Test d'ipotesi

- Prof. Giuseppe Verlato
- Sezione di Epidemiologia e Statistica Medica, Università di Verona

Nel '900 falsificazione di ipotesi (Karl Popper)

Nell' '800 dimostrazione di ipotesi (Claude Bernard)

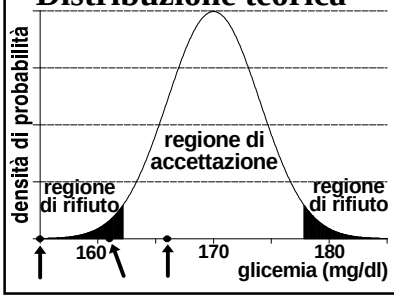
TEST D'IPOTESI

Ipotesi Nulla (H_0)
(Tutte le differenze osservate sono dovute al caso)

Ipotesi alternativa (H_1)

FALSIFICARE L'IPOTESI NULLA

Distribuzione teorica



~~**Distribuzione teorica**~~

Test non-parametrici

Metto i dati in ordine crescente:

AAABAABBABBBABBBBB

I valori del gruppo A tendono ad essere più bassi dei valori del gruppo B

Nel Medioevo l'uomo era AL CENTRO

. dello spazio (teoria geocentrica)

. del tempo (basandosi sulla Bibbia la Terra era nata 4000 anni a.C.)

. del mondo biologico (teoria fissista)

Attualmente l'uomo NON è più al centro

. dello spazio (teoria eliocentrica, periferia del Sole nella Via Lattea, esistenza di altri 100 miliardi di galassie)

. del tempo (la Terra è nata 5 miliardi di anni fa, il Big Bang ha avuto luogo 15 miliardi di anni fa)

. del mondo biologico (teoria dell'evoluzione della specie)

IPOTESI SCIENTIFICA:

Affermazione che si può sottoporre a verifica, che si può tentare di falsificare. Con una procedura che comporta delle misurazioni si può cercare di dimostrare che l'ipotesi non è vera.

Un'ipotesi scientifica viene ritenuta vera finché non si dimostra il contrario.

IPOTESI STATISTICA:

Affermazione circa una caratteristica di una popolazione che si cerca di supportare o di rifiutare sulla base delle informazioni disponibili, in genere ricavate da un campione.

TEST D'IPOTESI

H_0: IPOTESI NULLA
Tutte le differenze osservate sono delle semplici fluttuazioni casuali

H_1: IPOTESI ALTERNATIVA
Le differenze riscontrate nelle statistiche campionarie rispecchiano una reale differenza nei parametri delle popolazioni corrispondenti

Esempio:

La glicemia dei diabetici italiani è uguale alla glicemia dei diabetici americani

La glicemia dei diabetici italiani è diversa dalla glicemia dei diabetici americani

TEST STATISTICO:

Regola che consente di discriminare tra i risultati che portano a non rifiutare o a rifiutare l'ipotesi nulla (H_0).

Nel riportare la decisione si riporta anche la probabilità che questa sia corretta.

Dati del campione

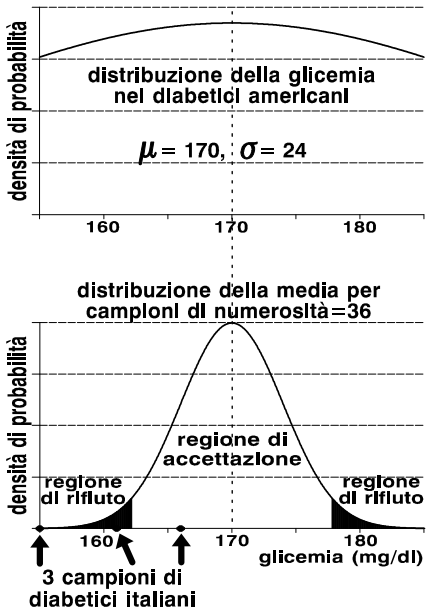
Test statistico

P>0,05 = la probabilità che le differenze osservate siano dovute al caso è superiore al 5%

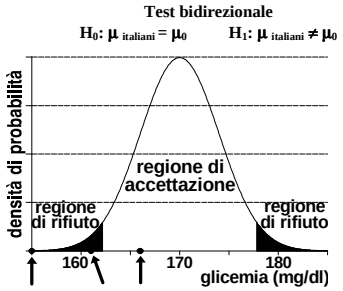
Accetto l'ipotesi nulla (H0)
tutte le differenze osservate tra i campioni possono essere attribuite al caso

P<0,05 = la probabilità che le differenze osservate siano dovute al caso è inferiore al 5%

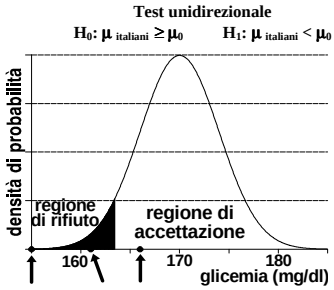
Accetto l'ipotesi alternativa, H1
le differenze osservate tra i campioni rispecchiano delle differenze reali tra le popolazioni



Two-sided



One-sided



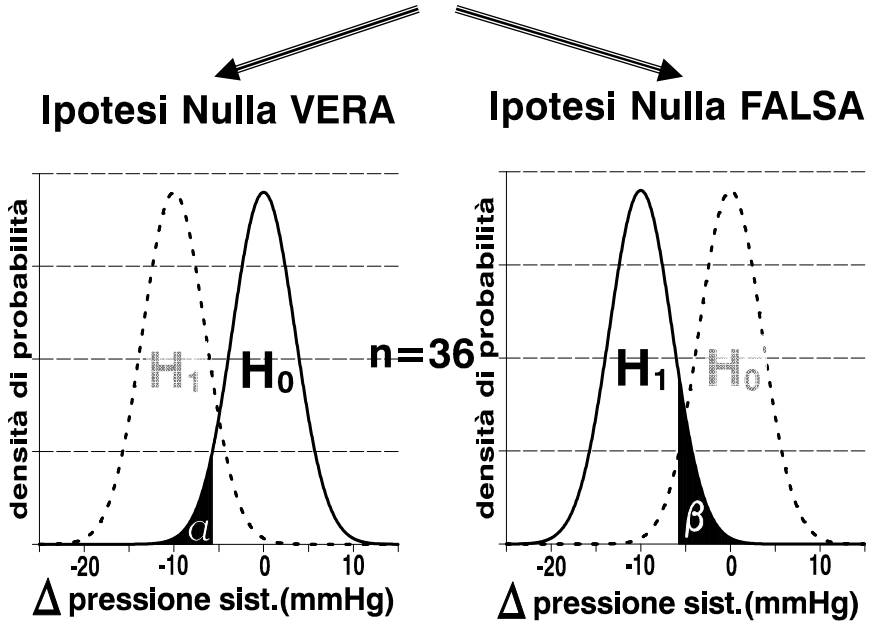
Ipotesi Nulla (H_0)

	vera	**falsa**
Accetto H_0	**Va bene**	**Errore del II tipo**
Rifiuto H_0	**Errore del I tipo**	**Va bene**

P(errore del I tipo) = α (alfa)
P(errore del II tipo) = β (beta)

In genere, nel test d'ipotesi la probabilità di errore del I tipo viene fissata al 5% (0,05). Pertanto in un caso su 20 si rifiuterà H_0 (ovvero il test risulterà significativo) per semplice effetto del caso, anche quando H_0 è vera. In termini statistici si sceglie un livello di significatività del 5%.

Ad esempio, se in un test d'ipotesi P<0,01, vuol dire che posso rifiutare H_0 con una probabilità di errore del I tipo inferiore all'1%; in altre parole la probabilità che le differenze osservate siano dovute al caso è inferiore all'1%.



Attualmente si preferisce riportare la probabilità esatta associata con un determinato test statistico.

Anziché scrivere **P<0,01** si preferisce riportare **P=0,003**.

Al posto di un **test d'ipotesi** si esegue un **test di significatività**.

Tuttavia,

mentre **P<0,01 è la probabilità prefissata che le differenze siano dovute al caso sotto H_0** ,

P=0,003 è la probabilità di osservare quel determinato risultato o un risultato più estremo (PTOME = Probability of This Or More Extreme) sotto H_0.

Con un livello di significatività del 5%, un test su 20 risulta significativo per semplice effetto del caso.

Se si eseguono 100 test statistici, per effetto del caso 5 risulteranno significativi. Pertanto c'è il rischio di inondare la letteratura scientifica con segnalazioni che poi risulteranno fallaci.

Si verifica un' "inflazione di alfa, un'inflazione della probabilità di errore del I tipo". Questa inflazione viene definita come "*multiple testing bias*" (bias da test ripetuti).

Il *multiple testing bias* si verifica, ad esempio, quando:

1) si prendono in considerazione molte variabili,
2) si ripetono le analisi in diversi sottogruppi
3) si ripetono le analisi a diversi intervalli di tempo.

ANALISI DEI SOTTOGRUPPI

Nello studio internazionale ISIS2 (1988) l'aspirina si è dimostrata superiore al placebo nel trattamento dell'infarto e nella prevenzione di ulteriori episodi ischemici.

Tuttavia nel sottogruppo del segno zodiacale dei Gemelli il placebo appariva più efficace dell'aspirina.

Bibliografia

ISIS-2 (Second International Study of Infarct Survival) Collaborative Group (1988) Randomized trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. Lancet ii: 349-360.

Come ovviare al *multiple testing bias*?

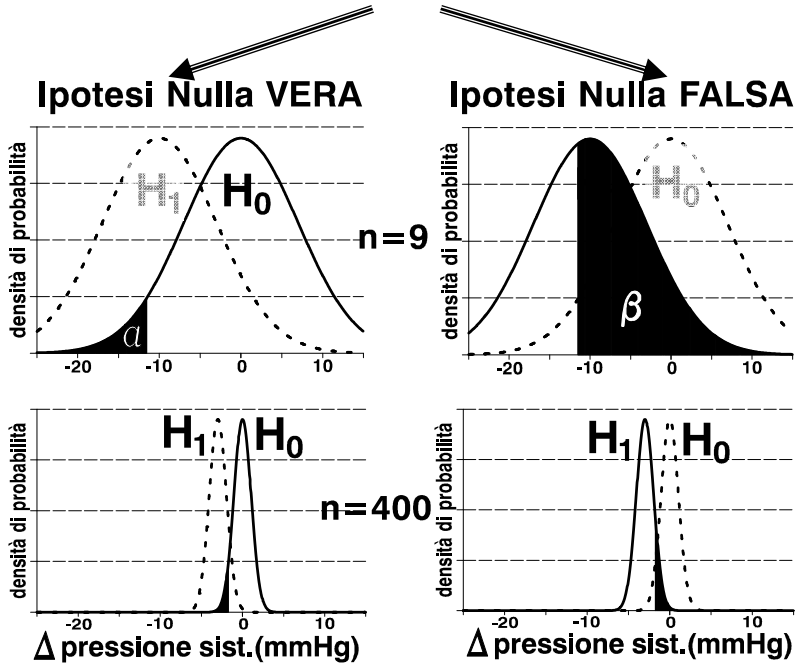
Il **medico di base** di fronte agli **esami di laboratorio**:	Il **biostatistico** di fronte a **molti test statistici**:
Individua degli **esami più importanti** e degli esami meno importanti	Individua degli **end-points (eventi) primari (pochi)** e degli end-points secondari
Prende in considerazione solo/ soprattutto gli **esami particolarmente sballati**	Adotta un **livello di significatività più conservativo** ($p<0.05 \rightarrow p<0.01$), o corregge la P osservata (correzione di Bonferroni)
Guarda la **concordanza** fra i vari esami (indicatori di sofferenza epatica, di infezione virale)	Guarda la **concordanza** fra i vari end-points

POTENZA di un test = 1- beta = 1 - P(errore del II tipo)

E' la probabilità che un test statistico ha di falsificare l'ipotesi nulla quando l'ipotesi nulla è effettivamente falsa.

In altre parole, la Potenza di un test è la sua capacità di cogliere delle differenze, quando queste differenze esistono.

Il test statistico è costruito in modo da mantenere costante il livello di significatività, indipendentemente dalla numerosità campionaria. Ma questo risultato viene raggiunto a spese della potenza del test, che aumenta all'aumentare della numerosità campionaria.



SIGNIFICATIVITA' STATISTICA e RILEVANZA CLINICA

Un'indagine epidemiologica, condotta su un gran numero di persone, ha messo in luce che i fumatori dormono meno della popolazione generale.

La differenza aveva una significatività elevata (P<0.001), ovvero ben difficilmente poteva essere attribuita al caso.

La differenza consisteva in 3 minuti di sonno in meno nei fumatori rispetto ai non-fumatori.

La POTENZA di un test dipende:

1) dalla numerosità del campione

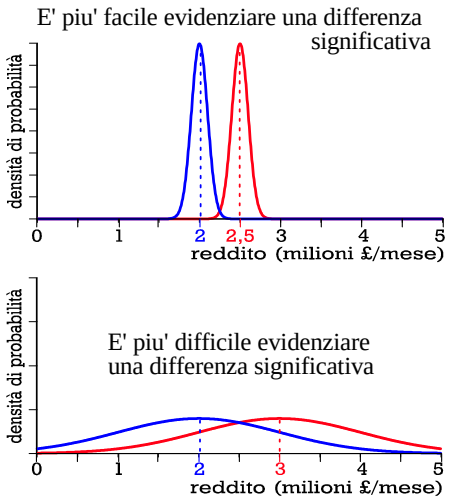
2) dalla variabilità del fenomeno in studio

3) dalla differenza minima che si vuole mettere in evidenza

4) dal livello di significatività adottato.

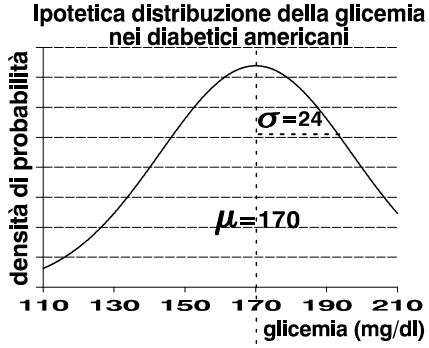
Il modo principale per raggiungere un'adeguata potenza è pianificare un'adeguata numerosità campionaria nel protocollo dello studio.

significativita' statistica $\approx$ differenza osservata / variabilita' casuale



L’intervallo di confidenza come test d’ipotesi

L’intervallo di confidenza è una stima intervallare, ma può anche essere considerato un vero e proprio test d’ipotesi.



IPOTESI NULLA: Nei diabetici italiani la distribuzione della glicemia è uguale a quella dei diabetici americani. Pertanto la media per campioni di numerosità 36 si distribuisce con:

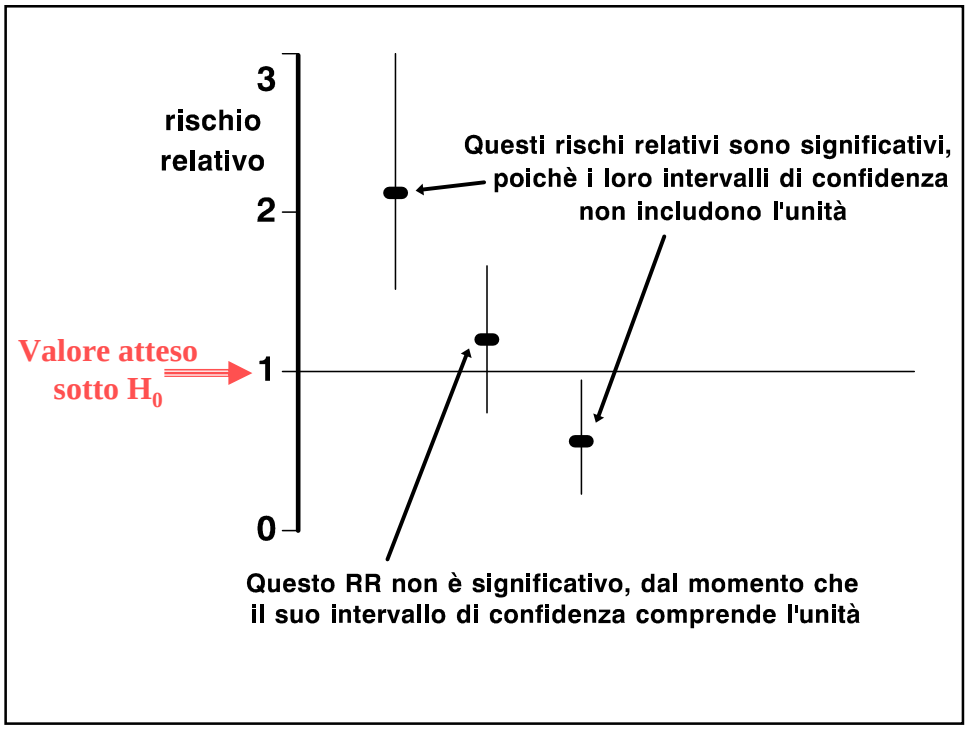
$\mu = 170$

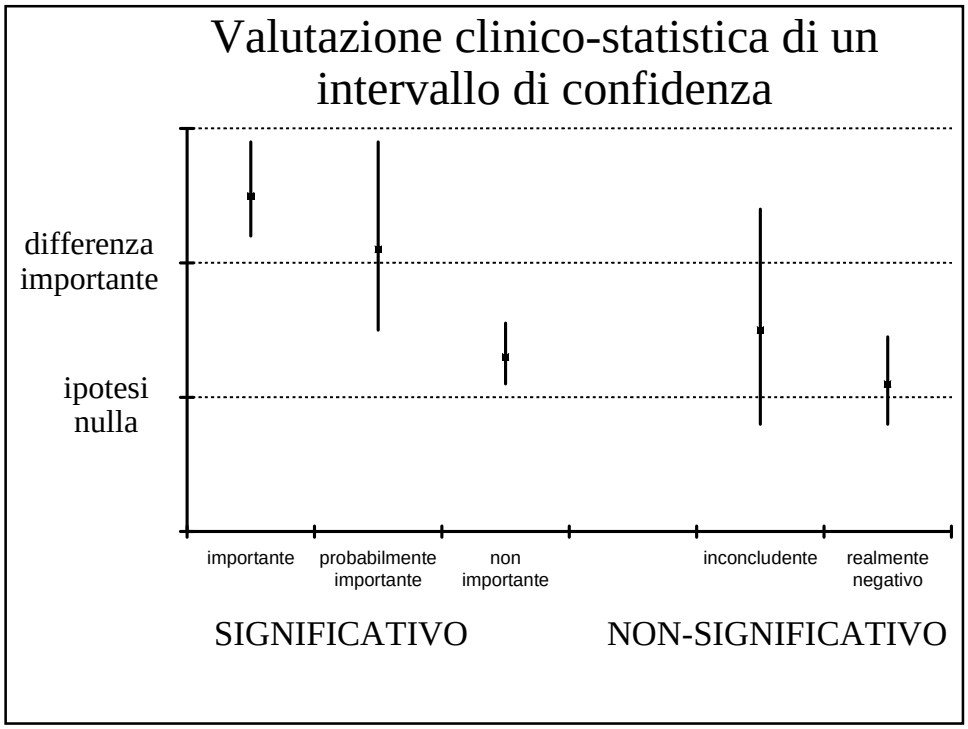
$\sigma/\sqrt{n} = 24/\sqrt{36} = 4$

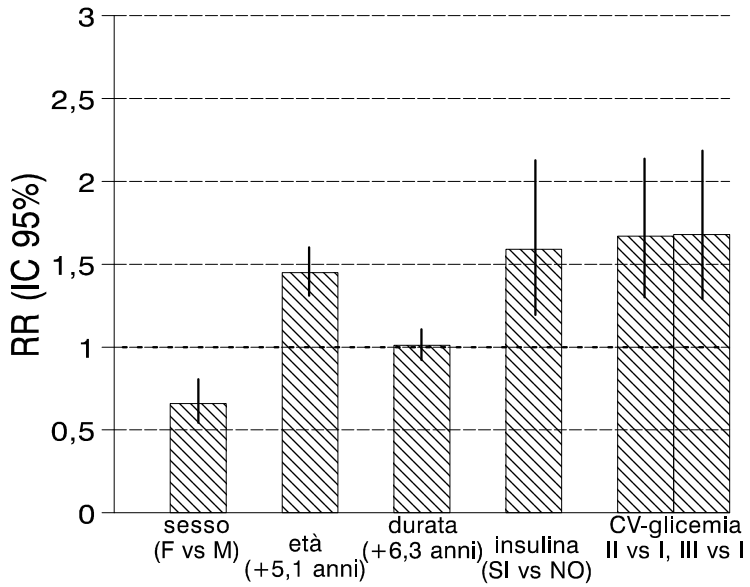
I primi due intervalli di confidenza sono significativi perché non contengono 170 mg/dl (valore atteso sotto H0)

stime intervallari di μ

$155 \pm 1{,}96*4$	$147{,}2 \text{ —— } 162{,}8$
$161 \pm 1{,}96*4$	$153{,}2 \text{ —— } 168{,}8$
$166 \pm 1{,}96*4$	$158{,}2 \text{ —— } 173{,}8$


3
rischio
relativo
2
1
0
Valore atteso
sotto H_0
Questi rischi relativi sono significativi,
poichè i loro intervalli di confidenza
non includono l'unità
Questo RR non è significativo, dal momento che
il suo intervallo di confidenza comprende l'unità


Valutazione clinico-statistica di un intervallo di confidenza
differenza importante
ipotesi nulla
importante
probabilmente importante
non importante
inconcludente
realmente negativo
SIGNIFICATIVO
NON-SIGNIFICATIVO



1) Muggeo M, Zoppini G, Bonora E, Brun E, Bonadonna RC, Moghetti P, Verlato G (2000) Effect of fasting plasma glucose variability on 10-year survival of type 2 diabetic patients from the Verona Diabetes Study. *Diabetes Care*, 23:45-50

"Overemphasis on hypothesis testing - and the use of P values to dichotomise significant or non-significant results - has detracted from more useful approaches to interpreting study results, such as estimation and confidence intervals.

In medical studies investigators are usually interested in determining the size of difference of a measured outcome between groups, rather than a simple indication of whether or not it is statistically significant ...

Confidence intervals, if appropriate to the type of study, should be used for major findings in both the main text of a paper and its abstract."

Gardner MJ, Altman DG (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. British Medical Journal, 292: 746-750

International Committee of Medical Journal Editors

"When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid sole reliance on statistical hypothesis testing, such as the use of P values, which fails to convey important quantitative information."

International Committee of Medical Journal Editors (1992) Uniform requirements for manuscripts submitted to biomedical journals [Special Report] N Engl J Med, 324: 424-428.

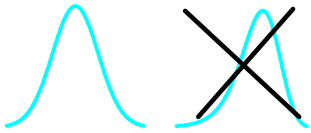
Scelta del test statistico

All'inizio di un'elaborazione statistica la prima domanda da porsi è: Di che tipo è la variabile?

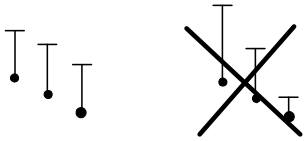
	NOMINALE	ORDINALE	QUANTITATIVA
Esempi	Stato di vita (vivo/morto) Sesso (M/F) Tipo di ventilazione (spontanea/assistita/artificiale)	Intensità del dolore Profondità del coma Wassermann (-, -+, +, ++, +++, ++++, +++++)	Peso (Kg) Età (anni) Glicemia (mmol)
Test indicati	Chi-quadrato (χ^2)	Test non-parametrici	T di Student per dati non-appaiati o per dati appaiati Analisi della varianza Regressione e correlazione

ASSUNZIONI PER TEST PARAMETRICI

NORMALITA'



OMOSCHEDASTICITA'
(STABILITA' della VARIANZA)



INDEPENDENZA delle OSSERVAZIONI (ERRORI)

N = numero animali, non numero di prove

t di Student
analisi della varianza
analisi della covarianza
regressione

Con una variabile di tipo quantitativo, qual è il test statistico da effettuare?

Confronto fra soggetti diversi		Misure ripetute sugli stessi soggetti		Confronto fra variabili diverse
↓	↓	↓	↓	↓
2 gruppi	Più di 2 gruppi	2 misurazioni	Più di 2 misurazioni	
↓	↓	↓	↓	
t di Student	ANOVA a 1 criterio	t di Student per dati appaiati	ANOVA per misure ripetute	Regressione e Correlazione

ANOVA = ANalysis Of VAriance (Analisi della varianza)

1) Viene condotto uno studio sugli studenti iscritti alla Facoltà di Farmacia. L'indice di massa corporea (peso/statura^2) delle matricole viene confrontato con l'indice di massa corporea degli iscritti al terzo anno. Che tipo di test si può utilizzare per questo confronto?

2) Nello stesso studio in un gruppo di studenti l'indice di massa corporea (peso/statura^2) viene misurato due volte, sia al momento dell'iscrizione che alla fine del terzo anno di corso. Che tipo di test si può utilizzare per confrontare queste due misurazioni successive?

3) Nella stessa indagine viene studiata la relazione tra peso e statura. Che tipo di test si può utilizzare?

4) Nella stessa indagine viene studiata la relazione tra colore degli occhi e colore dei capelli. Che tipo di test si può utilizzare?

A) test t di Student per dati non-appaiati
B) test t di Student per dati appaiati
C) test del chi-quadrato
D) regressione e correlazione
E) altro _____

TEST PARAMETRICI		TEST NON-PARAMETRICI
frequenza cardiaca, pressione arteriosa	Variabili	dolore, Glasgow coma score
Test t per dati non-appaiati	2 campioni indipendenti	Test di Wilcoxon-Mann-Whitney
ANOVA a 1 criterio	K campioni indipendenti	Test di Kruskal-Wallis
Test t per dati appaiati	2 campioni dipendenti	Test di Wilcoxon
ANOVA per misure ripetute	K campioni dipendenti	Test di Friedman
Correlazione e regressione	Associazione tra due variabili	Coeff. di correlazione ordinale di Spearman