- Copyright - Il Pensiero Scientifico Editore downloaded by IP 3.17.74.227 Sat, 27 Apr 2024, 04:25:33

poteva essere preventivato e, quando molti pazienti non sono valutati il dubbio della presenza di una distorsione da selezione fa cadere i presupposti per l'uso di un test unidirezionale (di ciò si parlerà in dettaglio nella rubrica Statistica per concetti).

Incompletezza dell'arruolamento

Come sottolineato più volte dagli autori (persino nel summary), il più importante limite del lavoro deriva dalla chiusura precoce dello studio e dalla possibile distorsione da selezione dovuta all'elevata frazione di pazienti per cui non fu valutato l'endpoint primario. Gli autori hanno reiteratamente avvertito dell'esistenza di tali problemi, e, quindi, della necessaria cautela con cui vanno interpretati i risultati, ma ciò, se attesta la loro onestà intellettuale, non rimuove certo il difetto.

Delle 416 pazienti che, in base al protocollo, si sarebbero dovute arruolare, ne furono randomizzate 233, ma di 15 non fu possibile la valutazione (9 avevano ritirato il consenso, 5 avevano subito un intervento di isterectomia, 1 di ooforectomia).

Delle restanti 218 pazienti che furono randomizzate e valutabili, solo per 135 fu valutata la perdita della funzione ovarica a 2 anni (endpoint principale). In conclusione, la preventivata dimensione del campione si è ridotta ad 1/3.

Conclusioni

Lo studio è randomizzato e appare meticolosamente condotto. Desta qualche sorpresa il fatto che un simile studio, che non ha prodotto rilevanti novità in materia, data l'esistenza di studi pubblicati in precedenza sullo stesso argomento, e con tutti i difetti che ha, sia stato pubblicato dal NEJM.

Malgrado tutto, si può ritenere che goserelin abbia una sua efficacia. Infatti, la perdita della funzione ovarica si è osservata in 5/66 (8%) nel gruppo GOS e in 15/69 (22%) pazienti nel gruppo S (differenza significativa), valori assai vicini a quelli ottenuti circa 4 anni prima dallo studio di Lucia Del Mastro et al., sopra citato, con triptorelin (8,9% nel gruppo delle pazienti trattate, 25,9% nel gruppo di sola chemioterapia). Va osservato che poco meno del 10% delle giovani donne sottoposte a chemioterapia e trattate con un farmaco che protegge la funzione ovarica, comunque la perdono. Pertanto andrebbe raccomandato alle pazienti che tengono particolarmente alla loro capacità riproduttiva di seguire, se possono, anche la strada della crioconservazione degli embrioni. •

Statistica per concetti

Test bidirezionali e test unidirezionali

Riassunto

Vengono anzitutto sintetizzate la logica e la costruzione di test statistici per l'analisi dei risultati di esperimenti comparativi, precedentemente trattate nei numeri 5 e 6 di CASCO, con lo scopo di procedere alla definizione dei test unidirezionali, più potenti di quelli bidirezionali, illustrando le ragioni del loro scarso utilizzo nella ricerca clinica.

Parole chiave. *Test unidirezionale, test bidirezionale, livello di significatività, valori critici del test.*

Summary

Two-sided and one-sided statistical tests

Methodological aspects of statistical tests used in comparative studies, previously described in CASCO 5 and 6, were synthetized to introduce the one-sided tests, more powerful than two-sided tests, highlighting the reasons of their not frequent use in clinical research.

Key words. *One-sided test, two-sided test, significance level, critical values of a statistical test.*

L'argomento riguarda tutti i tipi di test statistici che man mano sono stati descritti dal n. 5 al n. 11 di CASCO. Per facilitarne la comprensione, si consiglia di riguardare la costruzione dei test statistici esposta nel n. 6 di CASCO (estate 2013) e la logica del test statistico esposta nel n. 5 di CASCO. Comunque, per comodità del Lettore, riepiloghiamo i punti salienti della costruzione di un test statistico in modo tale che il contenuto della rubrica resti, come al solito, autosufficiente, nel senso che non richieda conoscenze pregresse.

Costruzione di un test statistico (sintesi)

1. Formulazione dell'ipotesi nulla, H_0. Negli esperimenti comparativi, l'ipotesi nulla è quella di uguale efficacia dei trattamenti. Siano A e B due trattamenti e P_A e P_B (parametri sconosciuti) le frequenze relative dei successi terapeutici nelle popolazioni target (cioè nelle popolazioni cui va riferito il risultato del test). Se i due trattamenti hanno la stessa efficacia, le percentuali di successi nelle due popolazioni sono uguali:

$$H_0 : P_A = P_B \quad \text{o anche} \quad H_0: P_A - P_B = 0.$$

2. Formulazione dell'ipotesi alternativa, H_1. L'ipotesi alternativa sia quella di differente efficacia; pertanto, se i

- Copyright - Il Pensiero Scientifico Editore downloaded by IP 3.17.74.227 Sat, 27 Apr 2024, 04:25:33

trattamenti non sono ugualmente efficaci, la percentuale dei successi nelle due popolazioni risulta diversa:

H_1: $P_A \neq P_B$, ossia H_1: $P_A - P_B \neq 0$.

3. Stima dei parametri. Le percentuali di successi nei due bracci di trattamento, f_A e f_B, costituiscono le migliori stime dei parametri. Pertanto, $f_A - f_B$ è la migliore stima di $P_A - P_B$.

4. Stimatore sotto H_0. Al variare del campione nell'universo dei campioni la stima del parametro varia. In altre parole, ripetendo lo stesso studio i pazienti arruolati sarebbero diversi e, poiché la risposta dipende non solo dal trattamento ma anche dal paziente, la differenza $f_A - f_B$ sarebbe diversa nei vari studi. L'insieme di tutti i possibili valori di una stima ha la natura di una variabile casuale e prende il nome di **stimatore** (v. Statistica per concetti in CASCO 5). Quindi, nel nostro caso, al variare del campione nell'universo dei campioni, $f_A - f_B$ varia e descrive una variabile casuale $F_A - F_B$, detta stimatore.

Se la numerosità dei due gruppi è sufficientemente grande ($n > 20$ per ciascun gruppo), si può dimostrare che lo stimatore $F_A - F_B$ è una variabile casuale normale (cioè rappresentabile con la curva di Gauss) con media $P_A - P_B$ ed un certo errore standard ES di cui, per semplicità, omettiamo la formula (si ricordi che ES è lo scarto quadratico medio dello stimatore e misura quanto ciascuna stima si discosta, in media, dalla media di tutte le stime).

Da questo punto in poi, il test prosegue come se l'ipotesi nulla (uguale efficacia dei trattamenti) fosse vera fino a giungere ad una conclusione che (probabilisticamente) la contraddica (in tal caso H_0 sarà respinta) o non la contraddica (in tal caso H_0 sarà accettata).

Quindi, se è vera l'ipotesi nulla, ES si trasforma nell'errore standard sotto H_0, ES_0, e la media dello stimatore, essendo pari a $P_A - P_B$, è uguale a 0.

5. Costruzione della statistica test. Una variabile casuale normale gode di due proprietà: è asintotica (nel senso che si avvicina indefinitamente all'asse delle ascisse man mano che ci si allontana dalla media) e simmetrica. Essa si trasforma nella variabile casuale normale standardizzata quando le si sottrae la media e la si divide per lo scarto quadratico medio (errore standard). Quindi, la variabile casuale normale standardizzata ha media 0 ed errore standard pari a 1.

Pertanto, essendo $f_A - f_B$ una determinazione di una normale (nel caso di grandi campioni), sottraendole la media e dividendo per lo scarto quadratico medio dello stimatore sotto H_0, si trasforma in una determinazione di una variabile casuale normale standardizzata.

Sotto l'ipotesi nulla, la media dello stimatore è uguale a 0 e l'errore standard è ES_0.

Pertanto, la quantità

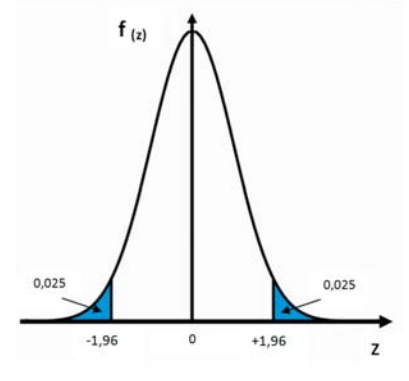
$$z = (f_A - f_B) / ES_0 \qquad [1]$$

è una determinazione di una normale standardizzata.

In altre parole, z è un qualsiasi punto dell'asse delle ascisse del grafico riportato nella figura 1.

Tenendo presente che le aree sotto la curva vanno interpretate come probabilità, si fissa una probabilità piccola di commettere un errore nel respingere l'ipotesi nulla (detta livello di significatività, generalmente il 5%, come nella figura 1).

Figura 1. ***Curva normale standardizzata nel caso di un test bidirezionale.***



Quindi, se è vera l'ipotesi nulla, la variabile normale standardizzata può assumere tutti i valori compresi tra $-\infty$ e $+\infty$, solo che quelli nelle code della distribuzione (tra $-\infty$ e $-1,96$ e tra $+1,96$ e $+\infty$) hanno complessivamente una probabilità piccola di presentarsi (il 5% nel grafico), mentre quelli tra $-1,96$ e $+1,96$ hanno un'alta probabilità di presentarsi (il 95%).

Collochiamo quindi nel grafico il valore z ottenuto con la [1]:

- Se cade tra $-1,96$ e $+1,96$ si accetta l'ipotesi nulla perché, sotto H_0, si è presentato uno degli eventi che avevano un'alta probabilità di presentarsi (il 95%); quindi non abbiamo elementi per respingerla. L'intervallo dell'asse delle ascisse compreso tra $-1,96$ e $+1,96$ si chiama **regione di accettazione**.
- Se cade tra $-\infty$ e $-1,96$ o tra $+1,96$ e $+\infty$ vuol dire che possono essersi presentate una delle due seguenti situazioni:
  a. l'ipotesi nulla è vera, ma si è presentato un evento raro, cioè uno degli eventi che complessivamente avevano il 5% di probabilità di presentarsi;
  b. l'ipotesi nulla è falsa.

Tra tali due possibilità si sceglie la seconda argomentando che, se l'ipotesi nulla fosse stata vera si sarebbe presentato un evento raro; ma, con pratica certezza, un evento raro non si presenta. La zona dell'asse delle ascisse compresa tra $-\infty$ e $-1,96$ e tra $+1,96$ e $+\infty$ si chiama **regione di rifiuto**.

I punti $-1,96$ e $+1,96$ si chiamano **valori critici del test** in quanto dividono la regione di accettazione da quella di rifiuto e, per convenzione, appartengono alla regione di rifiuto.

Pertanto, tra le due alternative sopra esposte, si sceglie la seconda, senza però dimenticare che poteva invece essere vera H_0 ed essersi presentato un evento raro. Quindi il 5% è la probabilità di sbagliare nel decidere di respingere l'ipotesi nulla

- Copyright - Il Pensiero Scientifico Editore downloaded by IP 3.17.74.227 Sat, 27 Apr 2024, 04:25:33

ed è detta **livello di significatività** (***significance level***) del test.

Test unidirezionali

Il test sopra esposto di chiama **bidirezionale** – o **a due code** – (***two-sided test***), nel senso che la regione di rifiuto è ugualmente ripartita nelle due code della distribuzione. In altre parole, se si respinge l'ipotesi nulla è perché il trattamento A è più efficace di B (se è $f_A > f_B$) o anche perché il trattamento B è più efficace di A (se è $f_A < f_B$). Quindi, il segno di z indica quale delle due situazioni sia da accogliere.

In alcune circostanze, però, una delle due situazioni può essere esclusa a priori (cioè, prima di eseguire il test). Se è così si può modificare l'ipotesi alternativa in modo più vantaggioso per il ricercatore (si ricordi che l'interesse del ricercatore è di respingere l'ipotesi nulla perché nel caso di accettazione non vorrebbe dire che H_0 è vera, ma solo che non ci sono evidenze per respingerla).

Ad esempio nel caso di sperimentazione contro placebo (indichiamolo con B), si può ritenere che l'effetto placebo sia incorporato in qualunque trattamento attivo. Il tal caso, l'ipotesi nulla è che il trattamento attivo (indicato con A) non abbia alcun effetto farmacologico, cioè abbia solo un effetto placebo. Pertanto, nell'ipotesi alternativa, sembra potersi escludere l'altra possibilità, cioè che sia $P_B > P_A$. Tale situazione si ha anche nel caso in cui ad un gruppo sperimentale è assegnato un trattamento che invece all'altro è negato (come nello studio esaminato in "Casi clinici" di questo numero).

In tali casi sembra potersi formulare un'ipotesi alternativa più vantaggiosa per il ricercatore:

$$H_1: P_A > P_B \text{ ovvero } H_1: P_A - P_B > 0$$

Tale ipotesi è quella che dà luogo ad un **test unidirezionale** (***one-sided test***) per la cui comprensione introduciamo la figura 2 che invitiamo il lettore di confrontare con la figura 1.

Come si può osservare, avendo escluso la possibilità che sia $P_A - P_B < 0$, il livello di significatività si concentra nella sola coda di destra e, pertanto, il valore critico del test diminuisce (passando da + 1,96 a + 1,645) rendendo più facile respingere l'ipotesi nulla.

Ad esempio, se avessimo ottenuto un valore della statistica-test pari a + 1,80, nel caso di un test bidirezionale risulterebbe non significativo (è inferiore a + 1,96), mentre con un test unidirezionale si giungerebbe ad una conclusione opposta (è superiore a + 1,645).

A parità delle altre circostanze il test unidirezionale è più potente di quello bidirezionale, ma, ricordiamolo, lo si può usare solo a condizione di poter escludere, prima di eseguirlo, una delle due alternative (quindi, mai lo si può usare quando si confrontano due trattamenti attivi).

Osservando la figura 2, nel caso di un test unidirezionale, la regione di accettazione va da $-\infty$ a + 1,645 (escluso) e quella di rifiuto va da + 1,645 (incluso) a $+\infty$. Quindi, se anche il risultato della statistica-test fosse pari a – 4 (un valore che per un test bidirezionale condurrebbe a respingere l'ipotesi nulla ad un livello di significatività pari a $P < 0{,}001$), l'ipotesi nulla dovrebbe essere accettata, sebbene con qualche perplessità.

Nelle applicazioni cliniche, un test unidirezionale si potrebbe usare solo quando si sperimenta verso placebo o quando ad un braccio si somministra un trattamento che all'altro è negato.

Nel secondo caso, come già sottolineato in "Casi clinici", il confronto di un trattamento attivo verso nessun trattamento non è accettabile perché, in caso si riscontri una significativamente maggior efficacia nel braccio di trattamento rispetto a quello di controllo, si potrebbe sempre pensare che essa sia riconducibile all'effetto placebo del farmaco e non ad una sua reale attività farmacologica. Comunque, non si dovrebbe pensare all'uso di un test unidirezionale, soprattutto in presenza di una forte selezione dei pazienti (come nello studio riportato nella scheda), in quanto potrebbe accadere che vi siano effetti indesiderati del trattamento che possano modificare la risposta – o essere modificati dalla risposta – così che l'uscita dei pazienti che ne soffrano possa giungere a privilegiare il gruppo di controllo. In tal caso (e forse anche nel caso di altri effetti di selezione) non si può escludere a priori che il trattamento di controllo risulterà nei fatti significativamente più efficace di quello sperimentale.

Nel caso di sperimentazione contro placebo, l'uso dei un test unidirezionale è comunque sconsigliato soprattutto perché, se i pazienti sono valutati in accordo al principio di intenzione a trattare (come viene quasi sempre richiesto ad uno studio clinico), se per la tossicità del trattamento molti pazienti sospendono la sua assunzione, nel gruppo di controllo tutti i pazienti ricevono un placebo, mentre in quello sperimentale molti pazienti non ricevono alcun trattamento e, pertanto, al termine dello studio nel braccio di trattamento si potrebbero riscontrare risultati significativamente peggiori di quelli osservati nel gruppo di controllo.

In conclusione, nella ricerca clinica l'uso di un test unidirezionale è quasi sempre da sconsigliare. Invece, esso trova utili impieghi in altri settori di ricerca (ad esempio, nelle applicazioni della Statistica all'analisi dei processi produttivi).

Enzo Ballatori

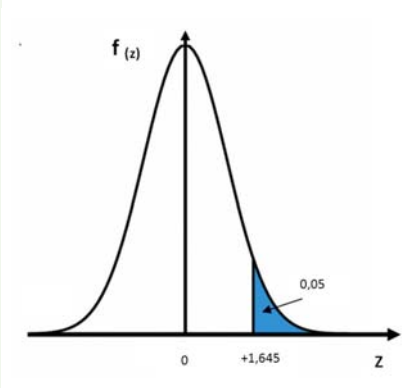



Figura 2. ***Curva normale standardizzata nel caso di un test unidirezionale.***