I test statistici in base al modello di randomizzazione

Nei testi di metodologia della ricerca clinica si legge spesso che le principali finalità della randomizzazione sono due:

a. consentire un accettabile bilanciamento dei bracci dello studio rispetto ai fattori prognostici, noti e sconosciuti;

b. fornire al test statistico che verrà usato per il confronto dei risultati tra i gruppi sperimentali una seconda base logica (dopo quella del modello di popolazione).

Ora, mentre la prima asserzione è abbastanza comprensibile (v. CASCO 2), la seconda risulta alquanto oscura e, spesso, non è ulteriormente chiarita. Scopo della presente nota è approfondire la seconda finalità della randomizzazione, sempre restando nei limiti di un modestissimo bagaglio formale. Come al solito, si è tentato di rendere la lettura della rubrica comprensibile in sé; tuttavia, per apprezzare al meglio quanto verrà esposto, si consiglia la rilettura di "Statistica per concetti" del numero precedente (Costruzione del test statistico in base al modello di popolazione, v. CASCO 6).

I test ottenuti dal modello di randomizzazione sono detti "test di randomizzazione" (randomization tests), e la loro costruzione non è condotta in base al modello di popolazione – ipotizzando, cioè, che i gruppi sperimentali siano campioni casuali estratti a sorte dalle popolazioni target – ma solo in base all'operazione di randomizzazione.

Il cardine della costruzione dei test di randomizzazione risiede nelle seguenti considerazioni.

La risposta (al trattamento), ossia ciò che si osserva sul paziente al termine dello studio, dipende non solo dal trattamento, ma anche dal paziente.

Allora, **se è vera l'ipotesi nulla di uguale efficacia dei trattamenti (H_0,** v. CASCO 6**), la risposta dipende solo dal paziente. Pertanto, sotto H_0, il paziente è randomizzato al trattamento con la sua risposta.**

Poiché, con alta probabilità, la randomizzazione garantisce un accettabile bilanciamento dei gruppi rispetto ad ogni caratteristica, nota o sconosciuta, un forte squilibrio tra i gruppi rispetto alla risposta (che può essere considerata una caratteristica sconosciuta al momento della randomizzazione) non può che dipendere dalla falsità dell'ipotesi nulla; il che porta a concludere che, con alta probabilità, i trattamenti a confronto non hanno la stessa efficacia.

Esempio. Si considerino 40 pazienti randomizzati a ricevere uno dei due trattamenti, A e B; siano le risposte "+" = successo terapeutico; "–" = insuccesso. Rappresentando il risultato dello studio nella seguente tabella:

	Trattamento		
Risposta	A	B	Tot.
+	10	2	12
–	10	18	28
Tot.	20	20	40

si può osservare che c'è un forte sbilanciamento a favore del trattamento A. Si tratta di verificare se lo squilibrio osservato (50% di successi ottenuti con A, 10% con B) possa essere attribuito al caso o alla differente efficacia dei trattamenti.

La risposta è fornita dalla probabilità che, sotto H_0, per puro effetto del caso, si presenti uno squilibrio non meno estremo di quello osservato, cioè dalla probabilità della tabella data + la somma delle probabilità delle tabelle più estreme di quella osservata, dove le "tabelle più estreme della osservata" sono quelle in cui lo sbilanciamento è maggiore di 50% vs 10%. Tecnicamente è assai semplice costruire le tabelle più estreme di quella osservata, **tenendo costanti le frequenze marginali**. Infatti, nella tabella considerata, è sufficiente diminuire, di volta in volta, di una unità la frequenza della casella [+, B] (nell'esempio, "2"), fino a che la frequenza della casella non assuma il valore minimo possibile (nell'esempio, 0); ad ogni diminuzione di una unità si ottiene una tabella via via sempre più estrema di quella osservata.

Inoltre, tenendo costanti le frequenze marginali, è possibile calcolare, sotto H_0, la probabilità della tabella osservata ($P_0 = 0{,}0063$; si omettono i calcoli per semplicità di esposizione). Le tabelle in cui A ottiene un risultato superiore a B ancor più accentuato di quello rilevato nello studio (tabelle più estreme), con le corrispondenti probabilità, sono:

Risp.\trattam.	A	B	Tot.
+	11	1	12
–	9	19	28
Tot.	20	20	40

$P_1 = 0{,}0006$;

Risp.\trattam.	A	B	Tot.
+	12	0	12
–	8	20	28
Tot.	20	20	40

$P_2 = 0{,}0000$.

Quindi, tenendo costanti le marginali, sotto H_0, la probabilità che, avendo eseguito la randomizzazione (quindi, per puro effetto del caso), si

presenti una tabella in cui lo sbilanciamento a favore di A è o come quello osservato o ancora più accentuato è P = 0,0063 + 0,0006 = 0,0069.

Se il test fosse unidirezionale (ad esempio, nel caso in cui B fosse un placebo e non fossero stati previsti drop-out), P sarebbe il livello di significatività: il trattamento A può essere considerato più efficace di B, con una probabilità di sbagliare nel prendere questa decisione inferiore al 7 per mille (P < 0,007).

Se, invece, il test fosse bidirezionale, occorrerebbe raddoppiare il livello di significatività ottenuto (procedura approssimata), in quanto, a priori, tra i "casi più estremi di quello osservato" dovrebbero essere annoverati anche quelli di direzione opposta (B più efficace di A). In conclusione (test bidirezionale), A è il trattamento più efficace, ad un livello di significatività inferiore all'1,4% (P < 0,014).

Il test che è stato costruito nell'esempio è noto come "test esatto di Fisher" (Fisher's exact test) e può essere ricavato sia in base al modello di popolazione che in base al principio di randomizzazione, nel senso che seguendo le due differenti logiche, si perviene allo stesso risultato[1].

Quanto esposto vale anche nel caso in cui il carattere di classificazione sia quantitativo, come, ad esempio, la riduzione della pressione diastolica in pazienti ipertesi trattati con due differenti anti-ipertensivi. Il punto di partenza è sempre lo stesso: dai risultati dello studio, si calcola la differenza di efficacia dei due trattamenti e poi, combinando, in tutti i modi possibili, i pazienti con la loro risposta, tra i due bracci di trattamento si calcola la probabilità di ottenere per puro effetto del caso, sotto H_0, uno sbilanciamento come quello osservato o ancora più estremo. Se tale probabilità è piccola (convenzionalmente, P ≤ 0,05) si respinge l'ipotesi nulla di uguale efficacia argomentando che, se fosse stata vera, si sarebbe presentato un evento raro che, come tale, con pratica certezza non si presenta. Viceversa, se è P > 0,05, si può concludere che si sia presentato uno degli eventi che, con alta probabilità (95%), ci si attendeva dovessero presentarsi sotto H_0; quindi, si accetta l'ipotesi nulla di uguale efficacia dei trattamenti perché non vi sono ragioni per respingerla.

In tal modo si perviene ad una classe di test di randomizzazione ampia quanto quella dei test statistici costruiti in base al modello di popolazione (includendo il test t di Student, l'analisi della varianza, ecc.). In tali casi, però, non vi è alcuna relazione tra test basati sul campionamento e test di randomizzazione, nel senso che, applicando entrambi agli stessi dati, si ottengono risultati differenti.

Riepiloghiamo i concetti principali.

a. La risposta (ciò che si osserva su un paziente al termine dello studio) dipende non solo dal trattamento, ma anche dal paziente.
b. Se è vera l'ipotesi nulla (di uguale efficacia dei trattamenti), la risposta dipende solo dal paziente.
c. Randomizzando i pazienti (ciascuno con la sua risposta) ai bracci di trattamento, se è vera l'ipotesi nulla, ci attende un buon equilibrio dei risultati tra i gruppi sperimentali. Se invece si osserva un forte squilibrio, si tratta di verificare se esso può essere attribuito al caso o, invece, è dovuto alla differente efficacia dei trattamenti.
d. Si calcola la probabilità di osservare per puro effetto del caso, sotto l'ipotesi nulla, uno squilibrio o come quello osservato o ancora più estremo.
e. Se tale probabilità è piccola (P ≤ 0,05) vuol dire che il risultato osservato appartiene all'insieme dei risultati che, complessivamente, sotto l'ipotesi nulla, avevano una bassa probabilità di presentarsi; quindi, si può concludere che, con alta probabilità, i trattamenti sono diversamente efficaci. Altrimenti (P > 0,05) non resta che accettare l'ipotesi nulla di uguale efficacia dei trattamenti perché non vi sono evidenze che inducano a respingerla.

L'importanza dei test di randomizzazione risiede nel fatto che non sempre è possibile considerare i due (o più) gruppi sperimentali come campioni estratti a sorte da popolazioni. Nella ricerca clinica, ad esempio, quando il principio di consecutività dell'arruolamento viene violato (come spesso accade quando, in un centro, sono aperti più protocolli su una stessa patologia), i gruppi sperimentali non possono essere riguardati come campioni casuali e, pertanto, i test costruiti in base al modello di popolazione non possono essere applicati: occorre far ricorso ai test di randomizzazione. Poiché il sospetto che i gruppi sperimentali non siano assimilabili a campioni casuali è abbastanza frequente, l'elemento di conforto è che, se l'endpoint primario ha la natura di una variabile binaria (ad esempio, successo o insuccesso), l'applicazione del test esatto di Fisher come test di randomizzazione conduce al medesimo risultato dello stesso test ricavato in base al modello di popolazione. Si ricordi, inoltre, che il test esatto di Fisher è quello di elezione quando le frequenze non sono grandi (come, ad esempio, spesso accade nell'analisi degli eventi avversi) e che qualunque software statistico calcola il test esatto di Fisher con la stessa facilità con cui calcola il chi-quadrato e, quindi, non c'è ragione per usare sue approssimazioni. Al riguardo, si noti che, oggi, il NEJM accetta solo lavori in cui siano usati test esatti.

I test di randomizzazione presentano alcuni problemi legati ai loro fondamenti (e, quindi, quasi filosofici, di cui non ci occuperemo in questa nota), ma il principale svantaggio rispetto ai test costruiti in

1. La dimostrazione di tale asserzione richiede nozioni di calcolo combinatorio che non si ritiene opportuno introdurre nella presente esposizione. Invito, però, il lettore desideroso di acquisire questa conoscenza a contattarmi per e-mail: e.ballatori@alice.it

base al modello di popolazione sta nel fatto che essi non hanno bisogno della conoscenza degli stimatori che, però, sono indispensabili per condurre analisi più sofisticate, come ad esempio l'analisi di covarianza o altre analisi multifattoriali (ad esempio, le analisi dei sottogruppi).

In conclusione, i test di randomizzazione sono semplici da costruire, ma sono anche poveri concettualmente poiché non fanno riferimento agli stimatori che, però, sono indispensabili per analisi più complesse. Poiché queste analisi sono molto frequenti e consentono di acquisire importanti conoscenze, occorre prestare molta attenzione per evitare che i gruppi sperimentali siano molto dissimili da campioni casuali (ad esempio, che sia rispettata la consecutività dell'arruolamento, che la selezione dei pazienti non avvenga introducendo arbitrariamente altri vincoli in aggiunta a quelli previsti dai criteri di eleggibilità e di esclusione), così da poter usare, per i confronti, test generati in base al modello di popolazione.

Comunque, se, malgrado tutte le attenzioni adottate nel management dello studio, i gruppi sperimentali non possono proprio essere considerati campioni casuali, allora è necessario ricorrere ai test di randomizzazione.

Enzo Ballatori