Metodi non parametrici negli studi osservazionali multivariati in presenza di fattori di confondimento

R. Arboretti Giancristoforo, M. Bolzan

2003.7

Dipartimento di Scienze Statistiche
Università degli Studi
Via C. Battisti 241-243
35121 Padova

Aprile 2003

METODI NON PARAMETRICI NEGLI STUDI OSSERVAZIONALI MULTIVARIATI IN PRESENZA DI FATTORI DI CONFONDIMENTO

R. Arboretti Giancristofaro, M. Bolzan

1. INTRODUZIONE

La valutazione dell'effetto di un fattore di esposizione o trattamento su una data patologia o su un dato evento di interesse, viene usualmente realizzata mediante il confronto dei profili di risposta di due o più gruppi di unità (o pazienti) individuati in corrispondenza di diversi livelli del trattamento.

Per poter attribuire al trattamento eventuali differenze riscontrate tra i gruppi confrontati, è necessario che i gruppi siano comparabili, ovvero simili rispetto a tutte le condizioni cosiddette "ambientali" che possono influire sulle risposte, ad eccezione del fattore oggetto di studio, cioè il trattamento. Qualsiasi confronto quindi, volto a valutare differenze negli esiti attribuibili al fattore di esposizione o trattamento, deve prevedere il controllo accurato della comparabilità dei gruppi.

Negli studi osservazionali, soprattutto a causa della ridotta formalizzazione metodologica, in particolare per la mancanza del requisito della assegnazione randomizzata delle unità ai livelli del trattamento, il problema della comparabilità tra i gruppi a confronto limita notevolmente l'applicabilità delle procedure inferenziali tradizionali e la conseguente interpretazione dei risultati.

Negli studi osservazionali è necessario quindi considerare la possibile presenza di potenziali fattori di confondimento. Si definisce fattore di confondimento una variabile che risulta associata sia alle variabili risposta, sia al trattamento. Se le distribuzioni di tali fattori differiscono tra i livelli del trattamento, un'eventuale differenza tra le risposte dei gruppi può non essere la conseguenza dell'effetto del trattamento ma dell'effetto delle variabili di confondimento. Il confronto diretto tra i gruppi di trattamento può quindi risultare distorto e portare ad attribuire erroneamente la possibile differenza tra i gruppi all'effetto del trattamento.

Relativamente alla distorsione indotta dai fattori di confondimento, i gruppi di trattamento possono differire in due modi:

- rispetto a variabili note e rilevate nel corso dello studio, nel qual caso si determina un bias evidente (overt bias);
- rispetto a variabili non note o non rilevate nel corso dello studio, nel qual caso si determina un bias nascosto (hidden bias).

La definizione e l'uso di metodi appropriati che consentano di rimuovere bias dovuti a variabili di confondimento, è quindi una delle questioni critiche per derivare inferenze valide da studi osservazionali. La maggior parte dei metodi proposti in letteratura si riferiscono al controllo dei bias evidenti. A questo proposito si può distinguere tra metodi che riguardano il disegno dello studio e metodi analitici applicabili a posteriori, successivamente alla raccolta dei dati.

In questo lavoro viene ripresa brevemente la tecnica del Propensity Score utilizzata per controllare un overt bias (Rosenbaum e Rubin, 1983a), che ha visto una ridotta applicazione rispetto ad altri metodi, spesso semplicemente per motivi pratici. La diffusione estesa di pacchetti statistici contenenti procedure già implementate, determina infatti sempre più l'applicazione, in alcuni casi anche poco critica, delle procedure più note.

L'interesse particolare che viene riservato nel presente lavoro alla metodologia del Propensity Score, è anche di tipo funzionale. Tale metodo infatti, riassumendo l'insieme delle variabili di confondimento in un unico fattore, consente di ricondursi ad un problema di confronto tra due o più gruppi di trattamento in presenza di un'unica variabile di confondimento, rispetto alla quale si può prevedere ad esempio una stratificazione a posteriori. Le procedure di verifica di ipotesi successivamente proposte faranno riferimento proprio ad una situazione di questo tipo.

Per il secondo tipo di problema, quello relativo alla possibile presenza di bias nascosti, si rinvia alle tecniche di analisi di sensitività per bias nascosti (Rosenbaum e Rubin, 1983b).

Successivamente alla presentazione della tecnica del Propensity Score, vengono presentate due procedure di verifica di ipotesi particolarmente adatte agli studi osservazionali: il test sui ranghi per ipotesi alternative coerenti dovuto a Rosenbaum (1994, 1995, 1997) e un nuovo test di permutazione multivariato e multistrato basato sulla metodologia di combinazione non parametrica NPC (NonParametric Combination) sviluppata da Pesarin (2001). Quest'ultimo test consente l'agevole trattamento della verifica di ipotesi in studi osservazionali multivariati anche in presenza di dati mancanti (anche non completamente a caso).

Infine viene presentato uno studio di simulazione comparativo tra i due test non parametrici, considerando anche il caso della presenza di dati mancanti completamente a caso.

2. IL PROPENSITY SCORE

L'idea alla base di tale metodologia è di sostituire all'insieme dei fattori di confondimento, una loro funzione che prende il nome di propensity score. Tale score viene poi utilizzato come singola variabile di confondimento. Il propensity score è la probabilità condizionata per un soggetto di essere assegnato ad uno specifico livello del trattamento dato un insieme di variabili osservate.

Si considerino inizialmente due livelli di trattamento che indicano rispettivamente l'esposizione e la non esposizione al trattamento o all'agente in studio. I gruppi confrontati corrispondono quindi rispettivamente all'esposizione al trattamento (soggetti "trattati") e alla non esposizione al trattamento (soggetti "di controllo"). Si indichino con t e con n rispettivamente il numero di soggetti esposti al trattamento e il numero di soggetti in totale. Sia c il numero di variabili di confondimento osservate sulle n unità e $\boldsymbol{U}$ il vettore c-dimensionale di tali variabili. Sia inoltre k il numero di variabili risposta osservate sulle n unità e $\boldsymbol{Y}$ il vettore k-dimensionale di tali variabili. Si considerino inoltre n variabili indicatrici dell'assegnazione al trattamento Z_j, con $j=1,...,n$, che assumono valore 1 per i soggetti appartenenti al gruppo dei trattati e valore 0 per gli individui appartenenti al gruppo di controllo. Si indichi con $\boldsymbol{Z}$ il vettore n-dimensionale delle variabili casuali Z_j. Il propensity score, indicato con $e(\boldsymbol{U})$, è definito come la probabilità di essere assegnati al gruppo di trattamento date le variabili osservate $\boldsymbol{U}$, ovvero la propensione ad appartenere al gruppo dei trattati determinata come funzione dei fattori di confondimento:

$$e(\boldsymbol{U}) = \Pr(\boldsymbol{Z} = 1 | \boldsymbol{U}). \tag{1}$$

Negli studi randomizzati, il propensity score è uno score di bilanciamento, ovvero una funzione delle variabili di confondimento $\boldsymbol{U}$, tale che la distribuzione condizionata di $\boldsymbol{U}$ dato lo score, è la stessa per le unità trattate e per le unità di controllo. In altre parole, utilizzando la notazione di indipendenza condizionata di Dawid (1979), si ha:

$$\boldsymbol{U} \perp \boldsymbol{Z} \mid e(\boldsymbol{U}). \tag{2}$$

Di conseguenza, se vengono costruiti strati o insiemi appaiati omogenei rispetto a $e(\boldsymbol{U})$, allora i due gruppi di trattamento avranno la stessa distribuzione delle variabili di confondimento. L'estensione di tali risultati agli studi osservazionali analitici non è immediata e presuppone l'introduzione di ipotesi circa la presenza o meno di possibili fattori di confondimento non noti o non rilevati. Rosenbaum e Rubin (1983a) hanno definito al riguardo, l'ipotesi di assegnazione al trattamento ignorabile. L'assegnazione al trattamento è ignorabile se $\boldsymbol{Y} \perp \boldsymbol{Z} \mid \boldsymbol{U}$. Se lo studio

non è soggetto a bias nascosti, il trattamento è ignorabile. Se il trattamento è ignorabile dato U, è ignorabile dato qualsiasi score di bilanciamento che dipende da U, e quindi è ignorabile rispetto al propensity score $e(U)$. Inoltre metodi basati sul propensity score (stratificazione, appaiamento, ecc.), producono un adeguato controllo dei fattori di confondimento inclusi in U, consentendo conseguentemente confronti non distorti dei gruppi di trattamento.

Negli studi osservazionali la funzione propensity score è quasi sempre non nota in quanto il meccanismo di assegnazione al trattamento è non noto. Comunque $e(U)$ può essere stimato dai dati osservati, utilizzando ad esempio un modello in cui Z è la variabile dipendente e i fattori di confondimento rappresentano le variabili indipendenti. Il propensity score può essere stimato ad esempio tramite un modello logistico o applicando un'analisi discriminante o utilizzando la metodologia degli alberi di classificazione. Si noti che per la stima del propensity score, le variabili risposta non hanno alcun ruolo, ma viene considerato solo l'insieme delle variabili di confondimento. Con più di due livelli di trattamento, in generale si definiscono diversi propensity score, uno per ciascun confronto tra coppie di trattamento (Rubin, 1997a, 1997b).

Si noti che il bilanciamento dei gruppi di trattamento rispetto a potenziali fattori di confondimento, ottenuto utilizzando i metodi basati sul propensity score, differisce dal bilanciamento che deriva dall'applicazione del meccanismo della randomizzazione negli studi sperimentali. La randomizzazione tende infatti a bilanciare i gruppi di trattamento rispetto a tutte le variabili di confondimento, osservate o non osservate. Negli studi osservazionali, i metodi basati sul propensity score hanno la proprietà di bilanciare i gruppi solo rispetto alle variabili di confondimento osservate. L'applicazione di tali metodi deve essere di conseguenza accompagnata da un'attenta valutazione della possibile presenza di ulteriori fattori di confondimento non adeguatamente considerati, nonché, dove possibile, dall'applicazione di metodi che esaminano la sensibilità dello studio osservazionale rispetto a bias nascosti.

La stratificazione per propensity score combina gli aspetti positivi della stratificazione e del modello parametrico, utilizzando inizialmente un modello per sintetizzare la quantità di confondimento in un singolo score, il propensity score appunto, e formando successivamente degli strati definiti in corrispondenza di intervalli di valori del propensity score e omogenei rispetto a $e(U)$. Come risultato i dati vengono quindi predisposti per un confronto tra gruppi di trattamento stratificati per un singolo fattore di confondimento.

Se il modello che stima la propensione al trattamento, cattura interamente la relazione tra l'insieme dei fattori di confondimento e il fattore di trattamento, allora all'interno di ciascuno strato, definito da valori simili del propensity score e quindi caratterizzato da una probabilità di assegnazione al trattamento

praticamente costante, soggetti trattati e di controllo tendono ad avere la stessa distribuzione dei fattori di confondimento inclusi in U.

Con riferimento alla scelta del numero di strati, alcuni risultati (Cochran, 1968; Rosenbaum e Rubin, 1984) mostrano come per relazioni monotone tra U e Y, in corrispondenza di un numero di strati pari a s = 2, 3, 4, 5, 6, le percentuali di riduzione della distorsione imputabile a U sono rispettivamente dell'ordine del 64%, 79%, 86%, 90% e 92%.

Riguardo ai vantaggi, la stratificazione per propensity score consente il controllo di un numero elevato di fattori di confondimento.

Tale metodo rende inoltre trasparente il controllo dei fattori di confondimento, permettendo al ricercatore di verificare il risultato della 'correzione'. Il modello per la stima del propensity score può essere ad esempio riformulato se si riscontrano ulteriori sbilanciamenti per alcune variabili di confondimento o possibili intervalli di valori di tali variabili per i quali non si ha un'adeguata sovrapposizione delle distribuzioni dei gruppi di trattamento. Ad esempio, se le varianze di un'importante variabile di confondimento differiscono in modo rilevante tra i due gruppi, allora si può considerare un secondo modello per il propensity score che includa il quadrato di questa variabile; oppure se le correlazioni tra due importanti variabili di confondimento differiscono tra i due gruppi allora si può aggiungere nel modello che stima il propensity score, il prodotto delle due variabili.

Un ulteriore aspetto rilevante della stratificazione per propensity score e in generale dei metodi basati sul propensity score, riguarda il trattamento dei dati mancanti. L'obiettivo in questo caso, è quello di ottenere il bilanciamento dei gruppi di trattamento sia rispetto ai valori osservati delle variabili di confondimento, sia rispetto al pattern di dati mancanti.

3. TEST NON PARAMETRICI PER IPOTESI ALTERNATIVE COERENTI

In questo paragrafo vengono presentati due metodi non parametrici per la verifica di ipotesi proposti specificamente per studi osservazionali multivariati.

Il contesto di riferimento è quello delineato nei paragrafi precedenti, ovvero del confronto dei profili di risposta multivariati di due o più gruppi di unità di osservazione stratificati rispetto ad un fattore di confondimento o rispetto a classi di valori del propensity score nel caso di più variabili di confondimento. Si suppone quindi che tutti gli accorgimenti e i criteri per migliorare la "qualità statistica" dello studio e quindi procedere al confronto tra i gruppi, con particolare riferimento alle tecniche di controllo per fattori di confondimento, siano stati adeguatamente considerati e applicati.

Un contributo rilevante circa la possibilità di interpretare in senso causale relazioni di tipo trattamento-risposta nell'ambito di studi osservazionali è dovuto a

Cochran (1965). Cochran discute l'importanza che l'ipotesi di ricerca negli studi osservazionali sia il più possibile "elaborata" ovvero estremamente particolareggiata, sia al momento della sua definizione che al momento della verifica statistica con i dati osservati e della conseguente interpretazione dei risultati. Per ipotesi di ricerca "elaborata" si intende la definizione dettagliata o "ristretta" di un pattern "multivariato" di risultati attesi coerente con l'ipotesi di effetto causale del trattamento. Il passo successivo è verificare nei dati la consistenza con il pattern anticipato. Tale pattern multivariato si traduce quindi nella definizione di una ipotesi alternativa multivariata di tipo direzionale, ove per ciascuna variabile viene decisa l'alternativa parziale più adeguata. Se tale pattern viene confermato nei dati, l'evidenza a favore della spiegazione causale dell'associazione trattamento-risposta è forte perché ogni possibile spiegazione alternativa a questa ipotesi deve poter spiegare l'intero pattern multivariato.

L'attenzione è quindi rivolta all'individuazione di test multivariati per saggiare l'ipotesi nulla di non causalità della relazione trattamento-risposta contro una ipotesi alternativa che traduca esattamente il pattern coerente specificato dal ricercatore e che per questo verrà definita ipotesi alternativa coerente. Il pattern coerente può essere complesso in quanto il trattamento può avere effetti su diversi tipi di variabili: categoriche, continue, miste; inoltre per ciascuna variabile risposta viene specificata la direzionalità dell'effetto, tenendo conto inoltre della possibile presenza di fattori di confondimento (eventualmente controllati mediante il propensity score) e di dati mancanti.

Con riferimento specifico alla verifica di ipotesi multivariata di due gruppi stratificati (ad esempio mediante propensity score), il primo importante problema riguarda la modellizzazione della struttura di dipendenza tra le variabili in esame. È noto, ad esempio, come tale modellizzazione risulti estremamente difficile per variabili di tipo categoriale (Joe, 1997). Inoltre nei casi di alternative ristrette o inferenza isotonica le soluzioni parametriche proposte richiedono massimizzazioni vincolate delle verosimiglianze, la cui distribuzione asintotica è spesso di difficile individuazione e la velocità di convergenza di tali soluzioni alla distribuzione asintotica è spesso molto lenta (si veda ad esempio Cohen e Sackrowitz, 1998; Hirotsu, 1998; Robertson *et al.*, 1988).

Nel seguito verranno considerate due soluzioni di tipo non parametrico. La prima suggerita da Rosenbaum (1994, 1995, 1997), prevede l'applicazione di test non parametrici basati sui ranghi per ipotesi alternative coerenti, basati su funzioni che confrontano insiemi di dati parzialmente ordinati (statistiche POSET-Partially Ordered SETs). La seconda soluzione rappresenta un'estensione dei test multivariati di permutazione (Pesarin, 2001) a test multivariati e multistrato.

3.1 Test sui ranghi per ipotesi alternative coerenti

Siano $Y_1, Y_2, \ldots, Y_k$, k variabili risposta osservate su un insieme di n unità statistiche suddivise in due gruppi definiti rispetto ai livelli di un qualche criterio di classificazione di cui si vuole valutare l'effetto sulle k variabili. Le unità di osservazione possono essere inoltre raggruppate in s, $s \geq 1$, strati omogenei rispetto ad una qualche variabile di stratificazione (es. fattore di confondimento o propensity score).

Per semplicità di trattazione, si supponga che il fattore di classificazione sia rappresentato da un trattamento e che i gruppi confrontati indichino rispettivamente l'esposizione al trattamento (soggetti 'trattati') e la non esposizione al trattamento (soggetti 'di controllo').

Si indichi con t_i e con n_i rispettivamente il numero di soggetti trattati e il numero di soggetti in totale dello strato i-esimo, e con $\mathbf{Y}_{ij}$ il vettore k-dimensionale delle risposte del soggetto j-esimo appartenente allo strato i-esimo, $j = 1, \ldots, n_i$, $i = 1, \ldots, s$. Si considerino inoltre n variabili indicatrici Z_{ij} tali che $Z_{ij} = 1$ se il soggetto j-esimo dello strato i-esimo appartiene al gruppo dei trattati e $Z_{ij} = 0$ se il soggetto è un individuo appartenente al gruppo di controllo, tali che:

$$\sum_{i=1}^{s}\sum_{j=1}^{n_i} Z_{ij} = n. \qquad (3)$$

Si supponga che il ricercatore abbia specificato il pattern di risultati coerente con l'ipotesi di un effetto del trattamento, nel senso di aver definito la direzionalità dell'effetto del trattamento su ciascuna variabile risposta per i soggetti sottoposti al trattamento rispetto ai soggetti di controllo. Obiettivo dell'analisi è quindi verificare se vi siano differenze statisticamente significative tra i profili di risposta multivariati dei due gruppi nella direzione prevista dall'ipotesi di effetto del trattamento. La costruzione della statistica test si articola nelle seguenti fasi:

i) si valuta la "distanza" tra il pattern globale delle risposte osservate e quello atteso, confrontando ciascuna unità con tutte le restanti unità appartenenti ad uno stesso strato, ordinando a coppie i vettori risposta a seconda di quale delle due unità confrontate si avvicina di più al pattern atteso per un soggetto trattato;
ii) si assegna a ciascuna unità un rango che è il risultato dei confronti di cui al punto precedente.

Per il confronto dei vettori risposta, essendo in un ambito multivariato è necessario definire una relazione di ordinamento parziale.

A questo fine si introducono i concetti di ordinamento semplice, parziale o quasi ordinamento su un insieme finito A.

Si definisce ordinamento semplice su A una relazione binaria indicata con il simbolo $\lesssim$ definita su A per la quale valgono le seguenti proprietà:

i) riflessiva: $\forall$ a $\in$ A, a $\lesssim$ a;

ii) transitiva: $\forall$ a, b, c $\in$ A, se a $\lesssim$ b e b $\lesssim$ c allora a $\lesssim$ c;

iii) antisimmetrica: $\forall$ a, b $\in$ A, se a $\lesssim$ b e b $\lesssim$ a allora a = b;

iv) ogni coppia di elementi di A è confrontabile: $\forall$ a, b $\in$ A, si verifica che a $\lesssim$ b oppure b $\lesssim$ a oppure a = b.

Una relazione binaria $\lesssim$ su A è un ordinamento parziale se gode delle proprietà riflessiva, transitiva e antisimmetrica, ma non viene soddisfatta la (iv), nel senso che possono esserci coppie di elementi di A tra i quali non è possibile istituire alcun confronto. Una relazione binaria $\lesssim$ su A è un quasi-ordinamento se gode delle proprietà riflessiva e transitiva, non necessariamente quella antisimmetrica e può ammettere elementi non confrontabili. L'insieme A si definisce insieme totalmente ordinato, parzialmente ordinato o quasi-ordinato se in esso è definito rispettivamente un ordinamento semplice, parziale o un quasi-ordinamento.

L'interesse specifico nel processo di costruzione del test per ipotesi alternative coerenti, è rivolto agli insiemi parzialmente ordinati, detti anche POSET, per i quali non è possibile confrontare tutte le coppie di elementi dell'insieme stesso, nel senso che per qualche a, b $\in$A, non risulta verificata nessuna delle tre condizioni: a = b, a $\lesssim$ b, b $\lesssim$ a. La relazione d'ordinamento parziale viene definita nell'insieme delle risposte R $\subset$ R^k (dove R^k è l'insieme dei reali k-dimensionale), tale che $\boldsymbol{Y}_{ij} \lesssim \boldsymbol{Y}_{ih}$, con $\boldsymbol{Y}_{ij}$, $\boldsymbol{Y}_{ih} \in$R, significa che il vettore di risposte $\boldsymbol{Y}_{ij}$ è più distante dal pattern atteso per un soggetto trattato rispetto al vettore risposta $\boldsymbol{Y}_{ih}$, per $j,h=1,...,n_i$, $i=1,...,s$. La definizione dell'ordinamento parziale $\lesssim$, varierà naturalmente a seconda della specifica situazione e del tipo di ipotesi causale formulata.

Proseguendo nella costruzione della statistica test, si definiscono le seguenti variabili indicatrici L_{ijh} tali che:

$$L_{ijh} = \begin{cases} 1 & se\ Y_{ih} \leq Y_{ij}\ con\ Y_{ij} \neq Y_{ih} \\ -1 & se\ Y_{ij} \leq Y_{ih}\ con\ Y_{ij} \neq Y_{ih} \quad , j, h=1,...,n_i,\ i=1,...,s \\ 0 & altrimenti \end{cases} \tag{4}$$

Il rango per il soggetto *j*-esimo appartenente allo strato *i*-esimo è dato da:

$$R_{ij} = \sum_{h=1}^{n_i} L_{ijh} \tag{5}$$

In altri termini, il rango del soggetto *j*-esimo è esprimibile come differenza tra il numero di soggetti appartenenti allo stesso strato che al confronto presentano una risposta più distante dai risultati attesi per un soggetto trattato e il numero di soggetti con risposta più vicina. Di conseguenza un alto valore positivo di R_{ij} è indicativo del fatto che ci sono molti soggetti che hanno al confronto una risposta più distante dall'atteso, mentre un basso valore negativo di R_{ij} suggerisce la presenza di soggetti con risposta più vicina al pattern ipotizzato per un soggetto trattato rispetto al soggetto *j*-esimo, appartenente al lo strato *i*-esimo. Meno chiaramente interpretabile è un valore di R_{ij} vicino a 0. Tale situazione si verifica o quando il numero di soggetti con risposta più distante dall'atteso rispetto a Y_{ij} è all'incirca pari al numero di soggetti con risposta meno distante dall'atteso, o quando il soggetto *j*-esimo risulta non confrontabile con la maggior parte dei soggetti restanti. Si noti che un soggetto con valore di R_{ij} pari a 0 non contribuisce al valore della statistica test definita sugli Y_{ijr} ma, come risulterà chiaro nel seguito, entra nella determinazione della stima della varianza della stessa.

La statistica test è definita come somma dei valori di L_{ijh} per cui il soggetto *j*-esimo riceve il trattamento e il soggetto *h*-esimo è di controllo, ovvero il numero di volte in cui un soggetto trattato presenta una risposta più vicina all'atteso rispetto ad un controllo, meno il numero di volte in cui un controllo mostra una risposta più vicina all'atteso rispetto ad un soggetto trattato:

$$T_R = \sum_{i=1}^{s}\sum_{j=1}^{n_i}\sum_{h=1}^{n_i} Z_{ij}(1 - Z_{ih})\, L_{ijh} \tag{6}$$

Utilizzando la notazione matriciale e sfruttando alcune proprietà della relazione $\lesssim$, è possibile esprimere alternativamente la statistica T come somma su tutti gli strati dei ranghi R_{ij} per i t_i soggetti trattati:

$$T_R = \boldsymbol{Z}^{\top}\boldsymbol{L}(\boldsymbol{1} - \boldsymbol{Z}) = \boldsymbol{Z}^{\top}\boldsymbol{L}\boldsymbol{1} - \boldsymbol{Z}^{\top}\boldsymbol{L}\boldsymbol{Z} = \boldsymbol{Z}^{\top}\boldsymbol{L}\boldsymbol{1} = \boldsymbol{Z}^{\top}\boldsymbol{R} \quad \text{con } \boldsymbol{R} = \boldsymbol{L}\boldsymbol{1}, \tag{7}$$

dove:

$$\begin{aligned}
&\underset{n\times 1}{\boldsymbol{Z}} = \left[Z_{ij},\ j=1,\ldots n_i; i=1,\ldots,s\right],\\
&\underset{n\times n}{\boldsymbol{L}} = \left[L_{ijh},\ j,h=1,\ldots n_i; i=1,\ldots,s\right],\\
&\underset{n\times 1}{\boldsymbol{R}} = \left[R_{ij},\ j=1,\ldots n_i, i=1,\ldots,s\right],\\
&\underset{n\times 1}{\boldsymbol{1}} = [1,1,\ldots,1]^{\top},
\end{aligned} \tag{8}$$

ovvero:

$$T_R = \sum_{i=1}^{s}\sum_{j=1}^{n_i} Z_{ij}R_{ij}\,, \tag{9}$$

in quanto $\boldsymbol{1}^{\top}\boldsymbol{L}\boldsymbol{1} = 0$ e $\boldsymbol{Z}^{\top}\boldsymbol{L}\boldsymbol{Z} = 0$ per il fatto che L_{ijh} e L_{ihj} sommati si elidono, e per la asimmetria di $\lesssim$ per cui $L_{iij} = 0$.

La statistica T_R è alla base del test per saggiare l'ipotesi nulla di assenza di effetto causale del trattamento contro l'ipotesi alternativa coerente secondo cui il trattamento induce la relazione $\lesssim$ tra i soggetti di controllo e i soggetti trattati.

Avendo espresso la statistica nella forma di una statistica somma, è possibile sfruttare le proprietà corrispondenti, in particolare con riferimento al calcolo del valore atteso e della varianza di T_R sotto l'ipotesi nulla.

Nell'ipotesi di assegnazione al trattamento ignorabile, sotto l'ipotesi nulla le risposte Y_{ij} non variano al variare dell'assegnazione all'uno o all'altro gruppo di trattamento ma sono strutture fisse della popolazione finita di n soggetti come avviene nella inferenza per studi randomizzati. Di conseguenza sotto l'ipotesi nulla la somma dei ranghi $z^{\top}R$ è semplicemente la somma di n ranghi fissi selezionati con campionamento casuale semplice con ripetizione da una popolazione di n ranghi. Utilizzando gli argomenti standard della teoria delle statistiche rango lineari, ne deriva che:

$$E\left(Z^{T}R\right) \quad = \sum_{i=1}^{s} t_i\overline{R}_i = 0\,, \tag{10}$$

dove:

$$\overline{R}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} R_{ij} = 0\,,\ i=1,\ldots,s, \tag{11}$$

e inoltre:

$$\begin{aligned}\mathrm{var}\left(Z^T R\right) &= \sum_{i=1}^{s} \frac{t_i(n_i - t_i)}{n_i(n_i-1)} \sum_{j=1}^{n_i} \left(R_{ij} - \overline{R}_i\right)^2 \\ &= \sum_{i=1}^{s} \frac{t_i(n_i - t_i)}{n_i(n_i-1)} \sum_{j=1}^{n_i} R_{ij}^2 .\end{aligned} \tag{12}$$

Applicando il teorema limite centrale nella versione di Lindeberg si ottiene che la versione standardizzata di T_R è asintoticamente normale. I quantili della distribuzione normale standardizzata possono essere usati per derivare i valori critici per il test T_R e calcolare i corrispondenti valori del *p*-value.

Il test T_R per ipotesi alternative coerenti così definito è un test corretto (Rosenbaum, 1995).

La statistica T_R è una generalizzazione di alcuni test noti. Se il numero delle variabili risposta e il numero di strati è pari ad uno e l'ordinamento parziale $\lesssim$ rappresenta la relazione di disuguaglianza ordinaria $\leq$, si ottiene il test somma dei ranghi di Wilcoxon-Mann-Whitney. Per osservazioni censurate e con un'opportuna definizione dell'ordinamento parziale $\lesssim$, si ha la statistica test di Gehan.

3.2 Test di permutazione multivariato e multistrato

Si introduce ora una estensione in ambito osservazionale della metodologia della combinazione non parametrica di test di permutazione dipendenti NPC (Pesarin, 2001), per la definizione di un test di permutazione multivariato e multistrato per ipotesi alternative coerenti per il confronto tra due gruppi di trattamento stratificati rispetto ad una variabile di stratificazione (es. fattore di confondimento). Nel contesto osservazionale, per valutare l'ammissibilità della condizione di scambiabilità, necessaria per l'applicazione dei test di permutazione, si richiede che i gruppi confrontati siano resi comparabili mediante l'applicazione di metodi statistici per il controllo dei fattori di confondimento. L'applicazione dei metodi di permutazione nell'analisi di studi osservazionali risulta rilevante in quanto l'assunzione di scambiabilità rappresenta una condizione molto debole.

Un ulteriore aspetto che rende l'approccio di permutazione una soluzione estremamente valida nel contesto osservazionale, è dato dalla possibilità di trovare

un'effettiva soluzione al problema non semplice riguardante la struttura di dipendenza tra le variabili risposta di interesse, mediante l'applicazione della metodologia della combinazione non parametrica di test di permutazione dipendenti NPC (Pesarin, 2001). Ricordiamo inoltre che in ambito osservazionale, come già accennato, ci si può trovare ad affrontare problemi complessi di verifica di ipotesi multidimensionale con ipotesi alternative ristrette. La NPC rappresenta una soluzione efficiente per questo tipo di problemi: le soluzioni possono essere di tipo esatto anche per le più realistiche numerosità finite e sono in generale asintoticamente coincidenti con le soluzioni ottime parametriche, qualora queste esistano (Pesarin, 2001).

La NPC prevede la scomposizione del problema di verifica di ipotesi k-dimensionale in due fasi: nella prima fase, si definisce in modo appropriato un insieme di k, con $k \geq 1$, test di permutazione unidimensionali detti *test parziali*. Ciascun test parziale è volto ad esaminare il contributo di ogni singola variabile risposta. La seconda fase consiste nella combinazione non parametrica dei test parziali in un unico *test combinato del secondo ordine*, che saggia se vi siano globalmente delle differenze tra le distribuzioni multivariate dei gruppi. In presenza di una variabile di stratificazione, vengono previsti due livelli di combinazione: la combinazione dei test parziali in s test combinati del secondo ordine, $s \geq 1$, ciascuno corrispondente allo strato i-esimo, $i=1,\ldots, s$, e una ulteriore combinazione di quest'ultimi in un unico *test combinato del terzo ordine*.

Indicando con A e B i due gruppi di trattamento, con $\boldsymbol{Y}$ la variabile risposta p-variata, $p \leq k$, con componenti continue, binarie o categoriali $\{Y_1,\ldots, Y_p\}$, il sistema di ipotesi prevede che l'ipotesi nulla globale di uguaglianza in distribuzione delle distribuzioni multivariate dei due gruppi A e B, sia adeguatamente scomponibile, come segue:

$$H_0 : \left\{ \boldsymbol{Y}_A \overset{d}{=} \boldsymbol{Y}_B \right\}, \text{con} H_0 : \bigcap_{i=1}^{s} \left[\bigcap_{h=1}^{p} \left(Y_{hAi} \overset{d}{=} Y_{hBi} \right) \right]. \tag{13}$$

Le sub-ipotesi alternative marginali possono essere di tipo direzionale o bilaterale a seconda del tipo di relazione attesa per ogni variabile all'interno del pattern specificato:

$$H_1 : \bigcup_{i=1}^{s} \left[\bigcup_{h=1}^{p} \left(Y_{hAi} <\overset{d}{\neq}> Y_{hBi} \right) \right]. \tag{14}$$

Si noti inoltre che il test di permutazione multivariato e multistrato permette di verificare anche il seguente sistema di ipotesi:

$$H_0 : \bigcap_{b=1}^{p}\left[\bigcap_{i=1}^{s}\left(Y_{hAi} \overset{d}{=} Y_{hBi}\right)\right], \tag{15}$$

contro l'alternativa:

$$H_1 : \bigcup_{b=1}^{p}\left[\bigcup_{i=1}^{s}\left(Y_{hAi} \overset{d}{<\neq>} Y_{hBi}\right)\right], \tag{16}$$

effettuando quindi dapprima p test combinati multistrato e successivamente un test di terzo ordine globale multivariato.

L'applicazione della metodologia NPC permette quindi una notevole flessibilità di analisi a seconda del pattern coerente specificato. Per ogni dettaglio concernente l'implementazione della metodologia NPC si rinvia a Pesarin (2001).

4. STUDIO COMPARATIVO DI SIMULAZIONE

Si consideri un problema di verifica di ipotesi per il confronto tra due gruppi di trattamento di numerosità n_1, n_2, suddivisi in s strati rispetto ad una variabile di stratificazione. Si supponga di rilevare k risposte sulle $n=n_1+n_2$ unità, e che queste rappresentino k misure ripetute di una stessa variabile risposta.

Per esaminare il comportamento del test basato sui ranghi T_R e del test di permutazione multivariato e multistrato T_P per ipotesi alternative coerenti sotto l'ipotesi nulla e in termini di potenza, sono state condotte alcune simulazioni comparativa considerando il seguente modello di risposta di tipo addittivo:

$$Y_{iuj}(t) = \mu_i + \eta_{iu}(t) + Z_{iuj}(t),\ j = 1,\ldots,n_i,\ u = 1, 2,\ i = 1,\ldots,s,\ t = 1,\ldots,k, \tag{17}$$

dove μ_i è la costante di popolazione per lo strato i -esimo; $\eta_{iu}(t)$ è l'effetto del trattamento dipendente dal tempo e $Z_{iuj}(t)$ le componenti di errore. I dati sono stati generati specificando i seguenti parametri:

$$\mu_i = 100;\ \eta_{i1}(t) = 1, \eta_{i2}(t) = 1.2\ ,\ i=1,\ldots,s,\ t=1,\ldots,k. \tag{18}$$

Per le componenti di errore si considera un modello AR(t -1) del tipo:

$$Z_{iuj}(t) = U_{iujt} + 1/2\sum_{r=1}^{t-1} U_{iujr};\ t = 1,\ldots,k;\ U_{iujt} \sim \mathrm{U}(0,1). \tag{19}$$

Il sistema di ipotesi che traduce il pattern causale ipotizzato, considera le seguenti ipotesi parziali:

$$H_{0it}:\left\{Y_{i1t} \overset{d}{=} Y_{i2t}\right\} \text{ vs } H_{1it}:\left\{Y_{i1t} \overset{d}{>} Y_{i2t}\right\},\ i=1,...,s,\ t=1,...,k. \tag{20}$$

Nelle prove di simulazione effettuate, il numero delle variabili e le ampiezze campionarie per ciascuno strato, sono specificate nelle didascalie delle tabelle contenenti i risultati. Il numero degli strati è pari a 2. Il numero di simulazioni in ogni prova è pari a 1000. Per il calcolo del test T_P, il numero di iterazioni Monte Carlo Condizionate (CMC, Pesarin, 2001) è pari a 1000 e la funzione di combinazione considerata è quella di Fisher. I risultati dello studio sono riportati nella tavola 1.

TAVOLA 1

Comportamento sotto H_0 e potenza dei test per ipotesi alternative coerenti (4 variabili, $n_{1i} = n_{2i} = 10$)

α	H_0		Potenza	
	T_P	T_R	T_P	T_R
0.010	0.010	0.011	0.413	0.401
0.025	0.027	0.033	0.591	0.558
0.050	0.053	0.054	0.708	0.697
0.100	0.102	0.109	0.829	0.821
0.200	0.199	0.215	0.920	0.912
0.300	0.307	0.314	0.956	0.954
0.400	0.408	0.405	0.978	0.976
0.500	0.501	0.512	0.990	0.989
0.600	0.609	0.607	0.998	0.997
0.700	0.696	0.702	0.999	0.999
0.800	0.799	0.801	0.999	1
0.900	0.897	0.903	1	1
1.000	1	1	1	1

I risultati mostrano come sia il test di permutazione multivariato e multistrato, sia il test basato sui ranghi rispettano il livello di significatività nominale. Riguardo la potenza, i due test mostrano un comportamento simile, con un lieve vantaggio del test T_P rispetto al test T_R.

Ulteriori prove di simulazione per valutare la potenza dei test T_P e T_R, sono state effettuate in presenza di dati mancanti nelle stesse condizioni presentate precedentemente. Riguardo al test basato sui ranghi, non vi è attualmente una

trattazione in letteratura per il trattamento dei dati mancanti. Vengono qui proposte due possibili soluzioni: i) si elimina l'intero vettore di valori dell'unità con almeno un dato mancante; ii) l'unità con almeno un dato mancante viene considerata nel calcolo della statistica test e tale unità ha quindi rango nullo in quanto nel confronto con le altre unità dello stesso strato risulta sempre non ordinabile.

Il test di permutazione multivariato e multistrato consente un trattamento efficiente dei dati mancanti, rendendo possibile l'inferenza su tutti i dati osservati, compresi i valori osservati delle unità che presentano dati mancanti. Al fine di confrontare la potenza del test T_P rispetto al test T_R, considerando per quest'ultimo entrambe le soluzioni (i) e (ii), sono state quindi effettuate prove di simulazione nell'ipotesi che i dati mancanti siano mancanti completamente a caso (MCAR). Si noti che il test T_P consente di trattare anche problemi di verifica di ipotesi in cui i dati mancanti sono supposti mancanti non a caso (not-MAR, Pesarin, 2001).

Si consideri nuovamente il modello di risposta introdotto precedentemente con le specificazioni sopra riportate. I risultati delle prove sono riportati nella tavola 2.

TAVOLA 2

Potenza dei test per ipotesi alternative coerenti in presenza di dati mancanti completamente a caso (MCAR)

	4 variabili, $n_{1i} = n_{2i} = 10$		
α	12 dati mancanti per strato		
	T_P	T_R (i)	T_R (ii)
0.010	0.078	0.059	0.076
0.025	0.171	0.154	0.168
0.050	0.279	0.251	0.253
0.100	0.441	0.385	0.389
0.200	0.611	0.583	0.587
0.300	0.719	0.711	0.713
0.400	0.800	0.786	0.789
0.500	0.876	0.867	0.869
0.600	0.923	0.899	0.900
0.700	0.953	0.947	0.948
0.800	0.979	0.972	0.974
0.900	0.995	0.989	0.990
1.000	1	1	1

I risultati indicano una perdita di potenza per il test T_R in entrambe le situazioni (i) e (ii) rispetto al test T_P. In particolare la perdita di potenza risulta ovviamente più marcata nel caso della soluzione (i).

Con riferimento alla valutazione della potenza non condizionata dei test di permutazione via NPC al confronto con soluzioni parametriche ottimali, sono presentati di seguito alcuni risultati di un confronto del test di permutazione multivariato e multistrato con il test di Hotelling, uniformemente più potente tra gli invarianti per ipotesi alternative di tipo bilaterale, per il confronto di due gruppi di unità non stratificate. Si consideri il seguente modello addittivo di risposta:

$$Y_{uj}(t)=\mu+\eta_u(t)+\delta_{uj}(t)+\sigma(t)\cdot Z_{uj}(t), j=1,...,n_u,\ u=1,2,\ t=1,...,k, \qquad (21)$$

dove μ è la costante di popolazione, $\eta_u(t)$ l'effetto del trattamento dipendente dal tempo, $\sigma(t)$ un coefficiente dipendente dal tempo e indipendente dalle unità e dal trattamento, $\delta_{uj}(t)$ gli effetti stocastici individuali e $Z_{uj}(t)$ le componenti di errore. Sia $\delta_{uj}(t)$ che $Z_{uj}(t)$ sono variabili casuali i.i.d. centrate, con distribuzione non nota e indipendenti rispetto alle unità e al trattamento ma non rispetto al tempo. Una delle possibili specificazioni per gli effetti individuali considera un modello AR(1) del tipo:

$$\delta_{uj}(t)=\upsilon(t)\cdot\delta_{uj}(t-1)+\beta(t)\cdot W_{uj}(t) \qquad (22)$$

dove $\upsilon(t)$ è un coefficiente autoregressivo, $W_{uj}(t)$ rappresentano le variazioni casuali i.i.d. del comportamento individuale e $\beta(t), t=1,...,k$ è un coefficiente variabile rispetto al tempo. Per semplicità è stato considerato un solo strato e i dati sono stati generati specificando per il modello i seguenti parametri:

$$\mu_1=\mu_2=100;\delta(t)=\delta^{t-1};Z_{uj}(t)\sim N(0,1),\ t=1,...,k,\ k=3. \qquad (23)$$

Per gli effetti individuali stocastici autoregressivi, si considera un modello AR(1) del tipo:

$$\delta_{uj}(t)=\upsilon^{t-1}\cdot\delta_{uj}(t-1)+\beta\cdot t\cdot W_{uj}(t); W_{uj}(t)\sim N(0,1), \qquad (24)$$

con

$$\eta_1(1)=1.25,\eta_1(2)=1,\eta_1(3)=1.5,\eta_2(1)=1,\eta_2(2)=1.75,\eta_2(3)=1. \qquad (25)$$

Il numero di simulazioni in ogni prova è pari a 1000. Per il calcolo del test di permutazione, il numero di ricampionamenti condizionati è pari a 1000.

Si noti che da ulteriori prove effettuate, il comportamento in potenza dei due test per k>3 è simile a quello riportato in tavola 3.

TAVOLA 3

Potenza del test di permutazione (T_P) e del test di Hotelling (T^2)

	$n_1 = n_2 = 10$, k=3						$n_1 = n_2 = 6$, k=3					
ν	1.1	4	1.1	1.1	2	ν	1.1	4	1.1	1.1	2
σ	1.2	1.2	2	1.2	1.2	σ	1.2	1.2	2	1.2	1.2
β	0.005	0.005	0.05	0.1	0.01	β	0.005	0.005	0.05	0.1	0.01
α						α					
T_P						T_P					
0.010	0.095	0.085	0.029	0.077	0.098	0.010	0.043	0.085	0.015	0.037	0.032
0.025	0.168	0.161	0.054	0.141	0.158	0.025	0.076	0.145	0.042	0.082	0.080
0.050	0.255	0.245	0.105	0.236	0.244	0.050	0.136	0.225	0.078	0.152	0.145
0.100	0.368	0.372	0.188	0.361	0.350	0.100	0.248	0.334	0.147	0.222	0.244
0.150	0.463	0.450	0.251	0.444	0.426	0.150	0.315	0.417	0.225	0.302	0.316
T^2						T^2					
0.010	0.087	0.069	0.021	0.064	0.077	0.010	0.030	0.082	0.012	0.023	0.033
0.025	0.147	0.158	0.049	0.131	0.148	0.025	0.065	0.137	0.039	0.068	0.065
0.050	0.224	0.243	0.100	0.221	0.220	0.050	0.110	0.197	0.069	0.123	0.130
0.100	0.340	0.355	0.183	0.342	0.337	0.100	0.212	0.312	0.129	0.231	0.233
0.150	0.441	0.435	0.254	0.440	0.428	0.150	0.292	0.400	0.209	0.293	0.314

In generale è possibile concludere che non ci sono grandi differenze nella potenza tra i due test e che in alcuni casi il test di permutazione ha un comportamento migliore rispetto al test di Hotelling (si veda anche Blair *et al.*, 1994).

5. CONCLUSIONI

I test per ipotesi alternative coerenti rispondono all'esigenza di individuare una procedura di verifica di ipotesi particolarmente indicata per l'analisi di studi osservazionali.

La soluzione di permutazione presenta al riguardo una struttura particolarmente congeniale nel soddisfare le esigenze concrete di ricerca, in quanto consente di analizzare il pattern causale a livello univariato tramite i *test parziali*, a livello multivariato per singolo strato mediante i *test combinati del secondo ordine* e infine a livello globale per l'insieme degli strati mediante il *test combinato del terzo ordine*. Il test multivariato e multistrato inoltre permette di specificare pattern di variabili risposta anche diversificati da strato a strato.

Il test basato sui ranghi proposto da Rosenbaum (1995) prevede al contrario solo la valutazione della significatività del pattern globale a livello multivariato per

tutti gli strati. Inoltre non prevede la possibilità di specificare pattern diversi per ciascuno strato.

Con riferimento al problema del trattamento dei dati mancanti, il test di permutazione multivariato e multistrato consente l'inferenza su tutti i dati osservati, compresi i valori osservati delle unità che presentano dati mancanti. Nel caso di dati mancanti non a caso, è inoltre possibile fare inferenza tenendo conto dell'informazione supplementare dovuta al dato mancante (Pesarin, 2001).

Il test multivariato e multistrato, sopra presentato, permette inoltre di affrontare problemi di verifica di ipotesi particolarmente complessi, ad esempio situazioni in cui il numero di unità di osservazione è minore del numero di variabili, o problemi di inferenza isotonica con alternative di tipo esclusivo (Arboretti, 2002).

A conclusione delle argomentazioni citate, va sottolineato che la flessibilità e la struttura della soluzione di permutazione, rispondono precisamente a quanto viene richiesto al ricercatore per l'analisi e l'interpretazione dei risultati di uno studio denso di implicazioni quale quello osservazionale.

Dipartimento di Scienze Statistiche
Università di Padova

ROSA ARBORETTI GIANCRISTOFARO
MARIO BOLZAN

RIFERIMENTI BIBLIOGRAFICI

R. ARBORETTI GIANCRISTOFARO, (2002), *Multivariate permutation tests in genetics*, "Statistica", LXII, 4 (to appear).

R.C. BLAIR, J.J. HIGGINS, W. KARNISKI, J.D. KROMREY, (1994), *A study of multivariate permutation tests which may replace Hotelling's T^2 test in prescribed circumstances*, "Multivariate Behavioral Research", 29, pp. 141-163.

W.G. COCHRAN, (1965), *The planning of observational studies of human populations (with Discussion)*, "Journal of the Royal Statistical Society", Series A, 128, pp. 134-155.

W.G. COCHRAN, (1968), *The effectiveness of adjustment by subclassification in removing bias in observational studies*, "Biometrics", 24, pp. 295-313.

A. COHEN, H.B. SACKROWITZ, (1998), *Directional tests for one-sided alternatives in multivariate models*, "The Annals of Statistics", 26, 6, pp. 2321-2338.

A.P. DAWID, (1979), *Conditional indipendence in statistical theory (with discussion)*, "Journal of the Royal Statistical Society", Series B, 41, pp. 1-31.

C. HIROTSU, (1998), *Isotonic inference*. In *Encyclopedia of Biostatistics*, Wiley, New York, pp. 2107-2115.

H. JOE, (1997), *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

F. PESARIN, (2001), *Multivariate permutation tests with applications in biostatistics*, Wiley, Chichester.

T. ROBERTSON, F.T. WRIGHT, R.L. DYKSTRA, (1988), *Ordered Restricted Statistical Inference*, Wiley, New York.

P.R. ROSENBAUM, (1994), *Coherence in observational studies*, "Biometrics", 50, pp. 368-374.

P.R. ROSENBAUM, (1995), *Observational Studies*, Springer-Verlag, New York.
P.R. ROSENBAUM, (1997), *Signed rank statistics for coherent predictions*, "Biometrics", 53, pp. 556-566.
P.R. ROSENBAUM, D.B. RUBIN, (1983a), *The central role of the propensity score in observational studies for causal effects*, "Biometrika", 70, pp. 41-55.
P.R. ROSENBAUM, D.B. RUBIN, (1983b), *Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome*, "Journal of the Royal Statistical Society", Series B, 45, pp. 212-218.
P.R. ROSENBAUM, D.B. RUBIN, (1984), *Reducing bias in observational studies using subclassification on the propensity score*, "Journal of the American Statistical Association", 79, 387, pp. 212-218.
D.B. RUBIN, (1997a), *Estimating causal effects from large data sets using propensity scores*, "Annals of Internal Medicine", 127, pp. 757-763.
D.B. RUBIN, (1997b), *Estimation from nonrandomized treatment comparisons using subclassification on propensity scores*, "Proceedings of the International Conference on Nonrandomized Comparative Clinical Studies", Aprile 10-11, Heidelberg.

RIASSUNTO

Metodi non parametrici negli studi osservazionali multivariati in presenza di fattori di confondimento

In questo lavoro viene proposto un percorso di analisi statistica per gli studi osservazionali multivariati in presenza di fattori di confondimento: dalla definizione di una opportuna post-stratificazione mediante il metodo del propensity score, alla verifica di ipotesi non parametrica per ipotesi alternative coerenti.

Infine viene presentato uno studio di simulazione comparativo per confrontare il test basato sui ranghi di Rosenbaum (Rosenbaum, 1995) e il test di permutazione multivariato e multistrato, anche in presenza di dati mancanti.

SUMMARY

Nonparametric methods for multivariate observational studies with confounding factors

In this paper we propose a path of analysis for multivariate observational studies in presence of confounding factors. We start from a suitable post-stratification by using the propensity score method in order to provide hypotheses testing for coherent alternatives based on nonparametric techniques.

Finally a comparative simulation study to compare the Rosenbaum test (Rosenbaum, 1995) and the multivariate and multistrata permutation test is also presented, even in presence of missing observations.