COS'È LA POTENZA DI UN TEST STATISTICO?

C. Limoni, O. Petrini

Generalmente gli studi clinici sono effettuati con l'intento di dimostrare la superiorità di una nuova terapia su quella in vigore[1]. Consideriamo uno studio che confronti l'efficacia di due terapie, la terapia A e la terapia B, in cui la terapia B è più efficace di A: la potenza di un test statistico è definita come *la probabilità di scoprire*, sulla base di un confronto tra *campioni* di pazienti, alcuni curati con A e altri curati con B, *che la terapia B è effettivamente migliore*.

Ipotesi di studio e errori possibili nel caso di studi scientifici

Per capire pienamente il significato della potenza di un test statistico è necessaria una breve introduzione su alcuni concetti chiave. Essi riguardano le *ipotesi statistiche* da verificare durante un lavoro scientifico, cioè l'*ipotesi nulla* (H_0) e l'*ipotesi alternativa* (H_A), così come gli errori che si possono commettere al momento di enunciarne le conclusioni.

Ad esempio, in un processo sappiamo che, quando si cerca di provare la colpevolezza di un indiziato, l'ipotesi di partenza (ipotesi nulla) è normalmente quella di "non colpevolezza". L'ipotesi alternativa, la cui validità si cerca di provare, è invece quella di "colpevolezza". Il giudice, nell'intento di dimostrare che l'imputato è colpevole (rifiutando quindi l'ipotesi nulla), raccoglie e valuta le prove in suo possesso, poi emette la sentenza. Esiste però la possibilità che commetta due tipi di errore: il primo consiste nel condannare l'imputato quando in realtà egli è innocente, ***rifiutando quindi a torto l'ipotesi nulla*** di non colpevolezza, mentre il secondo consiste nell'assolvere l'imputato quando egli è in realtà colpevole, rifiutando a torto, per insufficienza di prove, ***l'ipotesi alternativa***.

Anche negli studi clinici, al momento di confrontare l'efficacia di due terapie A e B, si enuncia un'ipotesi nulla di "non differenza" (H_0) con l'intento di rifiutarla: *si presuppone cioè che le due terapie diano dei risultati simili*. Le prove sono costituite dai risultati dello studio e, nel caso della durata di un certo sintomo, l'ipotesi nulla di "non differenza" tra le terapie è respinta solo se le prove raccolte per dimostrare la superiorità di B – ad esempio: il sintomo scompare più rapidamente – sono sufficienti. Il criterio di cui si serve il ricercatore per dichiarare che le prove sono sufficienti è il ***p-valore***. Esso è la probabilità che una differenza tra i risultati dei due gruppi di terapia sia dovuta al caso, cioè alla variabilità dei dati campionari. Statisticamente si accetta l'ipotesi alternativa (H_A) quando il p-valore è inferiore al 5% (considerato come livello di significatività α): in questo caso le prove sono considerate come sufficienti per rifiutare l'ipotesi nulla (H_0) di "non differenza".

L'errore consistente nel *rifiutare a torto l'ipotesi alternativa*, decretando che B non è migliore di A quando ciò è il caso, è statisticamente definito come errore β. Questo tipo di errore può essere ridotto utilizzando un'adeguata potenza del test.

Il primo tipo di errore (rifiuto a torto dell'ipotesi nulla) in statistica è spesso definito come errore di tipo 1, mentre il secondo (rifiuto a torto dell'ipotesi alternativa) è anche chiamato errore di tipo 2.

Generalmente si ammette che una potenza del test pari a 80%, avere cioè l'80% di probabilità di scoprire che una terapia è effettivamente migliore di un'altra, è ritenuta sufficiente. Con questa potenza, se B è veramente migliore di A, lo si vedrà nell'80% degli studi che confrontano l'efficacia di A e B, cioè in 8 studi su 10. Se la probabilità di scoprire che una terapia è effettivamente migliore di un'altra è dell'80%, la probabilità di non scoprirlo è necessariamente del 20%. Questa probabilità è l'errore β. L'errore β è quindi pari a 100%-potenza del test.

Fattori che influenzano la numerosità necessaria

Diversi fattori influenzano la probabilità di commettere l'errore β. Intuitivamente si può immaginare che se l'efficacia di B è superiore a quella di A, a parità di potenza, ***l'entità della differenza*** influisce sul numero di pazienti necessari a dimostrarla. In altre parole, se la superiorità è schiacciante, sarà più facile dimostrarla rispetto a situazioni in cui il beneficio supplementare di B è minimo. D'altra parte, la variabilità dei dati, misurata dalla deviazione standard nei campioni, influenza pure la numerosità, in quanto una variabilità maggiore richiederà una numerosità maggiore del campione. Anche il livello di significatività del test (α) influenza la numerosità richiesta: una maggiore "sicurezza" di accettare la "non differenza" tra i risultati si traduce in una maggiore numerosità dei gruppi. In altre parole, un livello di significatività α dell'1% richiederà una numerosità maggiore rispetto ad un livello α del 5%.

Considerazioni di tipo etico

Prima di passare all'illustrazione di un esempio concreto, dobbiamo soffer-

1 Esistono comunque anche studi clinici volti a dimostrare l'equivalenza di due terapie, oppure la non inferiorità di una terapia rispetto ad un'altra

marci sulle implicazioni etiche dell'errore β. Se la terapia B è effettivamente migliore di A e la potenza del nostro test non è sufficientemente elevata, in altre parole se la probabilità di scoprirlo è bassa, le conseguenze sono di due tipi: prima di tutto, con grande probabilità, includeremo inutilmente delle persone in uno studio clinico – esponendole eventualmente a potenziali eventi avversi – e d'altra parte la nuova terapia, più efficace, non potrà essere d'aiuto ai futuri pazienti, poiché non saremo riusciti a dimostrarne l'utilità.

Un esempio

Immaginiamo uno studio il cui obiettivo principale consista nel verificare l'efficacia di un nuovo prodotto nella terapia del dolore, tramite il confronto della durata del sintomo. Secondo uno studio pilota, il sintomo sembra scomparire mediamente dopo 15 o 16 minuti al massimo. Si sa che utilizzando un prodotto di riferimento – potrebbe essere un placebo oppure il "gold standard" sul mercato – il sintomo scompare in media dopo 20 minuti. Si tratta ora di valutare quale sia la numerosità che ci assicuri una probabilità dell'80% di scoprire che B è migliore di A.

La **Tabella 1** illustra alcuni casi possibili, a seconda dei valori assunti dai fattori che influiscono sulla numerosità da adottare, cioè l'entità della differenza media dei risultati e la variabilità dei dati. In questo esempio il livello di significatività α è posto sempre al 5%.

Vediamo dapprima la situazione 1, nella quale i ricercatori hanno deciso di arruolare arbitrariamente, senza effettuare quindi nessun calcolo di potenza, 10 pazienti per gruppo. Qui la potenza del test è pari a 56%. Questo significa che se B è effettivamente più efficace di A, la probabilità di scoprirlo è praticamente una su due, il che equivale al lancio di una monetina, dove se si ottiene testa B è migliore e se si ottiene croce A e B sono simili. Si sprecano dunque inutilmente e in modo non etico delle risorse.

	Scelta arbitraria di n	Influenza della variabilità nei dati, con medie uguali		Differenza tra medie minore	Differenza tra medie minore e maggiore variabilità
	Situazione 1	Situazione 2	Situazione 3	Situazione 4	Situazione 5
Livello di significatività del test, α	5%	5%	5%	5%	5%
Media gruppo 1, μ_1	15	15	15	16	16
Media gruppo 2, μ_2	20	20	20	20	20
Differenza tra medie, $\mu_1 - \mu_2$	-5	-5	-5	**-4**	-4
Deviazione standard comune,	5	5	6	5	6
Potenza del test (%)	56%	**80%**	**80%**	**80%**	**80%**
Numerosità per gruppo, n	**10**	**17**	**24**	**26**	**37**

Tab. 1: Simulazione di calcolo della potenza (confronto di due medie, Student-t test)

Nella situazione 2 la potenza del test è stata fissata preventivamente all'80%. Si osserva che per questo tipo di differenza (5 minuti tra A e B) il numero di pazienti da arruolare è di 17 per gruppo. In una situazione reale bisognerà poi tener conto di eventuali "drop-outs", aggiungendoli ai 34 citati.

Nella situazione 3 si può osservare l'influsso di un aumento della variabilità dei dati sulla numerosità richiesta. Qui la deviazione standard stimata aumenta di un punto, passando a 6, mentre la differenza tra le medie rimane 5. La numerosità aumenta a 24 pazienti per gruppo.

Nella situazione 4 la differenza di efficacia è stata ridotta di un punto, con la stessa variabilità della situazione 2. La numerosità in questa simulazione passa a 26 pazienti per gruppo.

Infine nella situazione 5 sia l'efficacia, sia la variabilità sono state modificate, per simulare la situazione più sfavorevole, dove l'efficacia è leggermente ridotta e congiuntamente la variabilità corrisponde a quella ipotizzata nella situazione 3. Qui la numerosità diventa 37 pazienti per gruppo, cioè più del doppio della situazione 2.

Con questo semplice esempio si può quindi constatare come la variazione nell'efficacia ipotizzata e (o) nella variabilità dei risultati possa influenzare in modo decisivo la numerosità richiesta per i campioni e di conseguenza aumentare anche i costi dello studio.

Conclusioni

La questione della potenza di un test statistico assume un'importanza particolare nel caso degli studi clinici. Un intero capitolo delle linee guida ICH E9 le è dedicato. In questa linea guida, tra l'altro, è espressamente specificato che questa numerosità deve permettere di rispondere in modo affidabile a tutte le domande poste e che il calcolo deve essere effettuato partendo dall'obiettivo principale dello studio. Convenzionalmente devono essere utilizzati un livello di significatività pari a α=5% e una potenza del test dell'80% al minimo.

La metodologia utilizzata per questi calcoli deve essere specificata nel protocollo dello studio, dove si dovranno anche indicare gli effetti previsti e la variabilità di cui si è tenuto conto.

Nell'ipotizzare la differenza minima utilizzata si deve tener conto della sua valenza clinica e per quanto pos-

sibile basarsi su dati pubblicati, che ne giustifichino la validità.
I calcoli di potenza sono sempre effettuati con programmi specifici, quali Nquery Advisor o Sample Power. Esistono anche dei programmi gratuiti, quali G*Power, disponibili in Internet.

C. Limoni
Alpha 5-Biometrics, Riva San Vitale
O. Petrini
Istituto cantonale di microbiologia, Bellinzona

Definizioni

- H_0: ipotesi nulla "di non differenza"
- H_A: ipotesi alternativa
- Potenza del test: probabilità che l'ipotesi alternativa di uno studio clinico sia accettata, quando essa è vera.
- **p-valore**: probabilità di osservare un certo risultato se è vera l'ipotesi statistica di non differenza. Per convenzione, se il p-valore è <0,05 si considera che il risultato è significativo, cioè non dovuto alla variabilità campionaria.
- **La significatività statistica (p<0,05)** significa che l'effetto della terapia è statisticamente, ma non necessariamente clinicamente rilevante.

Bibliografia

1 Altman DG, Machin D, Bryant TN, Gardner MJ, editors. Statistics with confidence. 2nd ed. London: BMJ Books; 2000.

2 Altman DG. Practical statistics for medical research. Chapman & Hall, 1991.

3 Bland M. An introduction to medical statistics. 2nd Edition. Oxford University Press. 1995.

4 Davies HTO, Crombie IK. What are confidence intervals and p-values? Accessibile nel sito http://www.whatisseries.co.uk/whatis/pdfs/What_are_conf_int.pdf

5 G*power 3, University of Düsseldorf.

6 CH E6. Guideline for good clinical practice. Accessibile nel sito http://www.emea.europa.eu/pdfs/human/ich/013595en.pdf

7 ICH E9. Statistical principles for clinical trials. Accessibile nel sito http://www.emea.europa.eu/pdfs/human/ich/036396en.pdf

8 Lwanga SK. (ed.). Adequacy of sample size in health studies. With contributions by Stanley Lemeshow, David W. Hosmer, Jr., Janelle Klar. Published on behalf of the World Health Organization by Wiley, New York, NY.1990.

9 Nquery Advisor 7, Statistical Solutions Ltd, Ireland.

10 Sample Power, SPSS Inc, Chicago, USA.