cune dimensioni della vita quotidiana, ma non tiene conto della sua durata, che, invece, ci sembra un elemento centrale per una sua appropriata valutazione. Un dolore assai severo sofferto per alcuni minuti, per il paziente, potrebbe essere più tollerabile di un dolore di media intensità che però si protrae a lungo.

In conclusione, i vantaggi del denosumab inducono a considerarlo più efficace ed accettabile dell'acido zoledronico nel trattamento delle metastasi ossee nei pazienti con carcinoma mammario, nelle condizioni previste dal protocollo. Il lavoro presentato nella scheda 2 sembra non aggiungere nulla, anche perché i pochi importanti risultati ottenuti sarebbero potuti essere agevolmente integrati nello studio descritto nella scheda 1. •

Statistica per concetti

La costruzione del test statistico nel modello di popolazione

Nel n. 5 di CASCO (3/2012), in questa stessa rubrica, è stata esposta la logica del test statistico, che preghiamo il lettore di riguardare perché l'articolo contiene concetti che, per motivi di spazio, non possono qui essere ripetuti per esteso, ma saranno riassunti all'inizio per rendere comprensibile il testo anche a coloro che non avessero sotto mano l'articolo precedente.
In questo numero ci occuperemo della costruzione del test statistico, che, per quasi tutti i test, segue la stessa impostazione. Per semplicità, tratteremo del problema del confronto tra due frequenze, ma, mutatis mutandis, quanto esposto potrà essere riferito anche ad altre situazioni, come, ad esempio, il test sul valore di una frequenza, particolarmente utile negli studi di fase 2, il test per il confronto tra due medie, e così via. La descrizione della sua costruzione consente, inoltre, alcune riflessioni su altri aspetti della logica del test che completano il quadro delineato nel numero precedente.

Sintesi dei concetti di base.

A e B siano i **trattamenti** a confronto in uno studio randomizzato di superiorità.

Esistono due popolazioni target, quella dei pazienti presenti e futuri che saranno trattati con A e quella dei pazienti che riceveranno B. L'interesse dello studio è riferire i risultati a tali popolazioni per decidere sulla diversa efficacia/tollerabilità dei due trattamenti.

La risposta sia binaria: successo o insuccesso terapeutico.

P_A e P_B sono i **parametri** (numeri sconosciuti), nel nostro caso le frequenze relative di successi nelle due popolazioni target. Poiché si tratta di uno studio comparativo, il parametro oggetto di interesse è $P_A - P_B$.

Da ciascuna delle due popolazioni target si estrae un campione costituito dal gruppo dei pazienti trattati con A e da quelli trattati con B.

f_A e f_B, stime dei parametri, sono le frazioni di successi osservate nel braccio A e nel braccio B, rispettivamente. Pertanto, la migliore stima del parametro $P_A - P_B$ è $f_A - f_B$.

La risposta al trattamento dipende dal trattamento, ma anche dal paziente, per cui, se ripetessimo lo studio su altri pazienti, otterremmo risultati differenti, semplicemente perché i pazienti sono diversi.

Immaginiamo l'insieme di tutti i possibili campioni diversi estratti a sorte dalle due popolazioni target (universo dei campioni). Al variare del campione nell'universo dei campioni le stime variano. La variabilità di tali stime si può calcolare per mezzo del Calcolo delle Probabilità.

La stima $f_A - f_B$ varia dunque da campione a campione e lo scarto quadratico medio della loro distribuzione prende il nome di **Errore Standard (ES)** e misura quanto ciascuna stima sia diversa, in media, dal valore del parametro $P_A - P_B$.
In altre parole, l'errore standard è una misura di quanto ci si attende che la stima, calcolata su ciascun campione dell'universo dei campioni, sia diversa, in media, dal valore del parametro per puro effetto del caso.

La costruzione di ogni test statistico procede toccando sempre i seguenti punti.

0. Descrizione delle popolazioni

Si esplicita tutto ciò che si conosce sulle popolazioni target.

Esempio. Riprendiamo l'esempio considerato nell'articolo precedente. I parametri sono le percentuali di pazienti che sono protetti dal vomito nella popolazione di pazienti presenti e futuri (nelle condizioni previste dallo studio) trattati con olanzapina (OLA, trattamento A), P_A, e nella popolazione di pazienti trattati con metoclopramide (METO, trattamento B), P_B. Tali parametri sono sconosciuti.

Le popolazioni target sono composte da pazienti presenti e futuri, trattati con chemioterapia altamente emetogena, che, malgrado la profilassi antiemetica standard ricevuta, riprendano a vomitare. Al primo episodio di vomito, i pazienti sono trattati con OLA (popolazione A) o METO (popolazione B). La risposta al trattamento è l'interruzione del vomito, nel senso che, nel periodo di osservazione, un paziente non vomiti più, o che invece continui a vomitare. La risposta è dunque binaria ed i parametri di interesse sono le percentuali di pazienti che, nelle due popolazioni, sono protetti dal vomito, P_A e P_B. Poiché lo studio è comparativo, vi è un unico parametro oggetto di inferenza: $P_A - P_B$.

1. *Formulazione delle ipotesi*

Questo primo punto riguarda unicamente i parametri. Si formula l'ipotesi nulla, di uguale efficacia dei trattamenti

$$H_0: P_A = P_B$$

osservando che una formulazione equivalente di H_0 è $P_A - P_B = 0$.

Si formula l'ipotesi alternativa, di differente efficacia dei trattamenti

$$H_1: P_A \neq P_B$$

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

Esempio (prosec.). L'ipotesi nulla è l'ipotesi di uguale efficacia dei trattamenti: se OLA avesse la stessa efficacia di METO, nelle due popolazioni target la percentuale dei pazienti che non vomitano più avendo assunto il farmaco sarebbe la stessa. L'ipotesi alternativa è quella di diversa efficacia: se i trattamenti avessero una differente efficacia, nelle popolazioni target la percentuale dei pazienti che non vomitano più sarebbe diversa.

2. *Stime dei parametri*

I due gruppi di pazienti sottoposti ai due trattamenti, A e B, possono essere riguardati come campioni casuali (cioè estratti a sorte) provenienti dalle due popolazioni target[1]. Le percentuali di successi terapeutici sono pertanto le migliori stime dei due parametri (sconosciuti) P_A e P_B. Pertanto, $f_A - f_B$ costituisce la migliore stima di $P_A - P_B$.

Esempio (prosec.). Dal campione di pazienti osservati si possono ottenere le stime: $f_A = 71\%$ (30/42) e $f_B = 32\%$ (12/38). La migliore stima di $P_A - P_B$ è $f_A - f_B = 0{,}71 - 0{,}32 = 0{,}39$: nel gruppo dei pazienti trattati con OLA è stato riscontrato il 39% di successi terapeutici in più rispetto al gruppo dei pazienti che hanno ricevuto METO.

3. *Stimatore sotto l'ipotesi nulla*

Dal Calcolo delle Probabilità è noto che, se le numerosità dei campioni sono sufficientemente grandi (maggiori di 20), al variare del campione nell'universo dei campioni, $f_A - f_B$ varia e descrive una variabile casuale normale (curva di Gauss) $F_A - F_B$ di media $P_A - P_B$ ed un certo ES (per semplicità, ne omettiamo la formula, inessenziale per la descrizione della logica del test).

Da questo punto in poi, procediamo come se l'ipotesi nulla fosse vera, salvo trovare, alla fine del ragionamento, una contraddizione che ci autorizzi a ritenere i due trattamenti diversamente efficaci. La logica è simile a quella della dimostrazione per assurdo di alcuni teoremi di geometria: si formula un'ipotesi e si procede con una catena di deduzioni che, se porta ad una contraddizione, induce a dichiarare falsa l'ipotesi da cui si era partiti. L'unica differenza tra tale procedimento e la logica del test statistico è che, nel secondo caso, la contraddizione è su base probabilistica, mentre è certa quella relativa al teorema di geometria.

Se è vera H_0, la media dello stimatore $F_A - F_B$ è uguale a 0, in quanto coincide con il valore del parametro. Quindi, l'ES, sotto l'ipotesi nulla, **ES°**, misura quanto ciascuna stima $f_A - f_B$ sia, in media, diversa da 0. In altre parole, ES° è una misura della variabilità attesa delle stime nel caso in cui i due trattamenti siano ugualmente efficaci.

4. *Costruzione della statistica test*

Ovviamente, se è $f_A - f_B = 0$, accettiamo subito l'ipotesi nulla senza bisogno di costruire il test. Consideriamo, quindi, il caso in cui vi sia una differenza tra le stime di successi nei due bracci di trattamento.

Da quanto esposto al punto precedente, il criterio decisionale si attua per mezzo della statistica test che consiste nel rapportare la differenza riscontrata sul campione con una stima dell'errore standard sotto H_0 (che misura, lo ripetiamo, quanto è diversa ciascuna stima, in media, da 0 per puro effetto del caso):

$$z = (f_A - f_B) / ES°.$$

Se è vera l'ipotesi nulla, ci si attende che z, in valore assoluto, sia all'incirca uguale a 1, perché numeratore e denominatore sono stime, rispettivamente, della differenza empirica riscontrata nel campione e della differenza teorica, cioè quella attesa se l'ipotesi nulla fosse stata vera.

Se, invece, in valore assoluto, z risulta molto maggiore di 1, allora qualcosa deve essere intervenuto a produrre una differenza empirica così tanto più grande di quella attesa. Poiché l'unico effetto sistematico che si sta controllando è il trattamento, non ci resta che respingere H_0 concludendo che i due trattamenti hanno una differente efficacia. Ma tale decisione non ha fondamento di certezza, per cui, nel respingere H_0, si ha una certa (seppur bassa) probabilità di sbagliare. Tale probabilità si chiama **livello di significatività** del test e, convenzionalmente, si fissa al 5%, così che la conclusione che i due trattamenti siano diversamente efficaci ha una probabilità al più

1. Si ricordi che tale ipotesi, nella ricerca clinica, è verosimile solo a condizione che i pazienti siano arruolati consecutivamente, altrimenti nulla di quanto esposto sarebbe applicabile.

uguale al 5% di essere sbagliata. Se la differenza è trovata significativa al 5% (ossia se il valore assoluto di z della statistica test è superiore a 1,96)[2], conviene fornire il valore esatto del livello di significatività (che si indica con "P <"), così che, essendo più basso del 5%, rassicuri maggiormente sulla possibilità di commettere un errore nel respingere l'ipotesi nulla.

Esempio (prosec.). Nell'esempio, la stima dell'ES sotto l'ipotesi nulla è ES° = 0,112. In tal modo la statistica test è z = 0,39/0,112 = 3,48, dove, a numeratore c'è la differenza tra la frequenza relativa dei pazienti protetti con OLA e quella dei pazienti protetti con METO. Sotto l'ipotesi nulla, la differenza osservata è oltre 3 volte superiore a quella attesa e ciò induce a respingere l'ipotesi di uguale efficacia dei due trattamenti, con una probabilità di sbagliare inferiore all'un per mille: P < 0,001. Ovviamente, dal segno positivo di z si evince che OLA è più efficace di METO (altrimenti, il segno di z sarebbe risultato negativo).

2. Il valore 1,96 è desunto dalle tavole della normale standardizzata, ovvero è fornito dal computer: la probabilità che, se è vera l'ipotesi nulla, il valore (assoluto) della statistica test sia maggiore o uguale a 1,96 è pari al 5%, nel senso che, se i due trattamenti sono ugualmente efficaci, la probabilità di sbagliare nel dichiararli diversamente efficaci è non superiore al 5%, e, quindi, costituisce un evento raro che, come tale, con pratica certezza, non si presenta.

Discussione

1. Il test descritto (noto come z-test) è equivalente al test χ^2 (chi-quadrato), nel senso che entrambi i test conducono sempre a prendere la stessa decisione.
2. Sia la procedura seguita, sia il chi-quadrato sono test validi solo a condizione che nessuna delle frequenze osservate sia molto piccola (vi sono 4 frequenze: successi o insuccessi terapeutici nei pazienti trattati con A e in quelli sottoposti a B). Vi è sufficiente accordo nel ritenere che le 4 frequenze debbono essere tutte maggiori di 5. Se una frequenza fosse minore di 5 sarebbe opportuno usare il test "esatto" di Fisher, di cui sia lo z-test che il χ^2, sono approssimazioni.
3. Nella ricerca clinica il test esposto si usa anche per la valutazione della sicurezza differenziale dei due trattamenti, ossia per valutare se l'incidenza di ciascuno degli eventi avversi considerati è significativamente diversa nei due bracci di trattamento.
4. L'ES della differenza tra due frequenze dipende anche dalle numerosità dei due campioni: quanto più queste sono elevate, tanto minore risulta ES°. Quindi, con campioni molto grandi risultano significative differenze anche molto piccole, mentre, con basse numerosità si corre il rischio che una differenza di efficacia osservata ($f_A - f_B$) anche molto grande non sia trovata significativa.
5. Come si è visto, ogni test statistico si basa sull'assunto (importantissimo) che tutte le circostanze sistematiche differenti dal trattamento siano ben bilanciate tra i due gruppi di pazienti. La randomizzazione evita sbilanciamenti sostanziali tra i due bracci rispetto a tutti i fattori sistematici, noti e sconosciuti, così che l'unico effetto sistematico che resta (quello che il test controlla) è la diversità dei trattamenti (per approfondimenti sull'argomento, riguardare la stessa rubrica nel n. 5 di CASCO, 3/2012).

Enzo Ballatori