I test di permutazione nell'individuazione dei fattori di rischio

Stefania Naddeo[1]
Dipartimento di Metodi Quantitativi, Università degli Studi di Siena

Riassunto: Nelle indagini statistiche dirette ad individuare i fattori di rischio associati ad una determinata patologia, i dati campionari raccolti vengono usualmente utilizzati per la verifica contemporanea di più ipotesi diverse, ognuna delle quali si riferisce all'incidenza di un singolo fattore.
In questa nota si valuta l'effetto della verifica di mancanza di incidenza di ciascun fattore con riferimento all'aumento nella probabilità di rifiutare almeno una di queste ipotesi quando sono vere. Il valore di questa probabilità, che è noto quando le statistiche test utilizzate per la verifica delle ipotesi sono indipendenti, è stato calcolato in modo approssimato per differenti livelli di dipendenza fra i fattori mediante una procedura di permutazione.
Si propone inoltre una procedura che consente di effettuare le verifiche controllando il livello di significatività globale e se ne valuta la potenza mediante uno studio di simulazione.
Parole chiave: Fattori di rischio, odds ratio, test multipli, livello di significatività, test di permutazione.

1. Introduzione

Nelle indagini statistiche in campo medico l'individuazione dei fattori di rischio legati a particolari patologie rappresenta uno degli argomenti di interesse primario. L'incidenza di un fattore in una determinata patologia viene generalmente valutata mediante il cosiddetto rischio relativo, che corrisponde al rapporto fra la probabilità di presentare la patologia in esame se si è soggetti al fattore considerato e la probabilità di presentare quella stessa patologia se, invece, non si è soggetti a tale fattore.

[1] Il presente lavoro è stato finanziato nell'ambito del progetto "La ricerca di determinanti del rischio mediante analisi di segmentazione di campioni", cofinanziato dal MIUR. Coordinatore nazionale è Luigi Fabbris, coordinatore del gruppo di Siena è Laura Carli Sardi.

Data una tabella di contingenza analoga a quella successiva, in cui si indica con M l'evento "l'individuo rilevato presenta la patologia in esame", con $\overline{\text{M}}$ l'evento complementare, con F l'evento "l'individuo rilevato presenta il fattore di rischio" e con $\overline{\text{F}}$ l'evento complementare, sia π_{11} la probabilità associata all'evento M∩F, π_{12} la probabilità di $\overline{\text{M}} \cap$F, π_{21} la probabilità di M$\cap \overline{\text{F}}$ e π_{22} quella di $\overline{\text{M}} \cap \overline{\text{F}}$.

Tabella 1
Classificazione degli individui a seconda della presenza/assenza della patologia e del fattore di rischio

	Patologia		
Fattore	M	$\overline{\text{M}}$	Totale
F	π_{11}	π_{12}	$\pi_{1.}$
$\overline{\text{F}}$	π_{21}	π_{22}	$\pi_{2.}$
Totale	$\pi_{.1}$	$\pi_{.2}$	1,0

Il rischio relativo Ω, dato dal rapporto

$$\Omega = \frac{\pi_{(1)}}{\pi_{(2)}} = \frac{\pi_{11}/\pi_{1.}}{\pi_{21}/\pi_{2.}} = \frac{\text{P}(\text{M} \mid \text{F})}{\text{P}(\text{M} \mid \overline{\text{F}})}, \qquad (1)$$

misura il rischio associato al fattore F. Se la (1) risulta pari ad uno, questo significa che al fattore F non è associato alcun rischio, se Ω risulta maggiore di uno F è un fattore il cui rischio è pari al $100(\Omega-1)$%, mentre se è inferiore ad uno F è un fattore protettivo e la diminuzione del rischio è pari a $100(1-\Omega)$% (cfr. AA.VV., 1991).

Nelle situazioni reali, però, spesso i dati campionari raccolti non consentono di stimare i valori delle probabilità $\pi_{(1)}$ e $\pi_{(2)}$. In generale, infatti, le informazioni disponibili derivano dalle cosiddette indagini *caso-controllo*, nelle quali i dati relativi agli individui che presentano una determinata patologia (casi) vengono rilevati su una popolazione di soggetti malati (per esempio su quella di coloro che si rivolgono ad una determinata struttura sanitaria), mentre i dati relativi ad individui sani (controlli) vengono rilevati su una popolazione di individui sani dalla quale si selezionano soggetti con le stesse caratteristiche sociali e demografiche dei malati.

Come si vede, non si dispone di un campione casuale proveniente dalla popolazione complessiva, ma di due campioni separati provenienti dalle due diverse popolazioni. In questa situazione la numerosità campionaria complessiva n è data dalla somma degli $n_{.1}$ soggetti malati e degli $n_{.2}$ soggetti sani. Gli individui dei due gruppi vengono poi classificati in base alla presenza/assenza del fattore di rischio, come nella tabella 2.

Tabella 2
Distribuzioni condizionate del fattore
a seconda della presenza/assenza della patologia

	Patologia	
Fattore	M	$\overline{\text{M}}$
F	n_{11}	n_{12}
$\overline{\text{F}}$	n_{21}	n_{22}
Totale	$n_{.1}$	$n_{.2}$

Non è quindi possibile, in queste condizioni, la stima del rischio relativo e i dati raccolti consentono solo di stimare le probabilità associate agli eventi F ed $\overline{\text{F}}$ "presenza o assenza del fattore considerato" subordinatamente agli eventi M "individuo affetto da patologia" e $\overline{\text{M}}$ "individuo sano". In pratica, quindi, le probabilità che è possibile stimare con questo tipo di dati considerano come dipendente la variabile che si vorrebbe considerare esplicativa e viceversa.

Nelle indagini caso-controllo l'individuazione dei fattori di rischio viene di solito effettuata mediante il calcolo del cosiddetto *odds-ratio R* che corrisponde a

$$R = \frac{P(M \mid F)/P(\overline{M} \mid F)}{P(M \mid \overline{F})/P(\overline{M} \mid \overline{F})} = \frac{\pi_{(1)}/(1-\pi_{(1)})}{\pi_{(2)}/(1-\pi_{(2)})}. \qquad (2)$$

Si vede subito che questa statistica è data dal rapporto fra due odds, dove quello posto al numeratore è relativo agli individui soggetti al fattore di rischio e valuta la probabilità di essere affetti dalla malattia rispetto alla probabilità di essere sani, mentre quello posto al denominatore ha un significato identico per i soggetti che non sono soggetti al fattore. La (2) può assumere valori non negativi e quando è pari ad 1 indica che le variabili considerate nella tabella 2×2 sono indipendenti fra di loro. Se, invece, il rapporto è maggiore di 1 questo significa che l'odds al numeratore è maggiore di quello al denominatore, cosicché la probabilità di presentare la malattia è più elevata per gli individui soggetti al fattore rispetto agli individui che non sono soggetti al fattore, dato che $\pi_{(1)}$ risulta maggiore di $\pi_{(2)}$. Considerazioni di segno opposto valgono quando l'odds-ratio risulta minore di 1.

Si dimostra facilmente (cfr. Agresti, 1996) che, data una tabella a doppia entrata 2×2, questo indice assume uno stesso valore sia quando si considera la variabile riportata sulla prima riga come dipendente da quella riportata sulla prima colonna, sia quando il ruolo di dipendenza fra le variabili viene scambiato. Per questo motivo il valore dell'odds-ratio può essere determinato anche quando i dati raccolti derivano da indagini caso-controllo.

In alcune circostanze che risultano abbastanza ragionevoli nei casi reali, questa statistica stima anche, con una certa attendibilità, il rischio relativo Ω. Dal confronto fra la (1) e (2) risulta infatti che

$$R = \frac{\pi_{(1)}}{\pi_{(2)}} \frac{1-\pi_{(2)}}{1-\pi_{(1)}} = \Omega \frac{1-\pi_{(2)}}{1-\pi_{(1)}}$$

per cui, quando le probabilità $\pi_{(1)}$ e $\pi_{(2)}$ risultano piccole, come spesso avviene nelle situazioni reali, il valore della (2) fornisce una stima approssimata della (1).

La stima campionaria di R, ottenuta sulla base di dati analoghi a quelli riportati nella tabella 2, è data da

$$\hat{R} = \frac{n_{11} n_{22}}{n_{12} n_{21}} \tag{3}$$

e si basa quindi sul prodotto delle numerosità poste sulla diagonale principale diviso per il prodotto delle numerosità riportate sulla diagonale secondaria.

Per evitare di utilizzare un indice che possa assumere un valore indeterminato, in questo lavoro è stata considerata una variante (cfr. Agresti, 1996) che si ottiene dalla (3) sommando il valore 1/2 a ogni numerosità contenuta nelle celle interne della tabella. Si è quindi utilizzata la statistica

$$\tilde{R} = \frac{(n_{11} + 0{,}5)(n_{22} + 0{,}5)}{(n_{12} + 0{,}5)(n_{21} + 0{,}5)} \tag{4}$$

che non presenta gli inconvenienti della precedente.

2. L'individuazione dei fattori di rischio

In generale, nelle indagini per individuare i fattori di rischio associati ad una particolare patologia, si utilizzano gli stessi dati campionari per verificare più ipotesi diverse, ognuna delle quali si riferisce all'incidenza di un singolo fattore.

Consideriamo per semplicità il caso in cui ciascuno dei k fattori di rischio possa assumere le due sole modalità "presente" e "assente" che indichiamo rispettivamente con il valore 1 ed il valore 0. Sia W_{1j} (j=1,2,…,k) la variabile Zero-Uno rilevata sugli individui sani e W_{2j} la corrispondente variabile Zero-Uno rilevata sugli individui affetti dalla patologia.

Indicate con $\wp_{ij}$ (i=1,2; j=1,2,…,k) le corrispondenti distribuzioni di probabilità ignote, la j-esima ipotesi parziale, utilizzata per verificare la significatività del j-esimo fattore, risulta

$$H_{0j}: \wp_{1j} = \wp_{2j}, \qquad j=1,2,\ldots,k \tag{5}$$

contro l'ipotesi alternativa

$$H_{1j}: \wp_{1j} \neq \wp_{2j}, \qquad j=1,2,\ldots,k.$$

Le stesse ipotesi possono anche essere espresse nella forma equivalente

$$H_{0j}: R_j = 1 \tag{6}$$

$$H_{1j}: R_j \neq 1, \qquad j = 1,2,...,k \tag{7}$$

La verifica delle k ipotesi nulle parziali (5) o (6) implica il calcolo simultaneo della significatività di k statistiche test, una per ciascuna delle ipotesi da verificare, che in genere risultano dipendenti fra di loro.

E' ben noto che in casi come questi si ottiene come conseguenza un aumento della probabilità di rifiutare almeno una di queste singole ipotesi quando sono vere. Pertanto con questo procedimento aumenta la probabilità di considerare significativo un fattore che in realtà non esercita un'influenza sulla probabilità di contrarre la malattia.

Se le statistiche test fossero indipendenti e la significatività di ciascuno di essi fosse pari ad α, la probabilità di rifiutare almeno una delle ipotesi nulle quando sono vere è data da

$$\alpha* = 1-(1-\alpha)^k. \tag{8}$$

Per ovviare a questo inconveniente si può fissare un livello di significatività per la verifica dell'ipotesi nulla globale H_0 di mancanza di effetti per tutti i fattori complessivamente considerati.

Questa ipotesi H_0 corrisponde all'intersezione delle singole ipotesi nulle parziali di uguaglianza delle distribuzioni all'interno dei gruppi dei sani e dei malati

$$H_0 = \bigcap_{j=1}^{k} H_{0j},$$

mentre ovviamente l'ipotesi alternativa globale H_1 corrisponde all'unione di tutte le singole ipotesi alternative parziali

$$H_1 = \bigcup_{j=1}^{k} H_{1j}.$$

L'ipotesi nulla globale viene accettata quando vengono accettate tutte le singole ipotesi nulle parziali H_{0j} e si rifiuta se si rifiuta almeno una di queste ipotesi.

Come si è detto, si vuole tenere sotto controllo la probabilità α che almeno una statistica sia significativa. Predeterminata questa probabilità, sarebbe necessario calcolare la significatività α' con cui effettuare i singoli test. Se questi test fossero indipendenti, questo valore α' risulterebbe pari a

$$\alpha' = 1-(1-\alpha)^{1/k}$$

da cui per esempio, nel caso di $\alpha = 0{,}10$ e di due soli fattori, si ottiene $\alpha' \cong 0{,}0513$.

Se invece i test sono correlati fra di loro, la determinazione del livello di significatività α' è più complessa. In questo caso, se le statistiche test avessero tutte la medesima distribuzione e, per ipotesi, la regione critica fosse sulla coda destra, il rispetto del livello di significatività α si potrebbe ottenere determinando la distribuzione della statistica massima e calcolando su di essa la significatività dei singoli test.

In casi più generali, cioè quando le statistiche test non hanno la stessa distribuzione, si potrebbe utilizzare come statistica test a livello globale il *p*-valore minimo determinando la sua distribuzione sotto l'ipotesi nulla globale. Questa distribuzione è però generalmente complicata da determinare per via analitica proprio a causa della dipendenza fra le diverse statistiche test. Il problema può essere risolto, secondo la proposta di Westfall and Young (1993), utilizzando procedure di ricampionamento.

Nel caso esaminato, se è vera l'ipotesi nulla globale di assenza di influenza di tutti i *k* fattori considerati, i dati campionari risultano scambiabili fra di loro e la verifica di ipotesi può basarsi sulla distribuzione di permutazione del *p*-valore minimo.

Infatti, indicando con n la numerosità campionaria complessiva e tenendo presente che si dispone di due campioni, quello relativo agli individui sani, di numerosità n_1, e quello relativo agli individui malati, di numerosità n_2 ($n=n_1+n_2$), sotto H_0 ognuna delle

$$Q = n!/\prod_{i=1}^{2} n_i!$$

assegnazioni delle n osservazioni ai due campioni è equiprobabile.

Si osservi, però, che la determinazione della distribuzione del *p*-valore minimo può risultare piuttosto complessa da un punto di vista computazionale. La procedura, infatti, comporta la memorizzazione delle k distribuzioni di permutazione, ognuna formata da Q valori, delle k statistiche test e, successivamente, un numero di confronti pari a Q^2 per determinare le k distribuzioni dei *p*-valori di tali statistiche da cui ottenere la distribuzione del *p*-valore minimo.

Una procedura equivalente a quella appena descritta, ma più semplice dal punto di vista dei calcoli, può essere utilizzata se le k statistiche hanno tutte, almeno approssimativamente, la medesima distribuzione sotto H_0.

Indicata con T_j (j=1,2,…,k) la statistica test utilizzata per verificare l'ipotesi H_{0j} e con t_j la sua determinazione campionaria, i *p*-valori che consentono di verificare H_0 ad un livello di significatività prestabilito possono essere calcolati immediatamente facendo riferimento alla distribuzione della statistica $T = \max_j T_j$. Indicata con $t_v^* = \max_j t_{jv}^*$ il massimo delle stesse statistiche calcolate sulla v-esima permutazione

(v=1,2,…,Q), i p-valori modificati, che consentono di verificare l'ipotesi nulla globale ad un livello di significatività α prestabilito, risultano

$$\tilde{p}_j = \frac{1}{Q}\sum_{v=1}^{Q} \mathrm{I}_{(0,\infty)}(t_v^* - t_j), \qquad v=1,2,\ldots, Q, \tag{9}$$

mentre il p-valore associato al valore t della statistica massima è ovviamente

$$\tilde{p} = \frac{1}{Q}\sum_{v=1}^{Q} \mathrm{I}_{(0,\infty)}(t_v^* - t), \; v=1,2,\ldots, Q. \tag{10}$$

Quando le distribuzioni di permutazione sono diverse si possono utilizzare le corrispondenti statistiche standardizzate (cfr. Giraldo and Pesarin, 1992 and Bertacche and Pesarin, 1997). Indicati rispettivamente con $\bar{t}_j^*$ e s_j^* (j=1,2,…,k) la media e lo scarto quadratico medio delle distribuzioni di permutazione, queste statistiche assumono la forma

$$z_j = (t - \bar{t}_j^*)/s_j^*$$

e

$$z_{jv}^* = (t_{jv}^* - \bar{t}_j^*)/s_j^*.$$

In questo modo, sotto H_0 le statistiche z_j risultano uguali almeno in media e varianza e i risultati ottenuti sono tanto migliori quanto più queste distribuzioni hanno una forma simile.

In analogia con le (9) e (10), i p-valori associati alle statistiche test z_j che consentono di verificare H_0 al livello di significatività α sono approssimati da

$$\hat{p}_j = \frac{1}{Q}\sum_{v=1}^{Q} \mathrm{I}_{(0,\infty)}(z_v^* - z_j), \tag{11}$$

mentre il p-valore associato alla statistica standardizzata massima è

$$\hat{p} = \frac{1}{Q}\sum_{v=1}^{Q} \mathrm{I}_{(0,\infty)}(z_v^* - z), \tag{12}$$

dove $z = \max_j z_j$.

Il calcolo delle (11) e (12) comporta solo un numero di confronti pari a Q, anche se è necessario determinare la media e la varianza delle k distribuzioni di permutazione.

Si osservi che quando il numero Q delle possibili permutazioni risulta troppo elevato, questi p-valori possono essere stimati sulla base di un campione di q permutazioni estratte in modo casuale dalle Q complessive.

3. Stima della probabilità di rifiutare almeno una delle ipotesi vere

E' stato effettuato uno studio di simulazione per valutare, sotto diversi livelli di dipendenza dei fattori fra di loro, l'aumento nella probabilità di rifiutare almeno una delle singole ipotesi nulle parziali, quando sono vere, se la significatività degli odds ratio viene calcolata sulla base della loro distribuzione di probabilità marginale. Si è inoltre calcolata la significatività degli odds-ratio standardizzati sulla base della distribuzione dell'odds-ratio standardizzato massimo.

Per semplicità, si sono considerati solo due fattori che assumono le modalità 0 ed 1 ed una variabile, anch'essa dicotomica, che assume valore 1 se l'individuo presenta la patologia e il valore 0 in caso contrario.

Le ipotesi di base parziali che i fattori non abbiano influenza sulla probabilità che la patologia si manifesti sono quindi date dalla (6) per j=1,2.

Come si vede dalla (7), si tratta di un test a due code, che vuole individuare quelle variabili che risultano essere fattori di rischio oppure fattori protettivi. Per poter posizionare la regione critica su una sola coda della distribuzione, si è preferito utilizzare, anzichè la (4), la statistica test

$$R^* = \max(\widetilde{R}, 1/\widetilde{R}) \qquad (13)$$

che assume valori elevati quando il fattore influenza in modo significativo la variabile dipendente, a prescindere che sia un fattore protettivo o di rischio. La regione critica, quindi, risulta posizionata lungo la coda destra della distribuzione.

Le modalità dei due fattori sono state generate considerando diversi gradi di dipendenza fra i fattori stessi. Più in particolare le coppie di osservazioni sono state generate in modo casuale da una tabella 2×2 su cui l'indice di dipendenza assoluta normalizzato χ^2/n assume i valori 0,00; 0,16; 0,36; 0,64 e 1,00.

Per ciascuna delle diverse situazioni considerate sono state effettuate 1.000 simulazioni per campioni di sani e di malati di pari numerosità ($n_1 = n_2 = n/2$) con n = 50, 100 e 200. Per ogni simulazione, dato l'elevato valore del numero Q di possibili permutazioni, le significatività effettive sono state calcolate in modo approssimato sulla base di un campione casuale di q = 5.000 permutazioni.

Dato che le distribuzioni di permutazione risultano discrete, si è effettuato un test randomizzato per garantire il rispetto del livello di significatività globale.

Scelto un valore α=0,10, nelle tabelle successive è indicata, per le diverse numerosità campionarie, la quota di simulazioni in cui si rifiutano le due ipotesi nulle parziali e l'ipotesi nulla globale, in modo da confrontare il livello di significatività effettivo con il livello α prefissato. In particolare, la seconda, terza e quarta colonna riportano rispettivamente la quota di simulazioni in cui si rifiuta H_{01}, H_{02} e H_0 se le significatività dei due odds-ratio vengono calcolate sulla distribuzione di permuta-

Tabella 3
Livelli di significatività per $\alpha = 0{,}10$. Fattori indipendenti

n	Significatività sulla distribuzione marginale			Significatività sulla distribuzione della statistica massima		
	Rifiuto H_{01}	Rifiuto H_{02}	Rifiuto H_0	Rifiuto H_{01}	Rifiuto H_{02}	Rifiuto H_0
50	0,102	0,094	0,179	0,049	0,055	0,099
100	0,112	0,109	0,201	0,049	0,058	0,098
200	0,104	0,122	0,210	0,046	0,055	0,095

zione marginale corrispondente. Le tre colonne successive riportano invece le medesime quote quando si utilizzano i due odds-ratio standardizzati sulla base della media e della varianza della corrispondente distribuzione di permutazione e la loro significatività viene calcolata sulla base della distribuzione di permutazione della statistica massima.

La tabella 3 riporta i risultati ottenuti nel caso in cui i fattori siano indipendenti.

Come si vede dai risultati riportati nella seconda e terza colonna, ciascuna delle ipotesi nulle parziali viene rifiutata una quota di volte pari a circa il 10%, mentre l'ipotesi nulla globale viene rifiutata all'incirca nel 19% dei casi, in linea con il risultato atteso dalla (8). Con questo tipo di procedura, quindi, all'ipotesi nulla globale è associato un livello di significatività effettivo molto superiore a quello prefissato.

Ovviamente, data la correttezza del test, se la significatività degli odds ratio viene calcolata sulla distribuzione della statistica massima la quota di simulazioni in cui si rifiuta H_0 è molto prossima al livello α=0,10, come risulta dai risultati contenuti nell'ultima colonna della tabella precedente. Dai dati contenuti nella quinta e sesta colonna si osserva che con questa procedura le singole ipotesi parziali vengono verificate ad un livello che risulta inferiore ad α.

Dai risultati riportati nelle tabelle successive si rileva che, al crescere del livello di correlazione fra i due fattori, la quota di simulazioni nelle quali si rifiuta almeno una delle due ipotesi parziali quando la significatività è calcolata con riferimento alle distribuzioni marginali tende al livello α prefissato.

Tabella 4
Livelli di significatività per $\alpha = 0{,}10$. Fattori dipendenti, $\chi^2/n = 0{,}16$

n	Significatività sulla distribuzione marginale			Significatività sulla distribuzione della statistica massima		
	Rifiuto H_{01}	Rifiuto H_{02}	Rifiuto H_0	Rifiuto H_{01}	Rifiuto H_{02}	Rifiuto H_0
50	0,103	0,097	0,171	0,058	0,055	0,094
100	0,112	0,107	0,181	0,066	0,062	0,109
200	0,104	0,081	0,162	0,064	0,045	0,098

Tabella 5
Livelli di significatività per $\alpha = 0{,}10$. Fattori dipendenti, $\chi^2/n = 0{,}36$

n	Significatività sulla distribuzione marginale			Significatività sulla distribuzione della statistica massima		
	Rifiuto H_{01}	Rifiuto H_{02}	Rifiuto H_0	Rifiuto H_{01}	Rifiuto H_{02}	Rifiuto H_0
50	0,100	0,106	0,166	0,067	0,062	0,109
100	0,095	0,099	0,155	0,056	0,055	0,093
200	0,096	0,107	0,155	0,050	0,069	0,096

Tabella 6
Livelli di significatività per $\alpha = 0{,}10$. Fattori dipendenti, $\chi^2/n = 0{,}64$

n	Significatività sulla distribuzione marginale			Significatività sulla distribuzione della statistica massima		
	Rifiuto H_{01}	Rifiuto H_{02}	Rifiuto H_0	Rifiuto H_{01}	Rifiuto H_{02}	Rifiuto H_0
50	0,100	0,090	0,135	0,070	0,066	0,103
100	0,107	0,109	0,150	0,065	0,070	0,092
200	0,080	0,098	0,128	0,056	0,060	0,087

Nel caso di perfetta connessione, infine, si ottengono i risultati riportati nella tabella successiva. Dai dati della quarta colonna si vede che la significatività effettiva associata al test per la verifica di ipotesi globale è molto prossima al livello α=0,10. In questo caso, infatti, fra i test per la verifica delle due ipotesi vi è una corrispondenza uno ad uno e le differenze nei livelli di significatività osservati dipendono solo dal fatto che si è usato un test randomizzato.

Tabella 7
Livelli di significatività per $\alpha = 0{,}10$. Fattori perfettamente correlati

n	Significatività sulla distribuzione marginale			Significatività sulla distribuzione della statistica massima		
	Rifiuto H_{01}	Rifiuto H_{02}	Rifiuto H_0	Rifiuto H_{01}	Rifiuto H_{02}	Rifiuto H_0
50	0,101	0,098	0,113	0,096	0,092	0,106
100	0,094	0,093	0,102	0,094	0,093	0,104
200	0,094	0,089	0,096	0,086	0,090	0,093

Dai risultati delle simulazioni risulta quindi evidente che quando la verifica congiunta di più ipotesi sui medesimi dati campionari viene effettuata sulla base delle distribuzioni marginali delle statistiche test, si ottiene un aumento significativo della probabilità di rifiutare almeno una di queste ipotesi quando sono vere e che questo aumento è via via più marcato al diminuire del livello di connessione fra i fattori. Per ovviare a questo inconveniente è possibile utilizzare la procedura descritta nel paragrafo precedente che, essendo un test corretto, permette di verificare l'ipotesi nulla globale ad un livello di significatività prestabilito.

4. Stima della potenza

Per valutare la potenza della procedura proposta, è stato condotto un altro studio di simulazione su campioni di 100 individui, di cui metà sono sani e metà affetti dalla patologia, estratti da una popolazione in cui sono presenti due fattori dicotomici. Il primo fattore è un fattore di rischio, mentre il secondo non esercita alcun effetto sulla probabilità di presentare la malattia. Più in particolare i campioni sono stati estratti da una popolazione in cui l'odds ratio (4) calcolato per il primo fattore assume i valori 1,1(0,1)1,5, mentre quello relativo al secondo fattore risulta sempre pari ad 1. Il livello di connessione fra i due fattori è stato fissato molto prossimo alla situazione di indipendenza assoluta.

In questa situazione è possibile confrontare la potenza del test basato sulle distribuzioni marginali con quella del test basato sulla distribuzione della statistica massima. In particolare è possibile calcolare per simulazione, per entrambe le procedure, la probabilità di rifiutare correttamente l'ipotesi falsa H_{01} e la probabilità di rifiutare erroneamente l'ipotesi vera H_{02}.

Per ciascuna delle diverse situazioni considerate sono state effettuate 1.000 simulazioni e le significatività effettive sono state calcolate in modo approssimato sulla base di un campione casuale di q = 5.000 permutazioni. Anche in questo caso si è effettuato un test randomizzato.

Scelto un valore α=0,10, nella tabella successiva è indicata, per i diversi valori dell'odds ratio relativo al primo fattore, la quota di simulazioni in cui si rifiutano le due ipotesi nulle parziali e l'ipotesi nulla globale. Nella seconda, terza e quarta colonna sono indicate la quota di simulazioni in cui si rifiuta H_{01}, H_{02} e H_0 se le significatività dei due odds-ratio vengono calcolate sulle distribuzioni di permutazione marginali, mentre nelle tre colonne seguenti sono riportate le medesime quote quando la significatività dei due odds-ratio standardizzati è calcolata sulla base della distribuzione di permutazione della statistica massima.

Tabella 8
Livelli di potenza per diversi valori dell'odds ratio R_1
e per un odds ratio R_2 pari ad 1. α=0,10

R_1	Significatività sulla distribuzione marginale			Significatività sulla distribuzione della statistica massima		
	Rifiuto H_{01}	Rifiuto H_{02}	Rifiuto H_0	Rifiuto H_{01}	Rifiuto H_{02}	Rifiuto H_0
1,1	0,157	0,096	0,235	0,111	0,055	0,158
1,2	0,343	0,103	0,417	0,260	0,056	0,302
1,3	0,569	0,100	0,608	0,477	0,056	0,509
1,4	0,770	0,109	0,795	0,700	0,063	0,718
1,5	0,913	0,087	0,920	0,876	0,055	0,883

Dai valori riportati nella seconda e quinta colonna si vede che la potenza della procedura proposta è inferiore a quella del test classico anche se, ovviamente, al crescere dei valori dell'odds ratio, le differenze tendono ad attenuarsi. Per una interpretazione corretta delle funzioni di potenza è necessario tuttavia tenere presente che, sotto H_0, la significatività del test basato sulla marginale è all'incirca il doppio di quella del test basato sulla statistica massima. Dai valori riportati nella terza e sesta colonna si nota inoltre che mediante la seconda procedura la probabilità di commettere l'errore che consiste nel rifiutare l'ipotesi vera H_{02} è quasi la metà di quella che si ottiene con la procedura classica standard.

Conclusioni

Dai risultati ottenuti si rileva che con il test classico, al crescere del numero di fattori considerati, tende ad aumentare la probabilità di considerare significativi anche fattori che in realtà non esercitano alcuna influenza sulla variabile dipendente. La procedura proposta in questo lavoro consente invece la verifica contemporanea dell'incidenza dei diversi fattori controllando la probabilità che almeno uno di essi risulti significativo ed, essendo un test corretto, permette quindi di verificare l'ipotesi nulla globale ad un livello α prestabilito.

Si osservi inoltre che la procedura utilizzata nel caso di fattori che assumono solo le modalità zero ed uno può essere estesa al caso generale di fattori che assumono più modalità diverse. La procedura, infatti, conserva le sue caratteristiche anche se le statistiche idonee per le verifiche dell'incidenza dei vari fattori sono diverse dagli odds-ratio.

Riferimenti bibliografici

AA.VV. (1991) *Le cause prevenibili di handicap*, a cura di L. Fabbris e F. Vian, Conselve, CEREF, Padova.

AGRESTI A. (1996) *An Introduction to Categorical Data Analysis*, Wiley and Sons.

BERTACCHE, R. and PESARIN, F. (1997) Treatments of missing data in multidimensional testing problems for categorical variables. *Metron*, **LV**: pp. 135-149.

GIRALDO, A. and PESARIN, F. (1992) Verifica d'ipotesi in presenza di dati mancanti e tecniche di ricampionamento. *Atti della XXXVI Riunione Scientifica SIS*, Vol.**2**: pp. 271-278.

GOOD, P. (1994) *Permutation Tests*, Springer-Verlag.

MAXWELL, A.E. (1937) *Analysing qualitative data*, Chapman & Hall, London.

WESTFALL, P.H. and YOUNG, S.S. (1993) *Resampling-based multiple testing*, Wiley, New York.

A Non-Parametric Analysis to Identify Risk Factors

Summary. *A common problem in statistical medical analyses is the identification of risk factors associated with a certain disease. The collected sampled data are used to assess simultaneously more hypotheses, each of which assesses the influence of one factor. It is well known that the simultaneous assessment of two or more hypotheses entails a rise in the probability of rejecting at least one of the true null partial hypotheses.*

In this paper the rise in this probability is approximately evaluated for different levels of dependance between factors by means of a permutation-based procedure. The paper also proposes a procedure which computes an adjusted p-value for each test of the partial hypotheses in such a way that the global hypothesis that none of the factors have influence is assessed at a prefixed significance level. A simulation study is performed to check the power of the proposed procedure.

Keywords. *factors, odds ratio, multiple tests, significance level, permutation test.*