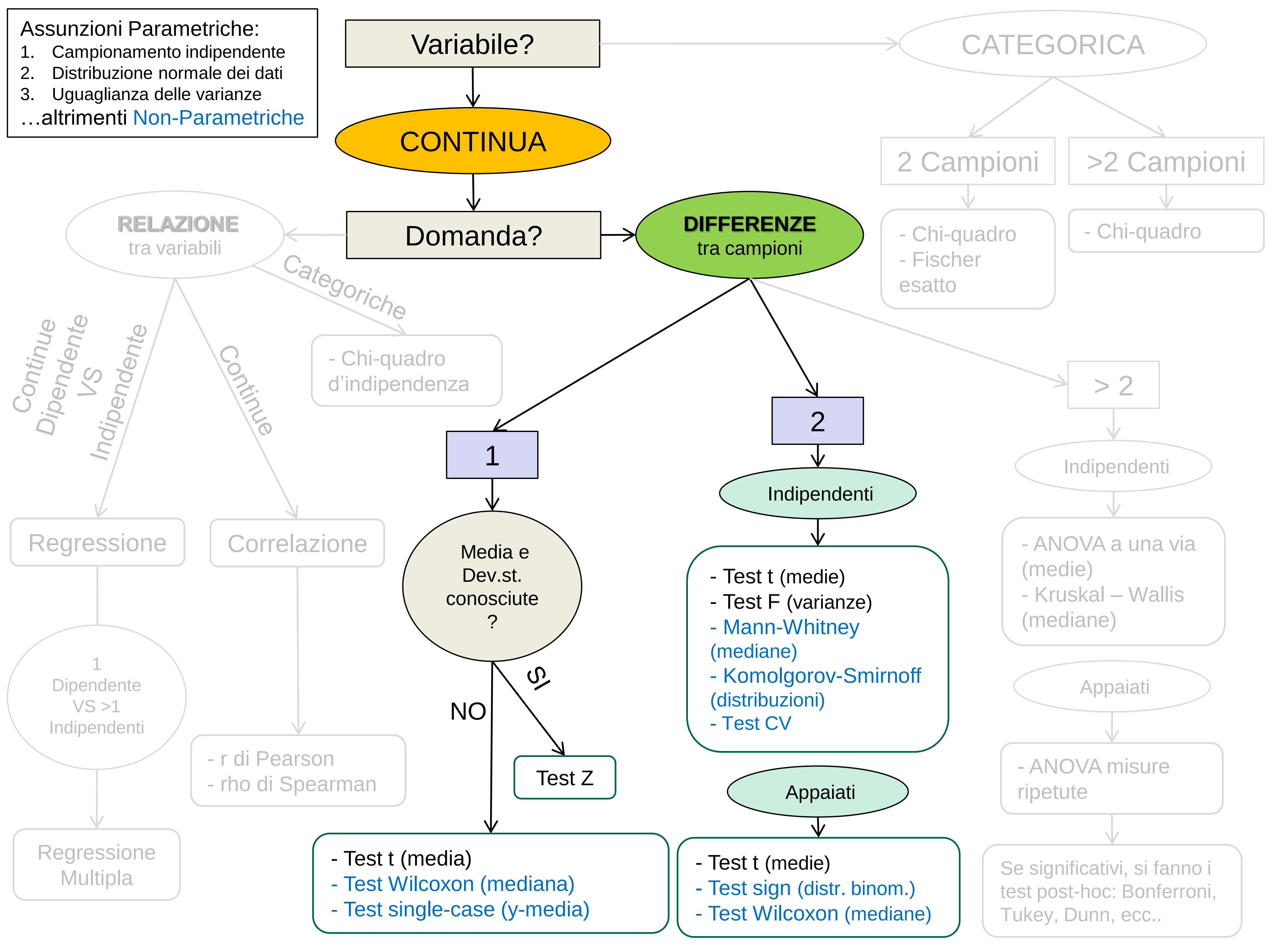
Ecologia e Statistica per l'ambiente

a.a. 2020-2021 - MATTEO GARBARINO - matteo.garbarino@unito.it

TEST 1 – 2 CAMPIONI

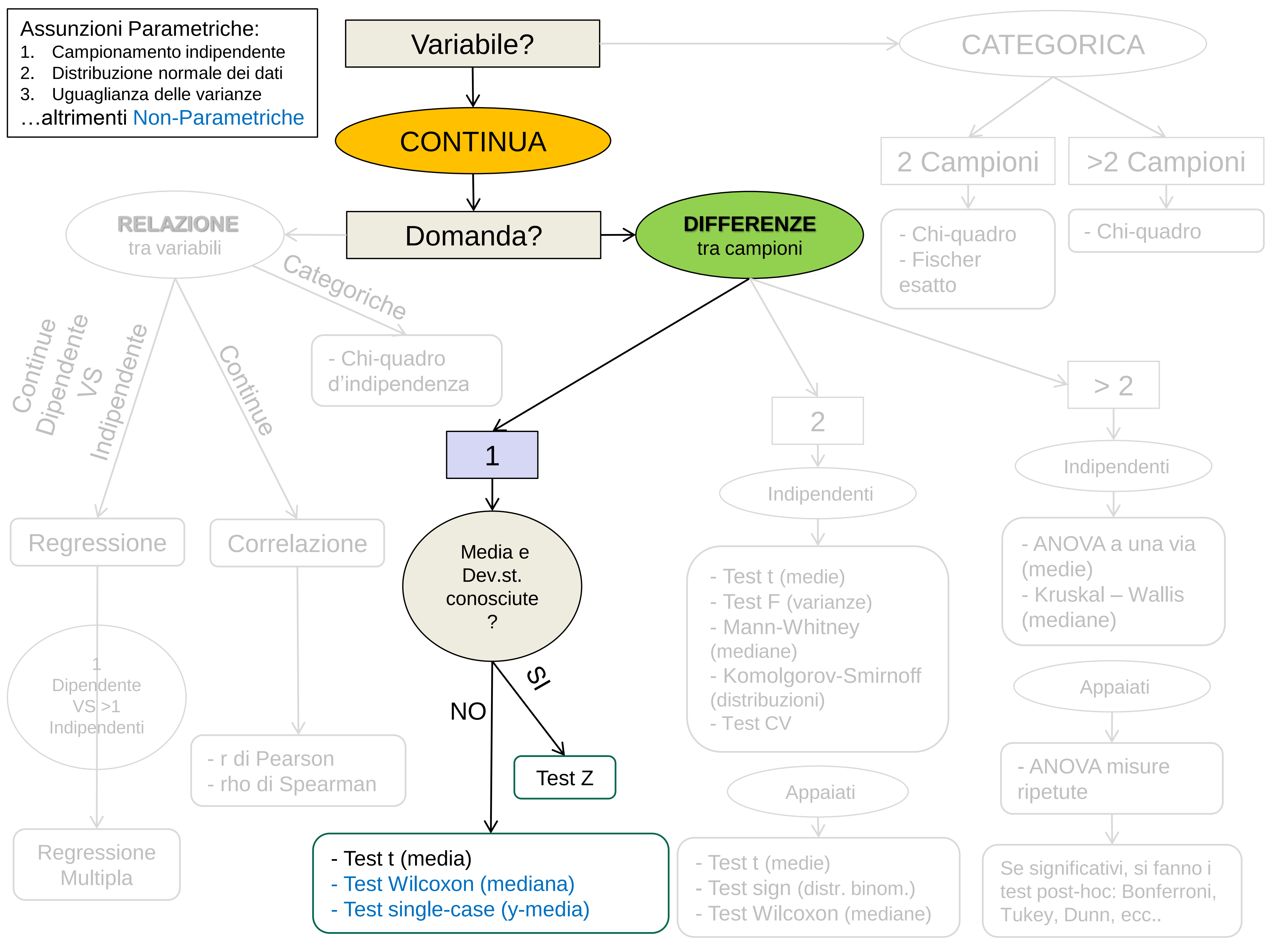



Assunzioni Parametriche:
1. Campionamento indipendente
2. Distribuzione normale dei dati
3. Uguaglianza delle varianze
…altrimenti Non-Parametriche
Variabile?
CATEGORICA
CONTINUA
2 Campioni
>2 Campioni
- Chi-quadro
- Fischer esatto
- Chi-quadro
Domanda?
RELAZIONE tra variabili
DIFFERENZE tra campioni
Categoriche
- Chi-quadro d'indipendenza
Continue
Dipendente VS Indipendente
Continue
Regressione
Correlazione
1 Dipendente VS >1 Indipendenti
Regressione Multipla
- r di Pearson
- rho di Spearman
1
Media e Dev.st. conosciute ?
SI
NO
Test Z
- Test t (media)
- Test Wilcoxon (mediana)
- Test single-case (y-media)
2
Indipendenti
- Test t (medie)
- Test F (varianze)
- Mann-Whitney (mediane)
- Komolgorov-Smirnoff (distribuzioni)
- Test CV
Appaiati
- Test t (medie)
- Test sign (distr. binom.)
- Test Wilcoxon (mediane)
> 2
Indipendenti
- ANOVA a una via (medie)
- Kruskal – Wallis (mediane)
Appaiati
- ANOVA misure ripetute
Se significativi, si fanno i test post-hoc: Bonferroni, Tukey, Dunn, ecc..


Assunzioni Parametriche:
1. Campionamento indipendente
2. Distribuzione normale dei dati
3. Uguaglianza delle varianze
…altrimenti Non-Parametriche
Variabile?
CATEGORICA
CONTINUA
2 Campioni
>2 Campioni
- Chi-quadro
- Fischer esatto
- Chi-quadro
Domanda?
RELAZIONE
tra variabili
DIFFERENZE
tra campioni
Categoriche
- Chi-quadro d'indipendenza
Continue Dipendente VS Indipendente
Continue
1
2
> 2
Indipendenti
Indipendenti
Regressione
Correlazione
Media e Dev.st. conosciute ?
- Test t (medie)
- Test F (varianze)
- Mann-Whitney (mediane)
- Komolgorov-Smirnoff (distribuzioni)
- Test CV
- ANOVA a una via (medie)
- Kruskal – Wallis (mediane)
1 Dipendente VS >1 Indipendenti
SI
NO
Test Z
Appaiati
Appaiati
- r di Pearson
- rho di Spearman
- ANOVA misure ripetute
Regressione Multipla
- Test t (media)
- Test Wilcoxon (mediana)
- Test single-case (y-media)
- Test t (medie)
- Test sign (distr. binom.)
- Test Wilcoxon (mediane)
Se significativi, si fanno i test post-hoc: Bonferroni, Tukey, Dunn, ecc..

Test per dati continui

Test tra 1 campione e una distribuzione teorica o popolazione (STIMA)


Assunzioni Parametriche:
1. Campionamento indipendente
2. Distribuzione normale dei dati
3. Uguaglianza delle varianze
…altrimenti Non-Parametriche
Variabile?
CATEGORICA
CONTINUA
2 Campioni
>2 Campioni
- Chi-quadro
- Fischer esatto
- Chi-quadro
Domanda?
RELAZIONE
tra variabili
DIFFERENZE
tra campioni
Categoriche
- Chi-quadro d'indipendenza
Continue Dipendente VS Indipendente
Continue
1
2
> 2
Indipendenti
Indipendenti
Regressione
Correlazione
Media e Dev.st. conosciute ?
- Test t (medie)
- Test F (varianze)
- Mann-Whitney (mediane)
- Komolgorov-Smirnoff (distribuzioni)
- Test CV
- ANOVA a una via (medie)
- Kruskal – Wallis (mediane)
1 Dipendente VS >1 Indipendenti
SI
NO
Appaiati
Test Z
- r di Pearson
- rho di Spearman
- ANOVA misure ripetute
Appaiati
Regressione Multipla
- Test t (media)
- Test Wilcoxon (mediana)
- Test single-case (y-media)
- Test t (medie)
- Test sign (distr. binom.)
- Test Wilcoxon (mediane)
Se significativi, si fanno i test post-hoc: Bonferroni, Tukey, Dunn, ecc..

Test Z

Se conosco media e deviazione standard della popolazione, allora utilizzo il test Z:

H_0: la media campionaria non è significativamente diversa dalla media della popolazione
H_1: la media campionaria è significativamente diversa dalla media della popolazione

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Dove:
X = media campionaria
μ = media della popolazione
σ = deviazione standard della popolazione
n = numerosità campionaria

Se il valore calcolato di Z è inferiore in valore assoluto al valore limite di Z (dato da n e p Didtrib.Z) allora tengo H_0

Altri Test 1 campione

<u>Test tra un campione e una distribuzione teorica o popolazione (STIMA)</u>

Test il cui obiettivo è verificare se un campione di dati (una Colonna) deriva da una popolazione di cui conosco media o mediana.
Ad esempio: l'altezza media di 200 abeti rossi rilevati in Valtournanche è la stessa di quella media dell'intera Valle d'Aosta?

Ecco i principali test:

- ✓ **Test t 1 campione per una data media μ_0** (parametrico)
- ✓ **Test Wilcoxon 1 campione per una data mediana M** (non-parametrico)
- ✓ **Test "singolo-caso"**

Test t
1 campione e media popolazione μ_0

L'intervallo di confidenza al 95% per la differenza tra le medie è basato sull'errore standard della media campionaria e la distribuzione *t*. Il test assume che la distribuzione del campione sia normale. Se *s* è la stima della deviazione standard del campione, l'intervallo di confidenza è:

$$\left[\left|\bar{x} - \mu_0\right| - t_{(\alpha/2, n-1)} \frac{s}{\sqrt{n}}, \quad \left|\bar{x} - \mu_0\right| + t_{(\alpha/2, n-1)} \frac{s}{\sqrt{n}} \right]$$

Dove, *t* ha *n*-1 gradi di libertà e 1-*α* = 0.95 per un intervallo di confidenza del 95%.

L'ipotesi nulla è: H_0: Il campione è estratto da una popolazione avente media μ_0.

Il test statistico è:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Test Single-case
1 campione e media popolazione M

L'ipotesi nulla è: H_0: Il singolo valore y è estratto dalla stessa popolazione da cui proviene il campione.

Test parametrico che assume la distribuzione normale. Di solito i softwares calcolano un semplice test z, che però risulta spesso inaccurato perché assume che media e deviazione standard (della popolazione) siano esatte. Nella realtà media e deviazione standard vengono stimate dal campione, quindi molti software calcolano una versione modificata del test t (Sokal & Rohlf 1995; Crawford & Howell 1998):

$$t = \frac{y - \bar{x}}{s\sqrt{\frac{n+1}{n}}}$$

Dove y è un valore del campione s è la deviazione standard del campione e n = numerosità campionaria.

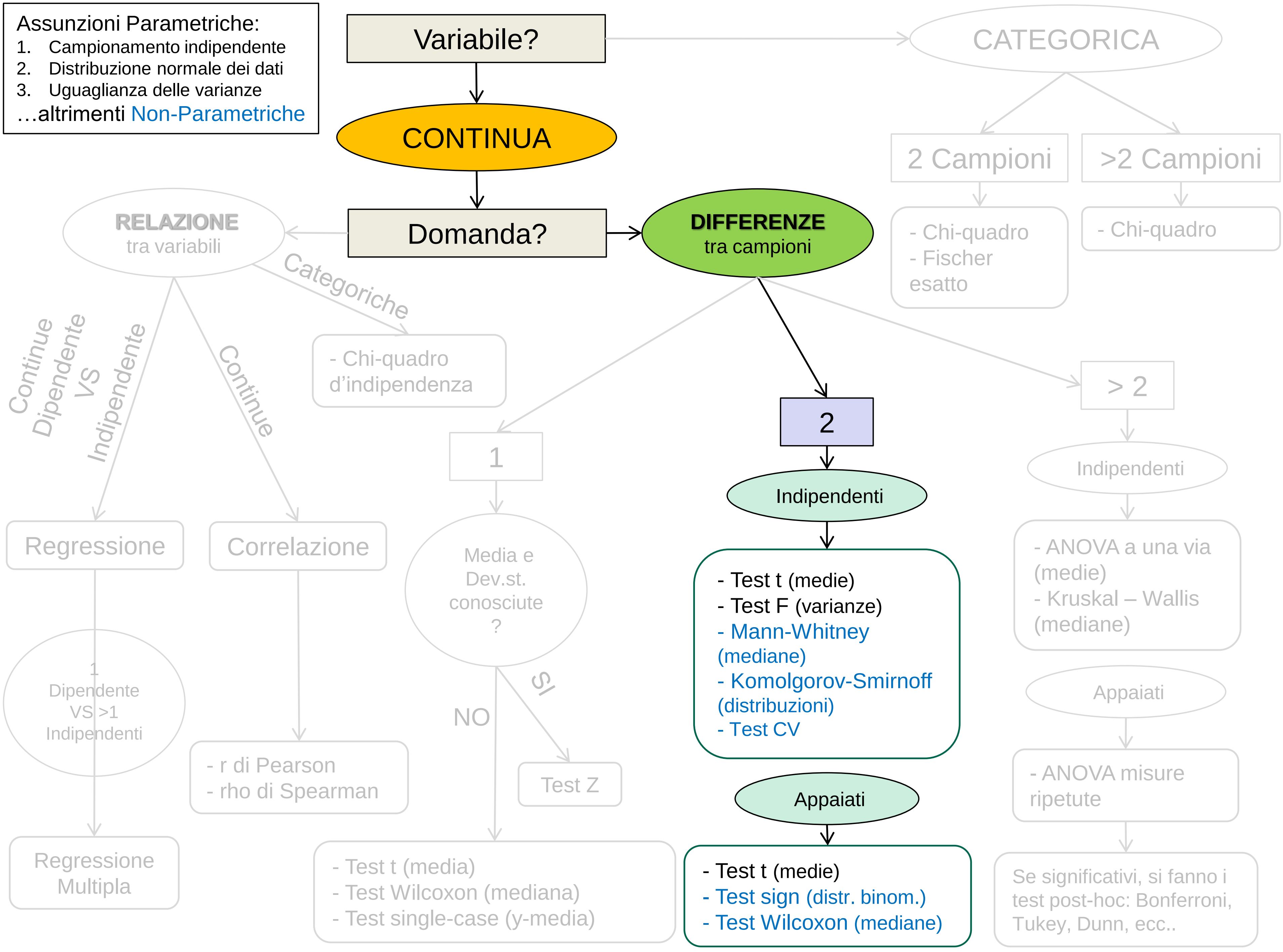
Test Wilcoxon
1 campione e mediana popolazione M

L'ipotesi nulla è:
H_0: Il campione è estratto da una popolazione avente mediana M.

Generalmente i software eliminano dal campione tutti i valori uguali ad M, poi i valori assoluti delle differenze $|d_i|$ vengono ordinati in ranghi (R_i), e la media dei ranghi assegnata per coppie. La somma dei ranghi per le coppie in cui d_i è positivo è W+. La somma dei ranghi per le coppie in cui d_i è negativo è W-.

Il test risulta dunque: $W = \max(W+, W-)$

Per $n < 10$ è possibile calcolare il valore di *p esatto*
Per $n > 10$ viene calcolato di norma un valore «approssimato» di *p*


Assunzioni Parametriche:
1. Campionamento indipendente
2. Distribuzione normale dei dati
3. Uguaglianza delle varianze
…altrimenti Non-Parametriche
Variabile?
CATEGORICA
2 Campioni
>2 Campioni
- Chi-quadro
- Fischer esatto
- Chi-quadro
CONTINUA
Domanda?
RELAZIONE tra variabili
DIFFERENZE tra campioni
Categoriche
- Chi-quadro d'indipendenza
Continue Dipendente VS Indipendente
Continue
Regressione
Correlazione
1 Dipendente VS >1 Indipendenti
Regressione Multipla
- r di Pearson
- rho di Spearman
1
Media e Dev.st. conosciute ?
NO
SI
Test Z
- Test t (media)
- Test Wilcoxon (mediana)
- Test single-case (y-media)
2
Indipendenti
- Test t (medie)
- Test F (varianze)
- Mann-Whitney (mediane)
- Komolgorov-Smirnoff (distribuzioni)
- Test CV
Appaiati
- Test t (medie)
- Test sign (distr. binom.)
- Test Wilcoxon (mediane)
> 2
Indipendenti
- ANOVA a una via (medie)
- Kruskal – Wallis (mediane)
Appaiati
- ANOVA misure ripetute
Se significativi, si fanno i test post-hoc: Bonferroni, Tukey, Dunn, ecc..

Test 2 campioni

Numerosi Test che hanno come obiettivo il confronto tra le medie di due campioni (o gruppi) (2 colonne).
Ad esempio: la distribuzione diametrica del bosco delle Navette misurata del 1980 è la stessa distribuzione che osserviamo oggi (2020)?

Ecco i principali:

- ✓ **Test t** tra medie (parametrico)
- ✓ **Test F** tra varianze (parametrico)
- ✓ **Test Mann-Whitney** tra mediane (non-parametrico)
- ✓ **Test Komolgorov-Smirnoff** tra distribuzioni (non-parametrico)

Test t
Tra 2 medie campionarie

Test parametrico che assume la distribuzione normale di entrambi i campioni e stesse varianze (esiste anche un test t per varianze differenti).

L'ipotesi nulla è: H_0: i due campioni sono estratti dalla stessa popolazione (popolazioni aventi la stessa media).

Il test statistico è:

$$t = \frac{\bar{x} - \bar{y}}{s_D}$$

Dove, S_D è l'errore standard della differenza tra le medie campionarie.

Nel caso in cui le varianze siano molto diverse è possibile usare il test di Welch, calcolato da molti softwares.

Test t
Tra 2 medie campionarie

Ulteriori dettagli sul test t di Student (qui S_D è S_E) nel caso di campioni indipendenti con differente numerosità campionaria (n1 ≠ n2):

$$H_0 : \mu_1 = \mu_2$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_e}$$

$$s_e = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\text{gdl} = n_1 + n_2 - 2$$

Test t per campioni appaiati (n1 = n2):
Vedi esercizi su Excel e slides più avanti

Test F
Tra 2 varianze

Test parametrico che assume la distribuzione normale di entrambi i campioni.

L'ipotesi nulla H_0 è: i due campioni sono estratti da popolazioni aventi la stessa varianza.

La statistica F è il rapporto tra la varianza maggiore (S^2_X) sulla minore (S^2_Y) , la significatività è a due code con n_1 e n_2 come gradi di libertà.

Sotto l'ipotesi $H_0 = (\sigma_X^2 = \sigma_Y^2)$, ovvero se le due popolazioni hanno la stessa varianza, allora la variabile aleatoria

$$F = \frac{S_X^2}{S_Y^2}$$

segue la distribuzione di Fisher-Snedecor

$$\mathcal{F}(n-1, m-1)$$

di parametri *n-1* e *m-1*, dove *n* e *m* sono le numerosità dei due campioni.

Test Mann-Whitney
Tra 2 mediane

Il test di (Wilcoxon) **Mann-Whitney U** viene utilizzato per testare se le mediane di due campioni indipendenti sono differenti.
È un test non-parametrico che non assume distribuzione normale, ma assume che la forma della distribuzione dei due campioni sia la stessa.

L'ipotesi nulla è: H_0: i due campioni sono estratti da popolazioni aventi la **stessa mediana**.

Procedura: per ogni valore del campione 1, conta il numero di valori del campione 2 che sono più piccoli (coppie contano 0.5). Il totale di questi conteggi è la statistica U. Se il valore di U è più piccolo invertendo l'ordine dei campioni (2 vs 1 invece di 1 vs 2), quest'ultimo viene scelto come valore di U (U1+U2=n1n2).
Il valore di p viene calcolato con una asintotica approssimazione della distribuzione normale (a due code), che è valida solo per un'alta numerosità campionaria. La formula include una correzione di continuità per le coppie:

$$z = \frac{U - n_1 n_2/2 + 0.5}{\sqrt{\frac{n_1 n_2 \left(n^3 - n - \sum_g f_g^3 - f_g \right)}{12n(n-1)}}}$$

Test Mann-Whitney
Tra 2 mediane

Il valore di p viene calcolato con una asintotica approssimazione della distribuzione normale (a due code), che è valida solo con un'alta numerosità campionaria. La formula include una correzione di continuità per le coppie:

$$z = \frac{U - n_1 n_2 / 2 + 0.5}{\sqrt{\frac{n_1 n_2 \left(n^3 - n - \sum_g f_g^3 - f_g \right)}{12 n (n-1)}}}$$

Dove $n = n_1 + n_2$ e f_g è il numero degli elementi della coppia g.

Per n_1+n_2<=30 (es. 15 elementi per campione o gruppo), può essere calcolato il valore di p esatto. Nel caso in cui si testi una numerosità campionaria più alta il test diventa meno robusto.

Test della mediana di Mood

Il test della mediana di Mood è un'alternativa a Mann-Whitney, ha un potere più basso e quindi, se possibile, è meglio utilizzare M-W U.

Tuttavia, nei casi in cui all'interno del campione ci siano forti outlier (valori aberranti) il test di Mood funziona meglio.

Procedura: il test conta il numero di valori in ogni campione che ricadono sopra o sotto la mediana aggregata, producendo una tabella di contingenza 2x2 che viene testata con un test del Chi-quadro con 2 gradi di libertà, senza la correzione di Yates.

Test di Kolmogorov-Smirnov
Tra 2 distribuzioni

Il test di è Kolmogorov-Smirnov, è un test non-parametrico per il confronto delle distribuzioni di 2 campioni. In altre parole, non testa direttamente la media, la varianza o altri indici in particolare.

L'ipotesi nulla è: H_0: I due campioni sono presi da popolazioni con uguale distribuzione.

Il test statistico è la massima differenza assoluta tra le due distribuzioni empiriche cumulative:

$$D = \max_x \left| S_{N_1}(x) - S_{N_2}(x) \right|$$

Il valore di p è calcolato sulla base della funzione: $$Q_{KS}(\lambda) = 2\sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2\lambda^2}$$

…e del valore Ne = N1N2/(N1+N2), ottenendo:

$$p = Q_{KS}\left(\left[\sqrt{N_e} + 0.12 + 0.11/\sqrt{N_e}\right] D\right)$$

Test di Kolmogorov-Smirnov
Alternative a K-S

Esistono alcune alternative al test K-S, ad esempio il test di Anderson-Darling ed il test di Epps-Singleton.

The **Anderson-Darling** test is a nonparametric test for overall equal distribution of two univariate samples. It is an alternative to the Kolmogorov-Smirnov test.

The **Epps-Singleton** test (Epps & Singleton 1986; Goerg & Kaiser 2009) is a nonparametric test for overall equal distribution of two univariate samples. It is typically more powerful than the Kolmogorov-Smirnov test, and unlike the Kolmogorov-Smirnov it can be used also for non-continuous (i.e. ordinal) data.

Test Coefficiente di Variazione
Fligner-Kileen test

Testa il coefficiente di variazione in due campioni a confronto.
CV è definito come il rapporto tra la deviazione standard e la media (in percentuale) ed è calcolato così:

$$CV = \frac{s}{\bar{x}} \cdot 100 = \frac{\sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}}{\bar{x}} \cdot 100$$

L’ipotesi nulla è: H_0: I due campioni sono presi da popolazioni con uguale coefficiente di variazione.

Con valori di $p<0.05$ posso rifiutare l’ipotesi nulla, quindi i campioni provengono da popolazioni differenti.

Questo test è potente e non influenzato dalla distribuzione dei due campioni (non-parametrico).

Gli output del test sono la statistica T e la expected T, il valore Z e il valore p.

Test 2 campioni appaiati
Two-sample paired tests

Tre test principali (1 parametrico e 2 non-parametrici) che confrontano 2 campioni (2 colonne) i cui elementi sono appaiati (per riga).
Ad esempio: la lunghezza delle braccia (sinistro e destro - colonne) di un certo numero di persone (righe) oppure la temperatura in estate e in inverno (colonne) di un certo numero di siti (righe). La possibilità di controllare un fattore di disturbo (come l'appartenenza ad un sito o ad una persona – fattore random) aumenta la potenza del test.

Ecco I principali:

- ✓ **Test t** per campioni appaiati (parametrico)
- ✓ **Test sign** per campioni appaiati (non-parametrico)
- ✓ **Test Wilcoxon** per campioni appaiati (non-parametrico)

Test t
Two-sample paired tests

L'ipotesi nulla è: H_0: la media della differenza tra i campioni è zero.

Procedura: verifica che la media della differenza sia uguale a zero utilizzando la formula del test t per un campione sulle differenze. Il test assume che vi sia distribuzione normale delle differenze tra i campioni appaiati.

$$s = \sqrt{\frac{1}{n-1}\sum\left(d_i - \bar{d}\right)^2} \qquad t = \frac{\bar{d}}{s/\sqrt{n}}$$

Dove $d_i = x_i - y_i$ e n-1 sono i gradi di libertà.

Test Wilcoxon
Two-sample paired tests

L'ipotesi nulla è: H_0: nessuna differenza tra le mediane.

Il test è non-parametrico e non assume che vi sia distribuzione normale.

Procedura: vengono calcolate le differenze per ogni riga, le righe con d = 0 vengono rimosse, le altre in valore assoluto $|d_i|$ sono ordinate in ranghi (R_i). La somma dei ranghi delle differenze positive d_i è W_+, la somma dei ranghi delle differenze negative è W_-.

Il test è quindi $W = \max(W_+, W_-)$

Il ***sign* test** (binomiale) conta il numero di casi n_1 in cui $x_i > y_i$ e n_2 in cui $y_i > x_i$. Il numero massimo (n_1, n_2) è poi riportato. Il valore p è esatto ed è calcolato dalla distribuzione binomiale. <u>Il *sign* test potrebbe essere meno potente rispetto agli altri test per campioni appaiati, ma non ha praticamente nessuna assunzione</u>.