L'ANGOLO DELLA STATISTICA

I test multipli

Catherine Klersy

G Ital Aritmol Cardiostim 2001;4:145-148

Servizio di Biometria ed Epidemiologia Clinica Direzione Scientifica, IRCCS Policlinico San Matteo, Pavia

La ricerca in campo biomedico spesso genera una *molteplicità di dati, di ipotesi e di analisi* che porta all'esecuzione di test statistici multipli. È comunemente accettata, almeno negli studi confermatori, la necessità di aggiustare per i confronti multipli eseguiti[1,2] ed è anche suggerito nelle corrispondenti linee guida biostatistiche[3] e nelle istruzioni per gli autori di alcune riviste. Tutti avranno incontrato nella lettura di lavori scientifici la "correzione di Bonferroni": questa rappresenta il metodo di aggiustamento attualmente più utilizzato,[4] ma ne esistono altri. In questa revisione si cercherà di rispondere a 3 domande:

- Perché è necessario aggiustare per la molteplicità dei test?
- Quando è necessario farlo?
- Come è possibile farlo?

Perché è necessario aggiustare per la molteplicità dei test?

Il tasso di errore rappresenta una misura probabilistica di effettuare un'inferenza erronea, dato un set di ipotesi da cui viene tratta una conclusione comune. Più in particolare, la probabilità di rifiutare erroneamente l'ipotesi nulla, quando questa è vera, viene chiamata errore di I tipo o errore α; in altri termini l'errore di I tipo è la probabilità di falsi positivi. Si possono calcolare tassi di errore α per ogni singola ipotesi valutata e tassi di errore relativi all'insieme delle ipotesi. Nel primo caso si parlerà di tasso di errore per singolo confronto (*comparisonwise*), nel secondo di tasso di errore per esperimento (*experimentwise*). La seconda strategia proposta è quella più consigliata. Infatti, aumentando il numero di test effettuati, aumenta la probabilità di errore di I tipo: se testiamo un'ipotesi nulla che in effetti è vera, utilizzando α come valore critico, la probabilità di ottenere un risultato non significativo (corretto) è $1-\alpha$; se testiamo 2 ipotesi *indipendenti* la probabilità che nessuno dei 2 test sia significativo è data, per un teorema del calcolo delle probabilità, dal prodotto delle probabilità $(1-\alpha)*(1-\alpha)$; più generalmente se testiamo K ipotesi indipendenti la probabilità che i test siano congiuntamente non significativi è data da $(1-\alpha)^K$; ne consegue che la probabilità di avere almeno un test significativo sarà $1-(1-\alpha)^K$. Esemplificando, se vengono testate 20 ipotesi indipendenti al livello di significatività $\alpha = 0,05$, la probabilità che nessuna sia significativa è $0,95^{20} = 0,36$. La probabilità che almeno

una sia significativa per errore sarà $1-(1-0,05)^{20} = 0,64$, ben superiore al valore nominale prescelto del 5%.

Riassumendo, un aggiustamento per i test multipli è necessario per mantenere norme di buona pratica clinica. Il non tenere conto della molteplicità dà luogo a un aumento della probabilità di trovare risultati significativi in favore dell'ipotesi alternativa, quando l'ipotesi nulla è vera.

Quando è necessario aggiustare per la molteplicità dei test?

L'essenza dell'interpretazione di uno studio clinico risiede nell'accordo fra aspetti clinici e aspetti statistici. Questi dipendono da una serie di fattori, quali la malattia in studio, la popolazione studiata, gli endpoint, il disegno dello studio, la conduzione dello studio, la scelta appropriata dell'analisi statistica in base al disegno e ai quesiti scientifici posti. La molteplicità dei test, in particolare, può rendere difficoltosa l'interpretazione dei risultati, soprattutto se non prevista nel disegno. Come abbiamo visto, se si continuano a effettuare test per un tempo sufficiente, inevitabilmente si finirà per trovare un risultato "significativo". È importante, pertanto, non attribuire un peso eccessivo a risultati significativi isolati, in mezzo alla massa di quelli non significativi.

In pratica, possiamo considerare la famiglia delle K ipotesi da testare come un esperimento. Dopo aver scelto di controllare l'errore per esperimento, dobbiamo decidere quali test di uno studio appartengono al particolare esperimento.[1] Se per esempio abbiamo uno studio con 3 diversi trattamenti confrontati con un trattamento standard, possiamo considerare come esperimento il set di tutti i confronti appaiati (6 ipotesi), oppure alternativamente l'esperimento che comprende il confronto di ognuno dei 3 trattamenti rispetto al controllo (3 ipotesi). In questi 2 esperimenti, l'aumento dell'errore di I tipo sarà diverso, essendo diverso il numero di ipotesi valutate e quindi diverso dovrà essere il controllo dell'errore per esperimento. Un procedimento così rigoroso è strettamente necessario solo in studi confermatori, cioè in studi in cui la dimostrazione di ipotesi predefinite condiziona un processo decisionale, ad esempio:

1. Un farmaco è efficace? Lo metto sul mercato.
2. Una caratteristica clinica o strumentale aumenta il rischio? Stratifico i pazienti in base al rischio e metto in opera una strategia di prevenzione.

In questi studi è richiesta una specificazione chiara delle ipotesi multiple e delle loro priorità. In particolare verranno trattenute solo ipotesi di rilevanza scientifica nel contesto specifico e si cercherà di ridurre il numero di ipotesi, soprattutto se ridondanti (correlate).

Negli studi esplorativi in cui i dati sono raccolti con uno scopo, ma senza ipotesi chiave prespecificate, un aggiustamento per test multipli non è così indispensabile. Questo è vero anche perché un tale aggiustamento non consente comunque di controllare il bias legato ai test di ipotesi generati dai dati. In questo caso i risultati dei test multipli di significatività possono essere utilizzati solo per descrivere, ma non per decidere, qualsiasi sia il metodo di controllo dell'errore di I tipo.

Le fonti di molteplicità sono diverse e includono:

a) *Gli endpoints multipli*: si possono considerare tali le misure ripetute nel tempo, i fenomeni caratterizzati da manifestazioni multidimensionali (ad es. malattie sindromiche, misure della qualità della vita, ecc.), le possibili misure di efficacia terapeutica (ad es. in pazienti con angina instabile, la prevalenza di morte, d'infarto miocardico, di rivascolarizzazione in urgenza o in emergenza possono essere altrettante misure di efficacia).
b) *Gli studi multipli*: in questi si cerca di dimostrare l'efficacia di un trattamento attraverso 2 o più studi fondamentali effettuati in centri diversi.
c) *I confronti fra più gruppi di pazienti o fra più braccia di trattamento*.
d) *Le analisi o i test multipli*: in particolare l'analisi dei sottogruppi per evidenziare caratteristiche particolari dei dati, l'analisi "per-protocol" (sui pazienti che hanno effettivamente seguito il trattamento) rispetto all'analisi "intention to treat" (in base al trattamento assegnato al momento della randomizzazione).
e) *Le analisi ad interim*: sono le analisi effettuate nel corso di un clinical trial per mettere in evidenza una superiorità marcata di uno dei trattamenti, con conseguente interruzione precoce dello studio.

Come è possibile aggiustare per la molteplicità dei test?

Esistono diverse procedure per aggiustare per la molteplicità dei test: procedure generali basate sull'aggiustamento del p-value, procedure ad hoc, statistiche-

test globali e procedure speciali a seconda della fonte di molteplicità considerata.

Procedure basate sul p-value: la loro caratteristica principale è di non dipendere dal tipo di dato analizzato (continuo, categorico, sopravvivenza) né dal tipo di test utilizzato (t, Chi2, logrank…). Includono la procedura di Bonferroni e i suoi miglioramenti.

Procedura di Bonferroni[1,2,4,5]

Per ognuna delle ipotesi si calcola:

1. valore soglia sotto cui si rifiuta l'ipotesi nulla, $p \leq \alpha/K$;
2. valore aggiustato $p_{aggiustato} = K^*p_{osservato}$

Il valore soglia sotto il quale verranno rifiutate le ipotesi nulle dipende dal numero di test effettuati; la correzione è indipendente dal tipo di dato; non utilizza la struttura di correlazione; è conservativa solo se vengono valutate numerose ipotesi (>5) e se i dati sono molto correlati (correlazione ≥0,7).

Procedure di Bonferroni modificate (stepwise)[1,2,6,7]

Sono state proposte diverse modifiche alla procedura di Bonferroni, tutte meno conservative. Si fondano sul concetto che delle K ipotesi nulle testate, le sole in cui è necessario proteggersi contro un rifiuto, sono quelle non ancora rifiutate. I metodi proposti si basano su un ordinamento crescente dei valori di p delle varie K ipotesi. Ognuna viene valutata in sequenza e confrontata con un valore soglia sempre meno conservativo.

Le procedure principali sono quelle di Holm, di Hochberg e di Hommel, elencate in ordine di potenza crescente. Illustriamo solo quella di Holm, rimandando alla bibliografia[2,7] per ulteriori dettagli.

- Siano $p_1>p_2>\ldots>p_K$ i valori ordinati di p relativi alle K ipotesi nulle.
- Rifiuta H_{0K} se $p_K<\alpha$ e passa allo step successivo, alternativamente fermati e accetta tutte K ipotesi.
- Rifiuta $H_{0(K-1)}$ se $p_{(K-1)}<\alpha/(K-1)$ e passa allo step successivo, alternativamente fermati e accetta tutte le (K-1) ipotesi.
- Ecc.
- I valori aggiustati di p sono dati da $p_{ak} = \max\{K^*p_K, (K-1)^*p_{(K-1)}, \ldots, k^*p_k\}$, k = 1,2…K.

Esempio: supponiamo di condurre uno studio in cui vengono valutate 4 ipotesi e in cui osserviamo i seguenti valori di p (ordinati):

p osservato	Valore soglia Bonferroni	p aggiustato Bonferroni	Valore soglia Holm	p aggiustato Holm = max{$4p_4$, $3p_3$, $2p_2$, p_1; kp_k} k = 1, 2, 3, 4
0,081	0,05/4 = 0,013	0,081*4 = 0,324	*stop*	0,081 = max (0,020, 0,045, 0,052, 0,081; *0,081*)
0,026	0,05/4 = 0,013	0,026*4 = 0,104	0,05/2 = 0,025	0,052 = max (0,020, 0,045, 0,052; *0,052*)
0,015	0,05/4 = 0,013	0,015*4 = 0,060	0,05/3 = 0,017^	0,045^ = max (0,020, 0,045; *0,045*)
0,005°	0,05/4 = 0,013	0,005*4 = 0,020°	0,05/4 = 0,013^	0,020^ = max (0,020; *0,020*)

Solo l'ultimo valore di p (°) risulta minore del valore soglia con la procedura di Bonferroni e quindi verrà rifiutata solo l'ipotesi 4, mentre con la procedura di Holm vengono rifiutate le ipotesi 3 e 4 (^), mantenendo un errore nominale di 0,05 per tutto l'esperimento.

Procedure ad hoc: utilizzano le informazioni relative alla struttura di correlazione che esiste fra le ipotesi/gli endpoint. Si citano le procedure di Tukey per endpoint molto correlati[2] e le procedure di Dubey e Armitage Pamar.[2]

Test globali: sono test basati sulla teoria della normalità; lo scopo di un test globale è di dimostrare un beneficio generale, per esempio di una terapia, considerando simultaneamente tutti i test/endpoint. Non sono indicati quando si è interessati a inferenze separate sulle diverse ipotesi. Esempi sono il T^2 di Hotelling, i test di O'Brien (OLS/GLS/non parametrico) e di Pocock.[2,5]

Procedure speciali: in alcune particolari situazioni sono state sviluppate procedure apposite.

Confronto di più gruppi (>2)[1,6,8]

Il caso patognomonico è il confronto delle medie con l'analisi della varianza. I confronti successivi dei gruppi 2 a 2 sono basati su test che controllano l'errore di I tipo per esperimento; esempi sono i test di Scheffé, il test di Tukey, il test di Dunnet, il test SNK e altri.

Analisi per sottogruppi[9,10,11]

Questa analisi si giustifica se è presente una eteroge-

neità di effetto nei diversi sottogruppi, che viene identificata valutando l'interazione in un modello di regressione. Solo interazioni di rilevanza scientifica vanno considerate e devono essere prespecificate in fase di disegno. È particolarmente importante interpretare clinicamente l'entità dell'effetto (e il suo intervallo di confidenza) al fine di concludere per la presenza di eterogeneità fra i gruppi. Se presente interazione, i successivi test all'interno dei sottogruppi vanno eseguiti controllando per l'aumentato errore di I tipo (con uno dei metodi riportati sopra).

Analisi ad interim[1,12]

Hanno come scopo di fermare il trial prima del termine se vi è sostanziale superiorità di un trattamento. Le analisi ad interim vanno pianificate in fase di disegno per avere rilevanza, e non essere occasionate dai dati (o peggio, da un congresso). La presenza e il numero di analisi ad interim condizioneranno la numerosità del campione. La soglia a cui si dichiara significativo l'effetto e per cui si interrompe il trial deve essere corretta per i test multipli. Esistono due gruppi di procedure sequenziali: in un primo gruppo la soglia viene aggiustata con un valore che è costante per tutto il periodo (ad es. metodo di Bonferroni); in un secondo gruppo, la soglia viene aggiustata mano a mano che procede il trial (ad es. metodi di O'Brien-Fleming, di Peto, di Lan-DeMets). In quest'ultimo caso si cerca di utilizzare criteri più stringenti per rifiutare l'ipotesi nulla nelle fasi precoci dello studio e di avvicinarsi il più possibile al valore nominale alla fine dello studio.

Endpoint multipli[1,2,5,13,14]

L'analisi di ogni endpoint separatamente condiziona, come abbiamo visto, un aumento dell'errore di I tipo; inoltre, viene ignorata la correlazione tra gli endpoint. Nel caso di endpoint multipli, si possono ipotizzare diverse strategie di analisi: identificare uno degli endpoint come primario e considerare gli altri come endpoint secondari da valutare controllando per i test multipli; generare un endpoint aggregato (ad es. morte e/o recidiva di infarto miocardico); utilizzare uno dei metodi riportati sopra basati sull'aggiustamento del p-value o su test globali.

Conclusioni

In questa breve rassegna abbiamo visto come aumenta la probabilità di falsi positivi quando aumenta il numero di ipotesi valutate e quindi di test statistici eseguiti. Le fonti di molteplicità sono diverse e si ritrovano in molti aspetti della ricerca biomedica. Vanno quindi identificate in fase di disegno al fine di utilizzare metodi statistici di analisi appropriati che minimizzano il rischio di dichiarare efficace un trattamento o fondamentale un fattore prognostico, quando in realtà è solo un effetto del caso. In tal senso appare utile considerare lo studio come un esperimento di cui vanno identificati tutti gli elementi come suggerito da Bender et al.[1]

Bibliografia

1. Bender R, Lange S. Adjusting for multiple testing-when and how? *J Clin Epidemiol* 2001;54:343-349.
2. Sankoh AJ, Huque MF, Dubey SD. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statist Med* 1997;16:2529-2542.
3. The CMP working party on efficacy of medical products. Biostatistical methodology in clinical trials in application for marketing authorization for medical products. *Statist Med* 1995;14:1659-1682.
4. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 1995;310:170.
5. Pocock JS, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987;43:487-498.
6. Wright SP. Adjusted P-values for simultaneous inference. *Biometrics* 1992;48:1005-1013.
7. Comelli M, Klersy C. Different methods to analyze clinical experiments with multiple endpoints: a comparison on real data. *J Biopharm Stat* 1996;6:115-125.
8. Altman DG, Bland JM. Statistics notes: comparing several groups using analysis of variance. *BMJ* 1996;312:1472-1473.
9. Altman DG, Matthews JNS. Statistics notes: Interaction 1: heterogeneity of effects. *BMJ* 1996;313:486.
10. Matthews JNS, Altman DG. Statistics notes: Interaction 2: compare effect sizes not P values. *BMJ* 1996;313:808.
11. Matthews JNS, Altman DG. Statistics notes: Interaction 3: how to examine heterogeneity. *BMJ* 1996;313:862.
12. Geller NL, Pocock SJ. Interim analysis in randomized clinical trials: ramifications and guidelines for practitioners. *Biometrics* 1987;43:213-223.
13. Cupples LA, Heeren T, Schatzkin A, Colton T. Multiple testing of hypotheses in comparing two groups. *Ann Intern Med* 1984;100:122-129.
14. Tang DE, Geller NL, Pocock SJ. On the design and analysis of randomized trials with multiple endpoints. *Biometrics* 1993; 49:23-30.

Indirizzo per la corrispondenza

Catherine Klersy
Servizio di Biometria ed Epidemiologia Clinica
IRCCS Policlinico S. Matteo
27100 Pavia
e-mail: klersy@smatteo.pv.it