Test non parametrici per campioni indipendenti

Riassunto

Sono presentati i test non parametrici utilizzabili nel caso di campioni indipendenti, cioè, nella ricerca clinica, nel caso di uno studio prospettico randomizzato a gruppi paralleli. L'esposizione è condotta in relazione alla natura dell'endpoint da analizzare, ossia se la risposta al trattamento è collocabile su una scala nominale, ordinale o di rapporti. Inoltre, si accenna anche al concetto di potenza-efficienza che lega i test parametrici ai corrispondenti test non parametrici.

Parole chiave. *Test non parametrici, ranghi, studi comparativi, potenza-efficienza.*

Summary

Nonparametric tests for independent samples.

Nonparametric tests can be used in analysing results of a prospective, parallel, randomised study. These tests are shown in relationship with the type of the considered endpoint: nominal, ordinal, or ratio scale. Moreover, the concept of power-efficiency with respect to the correspondent parametric test is outlined

Key words. *Nonparametric tests, ranks, comparative studies, power-efficiency.*

In "Statistica per concetti 2" del numero precedente di CASCO, si sono definiti i test parametrici (quelli che sono basati su assunzioni circa la popolazione target), in relazione al tipo di dati che possono presentarsi nella ricerca clinica.

Nella presente nota proseguiamo il discorso introducendo i test non parametrici, ma considerando solo quelli più utilizzati per l'analisi dei risultati di uno studio clinico randomizzato a gruppi paralleli. Questi test prendono il nome di "test per campioni indipendenti" in quanto ciò che accade in un braccio dello studio non è influenzato da quanto osservato negli altri bracci (i pazienti sono diversi). Tale restrizione, dettata unicamente da esigenze di spazio, non è esaustiva di tutti i test non parametrici, in quanto restano esclusi quelli appropriati per l'analisi dei dati sia nel caso di altri tipi di disegni comparativi (cross over, matching), sia nel caso di studi di fase 2, quando tutti i pazienti ricevono lo stesso trattamento.

Si dicono **non parametrici** i test per la cui costruzione non si fa ricorso ad ipotesi sulla popolazione target.

Scale nominali

Le scale nominali sono particolarmente importanti nella ricerca clinica, non solo per la loro ampia diffusione (si pensi ai caratteri dicotomici – ossia a due modalità – come guarito/non guarito, successo/insuccesso terapeutico), ma anche perché, all'occorrenza, dati di altra natura possono essere ricondotti a caratteri dicotomici.

Esempio 1. La sopravvivenza globale (Overall Survival, OS) è un carattere quantitativo continuo. Nei casi in cui è opportuno o conveniente, la OS può essere trasformata in carattere dicotomico considerando, ad es., la sopravvivenza a due anni: vale "alta" se supera due anni, vale "bassa" altrimenti.

Esempio 2. Il Functional Living Index-Emesis (FLI-E) per il vomito è composto da 9 item per ciascuno dei quali la risposta è collocata su una scala di Likert a 7 punti; quindi, per un paziente, FLI-E assume valori da un minimo di 9 (peggiore impatto del vomito sulla qualità di vita) ad un massimo di 63 (nessun impatto). Tali dati, per ciascun paziente, possono essere trasformati in un carattere dicotomico, considerando un cut-off di 54, definendo i valori non inferiori a tale soglia come No (or minimal) Impact (of vomiting) on patient's Daily Life (NIDL), mentre quelli al di sotto come Impact (of vomiting) on patient's Daily Life (IDL).

Come già esposto in "Statistica per concetti 2" del precedente numero di CASCO, nel caso di dati ottenuti con scale nominali, l'unica analisi possibile è quella basata sulle frequenze. Poiché i test basati su frequenze non richiedono alcuna assunzione sulla popolazione target, in caso di scale nominali i test da usare sono esclusivamente quelli non parametrici.

Nel caso di studi comparativi, il test basilare è il test "esatto" di Fisher, descritto in Statistica per concetti del n. 7 di CASCO (autunno 2013), cui vanno ad aggiungersi sue generalizzazioni (quando almeno uno dei due caratteri presenti un numero di modalità maggiore di 2: test di Freeman-Halton) ed approssimazioni (test chi-quadrato con o senza la correzione di Yates: cito i nomi perché consentono di individuarli nei package statistici). Tutti tali test sono non parametrici.

Scale Ordinali

Riguardando la rubrica "Casi clinici" del numero precedente (CASCO 9), si può osservare la corretta scelta degli autori di usare esclusivamente test non parametrici avendo a che fare con dati ottenuti mediante di scale di Likert.

Anche nel caso di scale ordinali, non possono essere usati test parametrici perché non si conosce la distanza tra due modalità consecutive. Ad esempio, per valutare l'intensità della nausea si può usare la scala di Likert: 0 = no nausea; 1 = nausea lieve

(compatibile con tutte le attività quotidiane); 2 = nausea moderata (rende impossibile l'esercizio di almeno alcune attività); 3 = nausea severa (costringe il paziente a letto). La distanza che c'è tra 1 e 2 non può essere considerata uguale a quella che c'è tra 2 e 3: le reali distanze non sono conoscibili a priori e variano da paziente a paziente, in relazione all'impatto della nausea sulla qualità di vita, come percepita dal paziente. Non avrebbe dunque senso calcolare la media delle intensità osservate (perché i dati, sebbene spesso esprimibili in forma numerica, non sono sommabili), mentre è possibile calcolare la mediana[1].

Nel caso di scale ordinali, i test non parametrici che consentono il confronto tra mediane sono basati sui "ranghi" che rappresentano i posti occupati nella distribuzione ordinata (detta "graduatoria") dai corrispondenti attributi.

Esempio 3.0. I ranghi. Siano 12, 8, 24, 4, 42 le osservazioni ottenute su 5 pazienti. Esse costituiscono una distribuzione. Ordinandoli in senso, ad esempio, crescente, si passa alla corrispondente graduatoria:
4, 8, 12, 24, 42.
Si assegni, ora, il posto che tali osservazioni ordinate occupano nella graduatoria:
1, 2, 3, 4, 5.
I valori così ottenuti si chiamano ranghi, e sono semplicemente i posti che le osservazioni ottenute occupano nella graduatoria.

Esempio 3.1. Ranghi in uno studio comparativo. I valori rilevati su una scala di Likert a 9 punti nei pazienti randomizzati a due trattamenti, A e B siano i seguenti:
A: 4, 3, 7, 1
B: 6, 8, 9.
Tali valori vanno sostituiti con i ranghi (cioè i posti che essi occupano nella graduatoria complessiva dei 7 attributi elencati). Ordinandoli senza riferimento al trattamento si ha:
1, 3, 4, 6, 7, 8, 9
cui corrispondono i ranghi (cioè i posti in graduatoria) 1, 2, 3, 4, 5, 6, 7.
Pertanto le risposte ai due trattamenti valutati con i ranghi sono:
A: 3, 2, 5, 1
B: 4, 6, 7.

Esempio 3.2. Ranghi legati (tied ranks). Nel caso che più attributi siano uguali (e quindi abbiano lo stesso rango) a ciascuno si sostituisce la media dei ranghi che avrebbero avuto nella graduatoria se fossero stati diversi. Ad esempio,
A: 5; 2; 2; 1
B: 5; 2; 9.
Ordinando i dati senza far riferimento al trattamento si ha:
1; 2; 2; 2; 5; 5; 9.
Complessivamente, i ranghi da assegnare ai dati ordinati sono sempre 7 (da 1 a 7). Pertanto, a 1 si assegna il rango 1, ai 2 la media dei ranghi che avremmo assegnato se fossero stati diversi: (2 + 3 + 4)/3 = 3, analogamente ai due 5 (che occupano il 5° e il 6° posto): (5 + 6)/2 = 5,5; infine a 9 il rango maggiore: 7. In tal modo sostituendo ai dati originali i ranghi si ha
1; 3; 3; 3; 5,5; 5,5; 7
che, ricollocati rispetto ai due trattamenti, danno luogo alle seguenti due distribuzioni:
A: 5,5; 3; 3; 1
B: 5,5; 3; 7.

L'analisi dei dati consiste nel valutare se la mediana dei ranghi corrispondenti alle osservazioni nel trattamento A (nel primo esempio, per mediana dei ranghi in corrispondenza di A, si può assumere 2,5) possa ritenersi significativamente diversa dalla mediana dei ranghi di B (nel primo esempio: 6). Nel secondo esempio (ranghi legati) la mediana dei ranghi per A è 3, quella per B è 5,5.

In tali casi, le due mediane sono confrontate per mezzo del **test U di Mann-Whitney** (*Mann-Witney U-test*) o anche mediante il **test di Wilcoxon per la somma dei ranghi** (*Wilcoxon rank sum test*). Tali due test, trovati indipendentemente da autori diversi, seguono logiche differenti, ma sono equivalenti, nel senso che a qualunque distribuzione siano applicati, danno sempre lo stesso risultato in termini di significatività della differenza tra le due mediane.

Nel caso che i trattamenti fossero più di due, un test complessivo che consente di determinare se almeno una mediana sia significativamente diversa dalle altre è il **test di Kruskall-Wallis**. Ad esempio, nel caso di 3 trattamenti, si calcola il test di Kruskall-Wallis. Se esso risulta significativo, vuol dire che almeno una mediana è significativamente diversa dalle altre. Per individuare a quale trattamento si riferisca, o si riferiscano, le mediane significativamente diverse, si può procedere confrontandole a due a due con il test di Wilcoxon per la somma dei ranghi, badando però a correggere il livello di significatività in base alla disuguaglianza di Bonferroni (v. CASCO 1): quindi occorre fare il minor numero possibile di confronti, scegliendo quelli essenziali, per non penalizzare troppo il livello di significatività per ciascuno di essi.

Scale di rapporti.

Come si è visto in "Statistica per concetti 2" del numero scorso di CASCO, in uno studio clinico randomizzato a due gruppi paralleli, per il confronto tra le medie potrebbero essere usati i test parametrici (t-test per campioni indipendenti), ma a condizione che siano rispettate le ipotesi su cui tali test si basano (normalità dell'errore, uguaglianza delle varianze), o che queste siano al più moderatamente violate (il t-test è un test robusto).

Per verificare se le suddette assunzioni sono rispettate esistono test statistici (test di normalità, test per l'uguaglianza delle varianze).

Potrebbe, però, anche essere usato il test U di Mann-Whitney (o, equivalentemente, il test di Wilcoxon per la somma dei ranghi).

Per decidere quale test sia più vantaggioso, e di quanto, occorre introdurre il concetto di Potenza-Efficienza (P-E).

Supponiamo di avere due test, S e T, ugualmente ammissibili per eseguire

1. Si definisce "mediana" il termine che divide la corrispondente graduatoria (cioè la distribuzione ordinata) in modo da lasciare a sinistra lo stesso numero di termini che lascia a destra.
Ad esempio:
distribuzione: 28, 12, 36, 54, 8
graduatoria: 8, 12, 28, 36, 54.
Mediana della distribuzione è 28 perché, nella graduatoria, lascia a sinistra due termini e a destra gli altri due.
Un altro esempio.
distribuzione: 44, 8, 4, 12
graduatoria: 4, 8, 12, 44.
Mediana è qualunque valore compreso tra 8 e 12, ma, convenzionalmente, per mediana si assume la semisomma dei termini centrali:
(8 + 12)/2 = 10.

un determinato confronto (ad es., t-test e test U). Supponiamo, inoltre, che occorrano 100 pazienti per braccio affinché il test S abbia una potenza dell'80% di individuare la differente efficacia tra i due trattamenti (ossia che ci sia l'80% di probabilità che il test S risulti significativo, se i due trattamenti hanno una diversa efficacia), mentre ne occorrano 120 per raggiungere con il test T la stessa potenza. In tal caso, la Potenza-Efficienza di T rispetto a S è pari a

P-E di T = (100/120) x 100 = 83,3%.

Si dice allora che il test T ha una P-E dell'83,3% del test S; in altre parole, per usare T con la stessa potenza di S, nelle condizioni dell'esempio, occorre arruolare il 20% di pazienti in più.

Supponiamo che siano rispettate le ipotesi di normalità e di uguaglianza delle varianze. In tal caso il test parametrico (t-test) è più potente del corrispondente non parametrico (test U), ma la Potenza-Efficienza del test U è circa il 95% di quella del t-test, cioè per avere un test U con la stessa potenza del t-test, sarebbe necessario arruolare solo poco più del 5% di pazienti in più. Tutto ciò vale solo se le assunzioni alla base del t-test sono rispettate, altrimenti il confronto sarebbe illogico e potrebbe anche accadere che una differenza non risulti significativa con il t-test mentre lo sia con il test U (mi è capitato più di una volta).

Nel caso di più di due trattamenti, il test parametrico da usare per valutare se almeno una media possa essere ritenuta significativamente diversa dalle altre è il test F di Fisher-Snedecor per l'analisi della varianza. Il corrispondente test non parametrico è il test di Kruskall-Wallis.

Come si è detto, sono più vantaggiosi i test parametrici (hanno una potenza superiore e quindi richiedono un minor numero di pazienti da arruolare), ma a condizione che siano rispettate le assunzioni su cui fondano. La verifica di tali assunti va condotta con appositi test statistici (di normalità, di uguaglianza delle varianze): se non risultano significativi, possono essere usati i test parametrici, altrimenti la scelta dovrebbe ricadere su quelli non parametrici.

Nel caso di piccoli campioni, però, non è possibile decidere se le assunzioni su cui si basano i test parametrici siano rispettate, perché non c'è una potenza sufficiente affinchè possano risultare significativi. In tali casi o si hanno informazioni esterne allo studio che convincano che le assunzioni siano verosimili (ad esempio, quando i dati si riferiscono a variabili biometriche di cui è nota la forma normale della distribuzione, come ad es., per la glicemia che si distribuisce normalmente nei soggetti non malati), oppure, per sicurezza, è preferibile usare i test non parametrici.

In conclusione, sarebbe sempre ragionevole usare i test non parametrici, proprio in quanto hanno una Potenza-Efficienza prossima a quella dei corrispondenti test parametrici; eppure il loro uso non è molto frequente. La ragione più plausibile è che, nell'analisi dei dati, oggi si tende ad usare, anziché test, modelli statistici che non solo consentono il confronto tra i gruppi sperimentali, ma permettono anche di ottenere preziose informazioni sull'importanza dei fattori prognostici e, quindi, di fare più accurate previsioni.

Brevissima guida bibliografica

Un manuale praticamente completo sui test non parametrici, molto chiaro e ricco di esempi è

– Siegel S. Non Parametric Statistics. Tokio: Mc Graw Hill-Kogakusha, 1956.

Molto più approfondito dal punto di vista teorico, ma tratta solo i test non parametrici basati sui ranghi, è il volume:

– Lehmann EL. Nonparametrics: statistical methods based on ranks. San Francisco: Mc Graw Hill, 1975.

Come si vede, tali opere sono molto datate, proprio perché oggi si ragiona soprattutto in termini di modelli statistici. Tuttavia per un clinico, ricercatore o utente dei risultati della ricerca, è importante avere qualche nozione sui test non parametrici soprattutto perché si usano spesso negli studi clinici per i confronti dell'endpoint primario e di quelli secondari tra i gruppi sperimentali.

Enzo Ballatori