se tali variabili siano state considerate in quanto tali o, invece, trasformate in fattori, v. scheda), per un totale di 9 tra fattori e variabili considerati. Non solo in tale situazione il proportional hazard diventa arduo persino da ipotizzare, ma sarebbe stato sufficiente dare un'occhiata alle corrispondenti curve relative ai due trattamenti per vedere una quasi perfetta sovrapposizione che avrebbe esonerato da qualsiasi altra analisi.

Conclusioni

Nella programmazione degli studi clinici e nell'analisi dei risultati, talvolta la Statistica viene utilizzata in maniera esagerata, quasi per nascondere, con la complessità degli strumenti adottati, la povertà dei risultati conseguiti. Il lavoro sintetizzato nella scheda e gli altri due sullo stesso argomento ci sembrano di questo tipo.

C'è una lunga catena che va dai medici ricercatori che accettano di firmare i protocolli proposti dall'industria, ai comitati etici che approvano i protocolli, alle riviste scientifiche che pubblicano i risultati, alle autorità regolatorie che approvano i farmaci, alle Consensus Conference che introducono i nuovi farmaci nelle linee guida, al medico che li prescrive; tutti i punti nodali di tale catena presentano problemi di varia natura. Dati gli scopi di questa rubrica, ci soffermeremo brevemente solo sul sistema di referaggio delle riviste.

Essendo i tre lavori considerati tutti pubblicati su riviste assai importanti (*Lancet* e *JCO*), sapendo che ciascuna di loro fa valutare ogni manoscritto da tre referee, di cui uno statistico, desta meraviglia il fatto che di nove (o poco meno) referee diversi che hanno analizzato i tre studi nessuno abbia avanzato obiezioni sostanziali, o che comunque queste non siano state considerate dall'Editor nel prendere la decisione finale di pubblicare il lavoro. Il lettore non solo dovrebbe sapere chi ha approvato la pubblicazione (in particolare se è stata un'iniziativa del solo *editor in chief* o se anche tutti o in parte i referee erano d'accordo), ma soprattutto dovrebbe conoscere i contenuti delle revisioni dei referee per comprendere meglio i possibili limiti dello studio (e forse anche dei referee!), onde poterne dare una propria valutazione critica. Ma la pubblicazione di tali revisioni (magari solo on-line) non è la sola cosa che andrebbe cambiata nell'attuale politica editoriale delle riviste scientifiche, anche dato il ruolo, sempre più pesante, rivestito dallo sponsor nella ricerca clinica (v. scheda, ultime righe), per non parlare degli abbondanti conflitti di interessi degli autori (indicati nell'articolo, ma non riportati nella scheda). •

Bibliografia

1. *Stopek AT, Lipton A, Body JJ, et al. Denosumab compared with zoledronic acid for the treatment of bone metastases in patients with advanced breast cancer: a randomized, double-blind study. JCO 2010; 28: 5132-9.*
2. *Henry DH, Costa L, Goldwasser F, et al. Randomized, double-blind study of denosumab versus zoledronic acid in the treatment of bone metastases in patients with advanced cancer (excluding breast and prostate cancer) or multiple myeloma. JCO 2011; 29: 1125-32.*

Statistica per concetti

Test statistici per dati appaiati

Riassunto

I test statistici per dati appaiati sono esposti in relazione alla scala in cui si colloca la risposta al trattamento. In particolare, sono descritti il test di Mc Nemar (scala nominale), il test di Wilcoxon per dati appaiati (scala ordinale) e il t-test per il confronto tra due medie nel caso di dati appaiati (scale di rapporti).

Parole chiave. *Disegno cross-over, dati appaiati, test di Mc Nemar, ranghi, test di Wilcoxon per dati appaiati, t-test per dati appaiati.*

Summary

Statistical tests for paired data

Statistical tests for paired data are described in relationship with the scale used to evaluate the response to the treatment. More precisely, we described Mc Nemar test (nominal scale), Wilcoxon matched pairs, signed ranks test (ordinal scale), t-test for paired data (ratio scale).

Key words. *Cross-over design, paired data, Mc Nemar test, Wilcoxon matched pairs signed ranks test, t-test for paired data.*

Si generano dati appaiati (*paired data*) quando uno stesso carattere (qualitativo o quantitativo) è rilevato due (o più) volte sulle stesse unità.

Il vantaggio di operare con dati appaiati consiste nel tenere costanti tutti i fattori (prognostici, predittivi, di rischio) che possono interferire con l'effetto del trattamento sulla risposta.

Nella ricerca clinica si ha spesso a che fare con dati appaiati, quando, ad esempio, si valuta la riproducibilità di un nuovo test diagnostico, per cui gli stessi pazienti sono esaminati da due osservatori indipendenti usando lo stesso strumento, o quando un test psicometrico viene somministrato due volte agli stessi soggetti per valutarne la riproducibilità. Ma l'esempio di dati appaiati più comune è in relazione al disegno cross-over, che pertanto è assunto come applicazione di riferimento nell'esposizione che segue.

Disegno cross-over

Siano A e B i due trattamenti a confronto (ma quanto esposto è generalizzabile al caso di più di due trattamenti). Il disegno cross-over consiste nel somministrare, in tempi diversi, entrambi i trattamenti agli stessi pazienti, valutando così due risposte in ogni paziente: una dopo la somministrazione di A, l'altra dopo quella di B.

Scopo del disegno dello studio è cercare di annullare (o ridurre) l'effetto dei fattori diversi dal trattamento sulla risposta (v. CASCO 9, Statistica per concetti 1), in modo tale che il confronto finale dipenda solo dalla diversa efficacia dei trattamenti (e dal caso, che però può essere controllato mediante gli strumenti dell'inferenza statistica, come il test o gli intervalli di confidenza). Quando siano verificate alcune assunzioni, il disegno cross-over è il più efficiente perché il confronto tra l'efficacia dei due trattamenti può essere eseguito nello stesso paziente (tenendo, quindi, costanti tutti i fattori che, oltre al trattamento, possono influenzare la risposta). Tale maggior efficienza si traduce, a parità delle altre condizioni, in una potenza più elevata che consente di ottenere i risultati programmati con un minor numero di pazienti (tanto per fissare le idee con dati inventati: se per ottenere un certo risultato con uno studio parallelo randomizzato occorressero 360 pazienti, con un disegno cross-over potremmo arruolarne molti meno, ad es., solo 140).

Vi è un non trascurabile secondo vantaggio nel disegno cross-over: è l'unico modo per stabilire quale dei due trattamenti sia più gradito al paziente.

Tecnicamente, alla metà dei pazienti si somministra prima il trattamento A e poi, dopo un certo periodo in cui il paziente viene lasciato privo di trattamento, il trattamento B (l'intervallo di tempo tra i due trattamenti è noto come periodo di *wash out*, importante per impedire, almeno in parte, che l'effetto del primo trattamento si trascini fino a modificare l'effetto del secondo). All'altra metà dei pazienti si somministra la sequenza BA, randomizzando i pazienti a ricevere AB o BA, per evitare che l'effetto "primo trattamento" possa essere sbilanciato a favore di uno dei due trattamenti.

Ovviamente, vi sono dei casi di impossibilità dell'uso del disegno cross-over, come nelle malattie acute o quando la risposta è in termini di sopravvivenza.

Il tipo di test per dati appaiati dipende, ovviamente, dalla scala di misura usata per valutare la risposta (v. CASCO 9, Statistica per concetti 2).

a. Scale nominali, carattere dicotomico.

La risposta è valutata in termini di successo (+) e insuccesso (–) terapeutico. Indicando con A e B i due trattamenti a confronto, i risultati dello studio possono essere rappresentati nella seguente tabella:

A↓ B→	+	–	Tot.
+	a	b	n_1
–	c	d	n_2
Tot.	m_1	m_2	N

dove "a" indica il numero di pazienti in cui si è osservato un successo in seguito alla somministrazione di entrambi i trattamenti (+, +), "d" il numero di pazienti che hanno avuto un insuccesso con entrambi i trattamenti (–, –), "b" il numero di pazienti in cui si è registrato un successo con A e un insuccesso con B (+, –), "c" il numero di pazienti che hanno presentato un insuccesso con A e un successo con B (–, +).

Esempio 1: controllo del vomito acuto in due successivi cicli di chemioterapia. Siano A e B i trattamenti antiemetici a confronto. Al termine dello studio i risultati sono i seguenti

A↓ B→	+	–	Tot.
+	60	34	94
–	10	20	30
Tot.	70	54	124

in 60 pazienti, entrambi i trattamenti sono stati in grado di prevenire il vomito e in 20 pazienti entrambi i trattamenti hanno fallito. In 34 pazienti si è osservato un successo con A e un insuccesso con B e in 10 pazienti è accaduto il contrario (successo con B, insuccesso con A).

Il test di Mc Nemar (1947) è lo strumento per provare se i due trattamenti hanno una diversa efficacia. Mc Nemar ritenne non informativi sull'efficacia differenziale dei due trattamenti i pazienti con segno uguale [(+, +) o (–, –)] che pertanto vanno esclusi. Restano così da considerare s = b + c pazienti con risultato discordante fra i due trattamenti [cioè, (+, –) o (–, +)]. L'ipotesi nulla (H_0) è quella di uguale efficacia dei trattamenti; pertanto, se è vera l'ipotesi nulla, la probabilità che uno degli s pazienti che presentano segni discordanti (i soli ad essere informativi ai fini della valutazione di efficacia differenziale fra i trattamenti) cada nella casella "b" è la stessa che cada in "c". In altre parole, sotto l'ipotesi nulla, un soggetto con segni discordanti ha probabilità pari a ½ di cadere in "b" e pari a ½ di cadere in "c". Il problema si riduce allora a valutare (tra gli s soggetti con segni discordanti) se lo squilibrio tra "b" e "c" possa ritenersi dovuto al caso o se, invece, è dovuto, con alta probabilità, alla diversa efficacia dei trattamenti (nell'esempio 1, occorre decidere se la differenza tra 34 e 10 possa essere attribuita al caso, o se, invece, è dovuta alla maggior efficacia di A). Per rispondere a tale quesito è sufficiente usare il test sul valore di una frequenza, cioè il test binomiale, o, nel caso dei grandi campioni, anche la sua approssimazione con la normale.

b. Scale ordinali

Le scale ordinali sono costituite da attributi (aggettivi) ordinabili in cui viene classificata la risposta osservata su ciascun paziente (per il concetto di rango, v. Statistica per concetti 2, CASCO 9). Quando la risposta è collocabile su una scala ordinale, l'analisi dei risultati di uno studio cross-over viene condotta mediante il

"test di Wilcoxon per dati appaiati" (*Wilcoxon matched-pairs signed ranks test*), che è un test basato sui ranghi (v. Statistica per concetti, CASCO 10) e prova l'uguaglianza delle mediane nelle popolazioni target. Come si ricorderà, "rango" indica il posto occupato dall'attributo considerato nella graduatoria, cioè nella distribuzione ordinata, ad es. in senso crescente.

Esempio 2: misura della severità della nausea. L'intensità della nausea può essere misurata in vari modi, ma il più semplice è una scala di Likert che consente di ancorare alla realtà la percezione del paziente: 0 = no nausea, L = nausea lieve (consente al paziente di svolgere tutte le sue attività quotidiane), M = nausea moderata (non permette al paziente di svolgere alcune delle sue abituali attività, S = nausea severa (il paziente è costretto a letto per la nausea).

La severità della nausea percepita dal paziente esplora una dimensione diversa dalla protezione dalla nausea e non può che riferirsi ai soli pazienti che hanno avuto nausea.

Come esempio didattico consideriamo uno studio con disegno cross-over in cui i pazienti, sottoposti a due cicli consecutivi di chemioterapia, hanno ricevuto in sequenza due diverse profilassi antiemetiche: A in un ciclo e B nell'altro. Si vuole valutare se negli 8 pazienti che hanno sofferto di nausea, questa sia stata più severa tra coloro che hanno ricevuto A o in quelli sottoposti a B. Le risposte ai trattamenti osservati negli 8 pazienti che hanno sofferto di nausea sono state le seguenti

Paziente	A	B
1	L	M
2	M	L
3	L	S
4	M	S
5	M	S
6	M	M
7	L	S
8	L	S

Il primo passo nella costruzione del test consiste nel trasformare gli attributi in ranghi (si tratta di assegnare in tutto 16 ranghi). Considerando che, in totale, gli attributi "L" (i più piccoli) sono 5, a ciascuno di essi si attribuisce lo stesso rango, pari alla media dei ranghi che avremmo dovuto assegnare (da 1 a 5: media 3). Gli "M" sono 6 cui avremmo dovuto attribuire i ranghi da 6 (i primi 5 li abbiamo assegnati a "L") a 11; quindi a ciascun M si assegna il rango medio 8,5 [=(6+7+8+9+10+11)/6]. Gli "S" sono 5; ad essi dobbiamo attribuire i ranghi da 12 a 16, media: 14.

Pertanto la tabella con i ranghi al posto degli attributi è la seguente:

Paziente	A	B
1	3	8,5
2	8,5	3
3	3	14
4	8,5	14
5	8,5	14
6	8,5	8,5
7	3	14
8	3	14

Il test di Wilcoxon per dati appaiati inizia con l'escludere le unità che presentano la stessa coppia di ranghi in quanto non informative sull'efficacia differenziale dei trattamenti (nell'esempio, il paziente 6 viene escluso; restano così 7 pazienti per ciascuno dei quali si calcola la differenza tra i ranghi, d (calcolando, ad es., la differenza tra il rango che compare nella colonna B e quello che compare in A: B – A).

Si determina quindi il rango di tali differenze, D = rango di d (indipendentemente dal segno), riattribuendogli poi il segno della differenza (da cui il nome di *signed ranks*).

Si invita il lettore a seguire le operazioni descritte nella tabella riportata nel seguente esempio.

Esempio 2, prosec. Nella tabella successiva, relativa ai 7 "pazienti informativi" dell'efficacia differenziale dei trattamenti, sono riportate le differenze (con segno) dei ranghi, d (= B – A).

L'ultima colonna (D = rango di d) si ottiene dalla colonna "d", attribuendo i ranghi alle quantità che compaiono in "d": le quantità più piccole (5,5) sono in numero di 4; pertanto attribuiamo a ciascun valore "5,5" lo stesso rango, pari alla media dei ranghi che avremmo dovuto attribuire se fossero stati diversi: (1+2+3+4)/4 = 2,5. Di "11" ce ne sono 3, che occupano nella graduatoria il 5°, il 6° e il 7° posto (i primi 4 sono occupati da "5,5"); quindi assegnamo ad essi lo stesso rango pari alla media dei ranghi che avremmo attribuito se fossero stati diversi: (5+6+7)/3 = 6. Infine, ai ranghi riportati nella colonna "D" riassegnamo i segni dei valori che compaiono nella colonna "d".

Si calcola, infine, la somma dei ranghi di d (riportati in D), negativi (–2,5) e positivi (25,5). La minore delle due costituisce la statistica-test necessaria per eseguire il test di Wilcoxon per dati appaiati.

Paziente	A	B	d	D = rango di d
1	3	8,5	5,5	2,5
2	8,5	3	-5,5	-2,5
3	3	14	11	6
4	8,5	14	5,5	2,5
5	8,5	14	5,5	2,5
7	3	14	11	6
8	3	14	11	6

Sotto l'ipotesi nulla di uguale efficacia dei trattamenti ci si attende che la somma dei ranghi positivi sia all'incirca pari a quella dei ranghi negativi (nell'esempio, come si vede, vi è un forte squilibrio, ma non sappiamo se possa comunque essere attribuito al caso o, invece, dipenda dalla differente efficacia dei trattamenti). Si procede, quindi, al calcolo del test di Wilcoxon per dati appaiati e alla conseguente individuazione del livello di significatività. Il calcolo del test è alquanto noioso ed va oltre gli scopi della presente nota, per cui si suggerisce di procedere in uno dei modi seguenti:

a. nel caso di piccoli campioni, si possono utilizzare le tavole che sono riportate in tutti i libri che trattano di test non parametrici (ad es., Sidney Siegel, Nonparametric statistics, Mc Graw Hill-Kogakusha, 1956). Entrando in tale tavola con la statistica test (–2,5) calcolata come sopra esposto, malgrado sia piccolo il numero di pazienti considerati (N = 7) nell'es. 2 il trattamento B può essere considerato più efficace di A nel ridurre la severità della nausea nei pazienti che ne hanno sofferto ad un livello di significatività del 5%.
b. nel caso di grandi campioni si può utilizzare l'opportuna approssimazione con la distribuzione normale di semplicissimo calcolo (v., ad es., Siegel 1956);
c. avvalersi di un software statistico (quasi tutti i packages statistici calcolano il test di Wilcoxon per dati appaiati e forniscono in automatico il livello di significatività).

c. Scale di rapporti (caratteri quantitativi).

In ogni paziente si rileva una coppia di risposte espresse da intensità.

Esempio 3. Consideriamo uno studio cross-over condotto su 8 pazienti ipertesi sottoposti a due trattamenti anti-ipertensivi, A e B. La risposta è misurata in termini di riduzione della pressione diastolica (mmHg) dopo 2 settimane di trattamento. I risultati dello studio, con le differenze, d, tra la risposta ad A e quella a B osservate in ciascun paziente siano i seguenti:

Paziente	A	B	d
1	18	12	6
2	4	4	0
3	8	10	-2
4	14	2	12
5	26	10	16
6	18	6	12
7	10	10	0
8	11	-1	12

Il paziente 1 ha avuto una riduzione di 18 mmHg con il trattamento A e di 12 con B. In tale paziente A si è dimostrato più efficace di B perché ha indotto una diminuzione della pressione diastolica superiore di 6 mmHg, e così via per gli altri pazienti. La differenza di efficacia a favore di A è risultata, in media, di 7 mmHg.

L'ipotesi nulla (H_0) è quella di uguale efficacia dei trattamenti. Quindi, se i trattamenti sono ugualmente efficaci, nelle due popolazioni target (definite dai pazienti presenti e futuri che verranno trattati con A e con B), le due medie sono uguali e, pertanto, la loro differenza è pari a 0. La media di d costituisce la migliore stima della differenza tra le medie nelle due popolazioni target. Se è vera H_0, ci si attende che la media di d sia prossima allo 0. Con i dati sopra riportati vale 7; occorre quindi provare se tale differenza possa essere attribuita al caso o se, invece, è dovuta al fatto che A è più efficace di B.

Il test più potente da usare in tale situazione è il test t di Student (t-test) per dati appaiati. Il t-test è un test parametrico perché si basa sull'assunzione di normalità dell'errore (v. CASCO 9, Statistica per concetti), ma resta valido anche per moderate violazioni di essa. Pertanto si può usare il t-test per dati appaiati solo quando si sia provato che il carattere si distribuisca normalmente mediante un test di normalità, oppure sia noto da fonti esterne che il carattere (nell'esempio, diminuzione della pressione diastolica) è distribuito all'incirca normalmente (quasi tutte le variabili biometriche lo sono).

In tutti gli altri casi occorre ricorrere al test di Wilcoxon per dati appaiati (test non-parametrico) costruendolo esattamente come esposto nell'esempio 2. In casi dubbi è sempre opportuno scegliere il *Wilcoxon matched-pairs signed ranks test* perché la sua potenza-efficienza (v. CASCO 10, Statistica per concetti) rispetto al t-test per dati appaiati è di circa il 95% (cioè, ove ricorrano le condizioni per l'uso del test parametrico, se usando il t-test per dati appaiati per provare una certa ipotesi occorrono 95 unità, con l'uso del test di Wilcoxon per dati appaiati ne occorrono 100: solo 5 di più).

Enzo Ballatori