FALSE DISCOVERY RATE: applicazione di un metodo alternativo per i confronti multipli con misure ripetute

MASSIMILIANO PASTORE*, MASSIMO NUCCI, GIOVANNI GALFANO
Università di Cagliari, *Università di Padova* e *Università di Trento*

17 marzo 2005

Sommario

In questo lavoro presentiamo l'applicazione di un metodo alternativo per i confronti multipli tra gruppi di medie ricavati da una serie di misure ripetute. Abbiamo confrontato tale metodo con due procedure. La prima (chiamata α non corretto) consiste in una serie di confronti a coppie separati senza controllo globale dell'errore di I tipo. La seconda è la versione standard del metodo Bonferroni. È noto che, al crescere del numero di confronti, il primo metodo tende ad aumentare l'errore di I tipo e il secondo metodo tende ad aumentare l'errore di II tipo. La procedura che presentiamo, chiamata False Discovery Rate (FDR; Benjamini & Hochberg, 1995), si propone come una soluzione di compromesso tra le due, minimizzando l'errore di II tipo e contemporaneamente tenendo sotto controllo l'errore di I tipo. Il confronto empirico, affiancato da una analisi Bootstrap, ha permesso di evidenziare i numerosi vantaggi del FDR e la sua semplicità di applicazione.

1 Introduzione

Quando si raccolgono dei dati in un esperimento, è pratica comune, per valutare gli effetti dei fattori considerati, ricorrere all'Analisi della Varianza (ANOVA) e, quando si ottiene un risultato statisticamente significativo, effettuare una serie di confronti multipli. In teoria, quando si predispone un esperimento si è guidati da precise ipotesi e si è interessati solo ad alcuni tra questi. Tuttavia, non si può escludere che il ricercatore possa essere interessato a valutare tutti i confronti adottando un approccio di carattere esplorativo. Ad esempio, qualora sia interessato ad investigare fenomeni poco noti oppure voglia condurre una analisi preliminare con l'obiettivo di selezionare gli stimoli più salienti da utilizzare successivamente in un secondo esperimento. È noto però che l'utilizzo di molti confronti multipli comporta un incremento dell'errore di I tipo in quanto, a livello globale, la probabilità di rigettare erroneamente un'ipotesi aumenta con l'aumentare dei confronti (si veda in proposito Keppel, 1991). Nel caso di un numero c di confronti eseguiti, tale probabilità globale (α_{FW}) è pari a

$$\alpha_{FW} = 1 - (1 - \alpha)^c \tag{1}$$

*Gli interessati potranno ottenere ulteriori informazioni circa l'argomento trattato nella presente ricerca scrivendo a Massimiliano Pastore, Dipartimento di Psicologia, Via Is Mirrionis - Loc. Sa Duchessa 09100 Cagliari. E-mail: massimiliano.pastore@unica.it

in cui α è generalmente assunto pari a .05. Se il disegno sperimentale fa uso di più fattori, il numero di confronti possibili, tende inevitabilmente ad aumentare. Ad esempio, in un esperimento con disegno fattoriale 2×3 si ottengono 6 medie. A partire da queste, il numero possibile di confronti è pari a $\binom{6}{2} = 15$ e, pertanto, il valore di α_{FW} è pari a .54, come a dire che la probabilità di trovare significativo per errore almeno uno dei confronti è superiore al 50%. Nel caso di un disegno 3×3, le 9 medie danno luogo a 36 confronti e α_{FW} pari a .84.

La letteratura che tratta il problema dei confronti multipli è vasta e articolata e, già nel 1977, Miller riportava una ricca bibliografia in proposito (per rassegne più recenti si vedano Shaffer, 1995; Westfal, Tobias, Rom, Wolfinger, & Hochberg, 1999).

I classici test utilizzati per i confronti multipli, ad esempio Fisher LSD (Fisher, 1935), Scheffé (1953), Tukey (1951), e Newman-Keuls (Newman, 1939; Keuls, 1952), fanno riferimento a gruppi di medie rilevati su misurazioni indipendenti e campioni che rispettino le assunzioni base dell'ANOVA. Si deve ricordare, però, che molti di questi test, in certi contesti, risultano sconsigliati in quanto non garantiscono sempre un adeguato rapporto tra protezione dall'errore di I tipo e la potenza (si vedano, in proposito, Keselman & Rogan, 1978; Ramsey, 1978, 1981; Einot & Gabriel, 1975; Cramer & Swanson, 1973).

Quando si fa riferimento a disegni sperimentali con misure ripetute, poiché i dati non sono indipendenti, vengono presi in considerazione altri tipi di confronti (Keselman, Keselman & Schaffer, 1991; Keselman & Keselman, 1988; Maxwell, 1980, Keselman, 1982; Keselman, Algina, Kowalchuk, & Wolfinger, 1999). Per esempio, Keselman et al. (1991) presentano delle modifiche al classico metodo di correzione dell'errore di I tipo introdotto da Bonferroni (1936), e che consiste nel dividere la probabilità di soglia desiderata complessivamente (α_{FW}, generalmente .05) per il numero di confronti da effettuare.

Benjamini e Hochberg nel 1995 hanno proposto un metodo che si propone di individuare un buon compromesso tra l'esigenza di tenere sotto controllo il rischio di commettere errori di I tipo, che come noto aumenta all'aumentare dei confronti, e la necessità di evitare una eccessiva riduzione della potenza del test, che diminuisce quanto più si abbassa la probabilità di soglia considerata per l'errore di I tipo.

L'idea cardine di questo metodo è quella di controllare il rapporto tra il numero di ipotesi H^0 rigettate per errore e quelle complessivamente rigettate. Per questa ragione il metodo è stato chiamato *False Discovery Rate* (FDR). La procedura proposta da Benjamini e Hochberg consente in origine di gestire dei confronti multipli su test indipendenti. In un successivo lavoro, Benjamini e Yekutieli (2001) hanno dimostrato che, con una semplice correzione, lo stesso metodo può essere utilizzato anche nel caso di misurazioni non indipendenti e test correlati. FDR presenta almeno tre vantaggi:

1. Funziona bene nei contesti di tipo esplorativo o, come li definisce Tukey (1977), contesti di analisi esplorativa dei dati, quando cioè si è interessati ad esplorare tutte le possibili comparazioni. In questi contesti ci si trova spesso nella necessità di effettuare molti confronti: FDR si offre come uno strumento particolarmente efficace, perché la sua potenza aumenta con l'aumentare dei gruppi sperimentali considerati (Benjamini & Hochberg, 1995; Keselman, Cribbie & Holland, 1999; Williams, Jones & Tukey, 1999).

2. Può essere utilizzato con molti tipi di statistiche, e non solo per valutare differenze tra medie (si vedano Keselman, Cribbie & Holland, 2002), inoltre è *distribution-free* (Genovese & Wasserman, 2002), cioè non necessita di alcuna assunzione a priori sul tipo di distribuzione dei dati.

3. Può essere utilizzato con estrema semplicità sia per confronti tra gruppi indipendenti che nei casi di misurazioni non indipendenti.

Nel nostro lavoro abbiamo voluto prendere in esame l'applicazione di FDR ad una serie tempi di reazione rilevati in un esperimento con disegno a misure ripetute. Siamo convinti che il contesto empirico cui facciamo riferimento sia particolarmente utile in quanto esemplifica delle situazioni piuttosto frequenti nella ricerca psicologica dove spesso le variabili rilevate non sono distribuite normalmente e non sono tra loro indipendenti. Oltre a ricordare che i tempi di reazione sono notoriamente distribuiti in forma non-normale (Luce, 1986; Van Zandt, 2000), vorremmo anche sottolineare che non risultano essere disponibili in letteratura lavori che abbiano applicato FDR nel contesto della cronometria mentale.

Nei paragrafi che seguono, illustreremo il metodo FDR e proporremo un esempio di applicazione per evidenziarne la funzionalità e la semplicità d'uso. Per offrire degli elementi di comparazione, abbiamo utilizzato altri due criteri di confronto. Nel primo, che in realtà è un non-criterio, non è stata apportata alcuna correzione all'errore di I tipo. Chiameremo questo criterio α non corretto (in quanto il valore di α rimane fisso a .05). Nel secondo criterio, comunemente noto come metodo Bonferroni, la probabilità di soglia è stata calcolata in relazione al numero di confronti effettuati. Abbiamo scelto questi due criteri, volutamente estremi, per evidenziare con maggiore chiarezza il comportamento di FDR, che si propone come soluzione di compromesso tra la perdita di potenza e l'incremento dell'errore di I tipo quando si ricorre ai confronti multipli. Infine abbiamo effettuato un'ulteriore analisi con metodo Bootstrap per garantire una maggiore generalizzabilità ai risultati ottenuti dai dati empirici.

2 False Discovery Rate

Supponiamo di voler testare m ipotesi, siano esse $\{H_1^0, H_2^0, ..., H_m^0\}$. Possiamo ipotizzare che m_0 di esse siano vere, anche se non sappiamo quali e quante, e di conseguenza le altre $m - m_0$ siano false. Effettuiamo, per ciascuna di queste ipotesi, un test che ci permetta di decidere se rigettarle o meno. Indichiamo con R il numero di ipotesi rigettate, di conseguenza avremo $m - R$ ipotesi per le quali l'esito del test non è risultato significativo.

È lecito attendersi che, tra le R ipotesi rigettate, possano essercene alcune che sono state respinte per errore. In altre parole, può capitare che una certa ipotesi H_j^0 sia vera, ma l'esito del test ad essa associato sia tale da portarci alla conclusione di considerarla falsa. Se indichiamo con V il numero, ovviamente ignoto, di tali ipotesi, possiamo definire FDR come valore atteso della proporzione di ipotesi rigettate per errore sul totale di ipotesi rigettate, in sintesi: $E(V/R)$. Nella pratica, il controllo FDR si propone di calcolare la probabilità soglia per decidere se rigettare o meno un insieme di ipotesi. Tale probabilità si ottiene nel modo seguente:

1. Si calcola per tutte le ipotesi considerate l'insieme delle statistiche test $\{X_1, X_2, ..., X_m\}$ con le relative probabilità associate a ciascuna di esse $\{P_1, P_2, ..., P_m\}$.

2. Si dispongono le probabilità calcolate in ordine crescente: $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$

3. Si individua quel valore k per cui sia vera la condizione

$$k = max\{i : p_{(i)} \leq \frac{i}{m} q\} \tag{2}$$

in cui q rappresenta la probabilità scelta (generalmente pari a .05).

4. Si rigettano tutte le ipotesi corrispondenti alle probabilità $p_{(1)}, ..., p_{(k)}$, cioè $H^0_{(1)}, ..., H^0_{(k)}$

Il criterio illustrato vale nel caso in cui le ipotesi si basino su osservazioni tra loro indipendenti, qualora ciò non fosse, la (2) viene modificata nel seguente modo:

$$k = max\{i : p_{(i)} \leq \frac{i}{m} \frac{q}{\sum_{i=1}^{m} \frac{1}{i}}\} \tag{3}$$

3 Esperimento

L'esperimento, cui hanno preso parte 20 studenti dell'Università di Padova, consisteva nell'esecuzione di un compito di detezione semplice a stimoli presentati sul monitor di un computer. La variabile dipendente era costituita dai tempi di reazione. Sono stati manipolati due fattori, ciascuno dei quali a tre livelli, ottenendo di conseguenza un disegno 3×3. Non riportiamo i dettagli relativi agli stimoli e ai contenuti dei fattori in quanto non rilevanti per l'oggetto di indagine del presente contributo; basti sapere che ogni soggetto è stato sottoposto a tutte le combinazioni possibili di stimoli.

Dalle analisi preliminari sui dati in oggetto, abbiamo rilevato le medie e le deviazioni standard riportate in Tabella 1. Le correlazioni tra le nove possibili combinazioni dei due fattori manipolati erano comprese tra .71 e .93. Tre di esse non mostravano una distribuzione normale secondo il test di Kolmogorov-Smirnov. Dall'ANOVA, sono risultati significativi sia i fattori principali ($F_{(2,38)} = 3.613$, $p < .05$; $F_{(2,38)} = 198.701$, $p < .001$) sia la loro interazione ($F_{(4,76)} = 3.120$, $p < .05$).

stimolo	N	Minimo	Massimo	Media	Dev. std.
T011	20	594.07	994.20	762.70	112.84
T012	20	520.81	895.56	669.83	100.81
T013	20	476.78	776.50	595.49	98.68
T021	20	571.11	966.18	734.95	121.91
T022	20	469.29	831.72	637.90	106.32
T023	20	477.47	769.39	596.18	78.45
T031	20	585.71	872.64	715.13	85.09
T032	20	529.00	902.72	653.78	107.18
T033	20	477.94	906.47	619.60	113.38

Tabella 1: Statistiche descrittive relative ai tempi di reazione rilevati nei 9 stimoli dell'esperimento.

4 Applicazione del FALSE DISCOVERY RATE

Nell'esperimento da noi preso in esame abbiamo rilevato 9 medie. Avendo ottenuto una interazione significativa, e supponendo di essere interessati a tutti i confronti a coppie possibili, avremo $m = \binom{9}{2} = 36$ ipotesi. Ciascuna di esse si traduce in un confronto tra due medie e $\overline{x}_i$ e $\overline{x}_j$ con $i, j \in \{1, ..., 9\}$, $i \neq j$, e si può esprimere come $H^0_k : \mu_i = \mu_j$, con $k \in \{1, ..., 36\}$, $i \neq j$. Abbiamo pertanto eseguito per 36 volte un test t per campioni appaiati, uno per ciascuna delle coppie ottenibili, e disposto in ordine crescente le probabilità associate ai vari test.

Utilizzando il criterio espresso dalla (3) abbiamo poi individuato la probabilità soglia per il controllo FDR, che risulta essere pari a .0079 e, di conseguenza, rigettato tutte le ipotesi con probabilità associate minori o uguali a tale valore.

			tipo di controllo		
		$P(2code)$	.05	Bonferroni	FDR
Coppia 1	T011 - T013	0.0000	**	**	**
Coppia 2	T013 - T021	0.0000	**	**	**
Coppia 3	T023 - T031	0.0000	**	**	**
Coppia 4	T021 - T033	0.0000	**	**	**
Coppia 5	T013 - T031	0.0000	**	**	**
Coppia 6	T011 - T023	0.0000	**	**	**
Coppia 7	T011 - T022	0.0000	**	**	**
Coppia 8	T021 - T023	0.0000	**	**	**
Coppia 9	T011 - T033	0.0000	**	**	**
Coppia 10	T021 - T032	0.0000	**	**	**
Coppia 11	T031 - T033	0.0000	**	**	**
Coppia 12	T011 - T032	0.0000	**	**	**
Coppia 13	T012 - T023	0.0000	**	**	**
Coppia 14	T021 - T022	0.0000	**	**	**
Coppia 15	T011 - T012	0.0000	**	**	**
Coppia 16	T012 - T013	0.0000	**	**	**
Coppia 17	T031 - T032	0.0000	**	**	**
Coppia 18	T022 - T031	0.0001	**	**	**
Coppia 19	T013 - T032	0.0001	**	**	**
Coppia 20	T023 - T032	0.0001	**	**	**
Coppia 21	T012 - T021	0.0002	**	**	**
Coppia 22	T012 - T033	0.0004	**	**	**
Coppia 23	T032 - T033	0.0030	**	n.s.	**
Coppia 24	T012 - T031	0.0030	**	n.s.	**
Coppia 25	T011 - T031	0.0044	**	n.s.	**
Coppia 26	T013 - T022	0.0079	**	n.s.	**
Coppia 27	T022 - T023	0.0153	**	n.s.	n.s.
Coppia 28	T011 - T021	0.0582	n.s.	n.s.	n.s.
Coppia 29	T013 - T033	0.0669	n.s.	n.s.	n.s.
Coppia 30	T023 - T033	0.0682	n.s.	n.s.	n.s.
Coppia 31	T012 - T022	0.0872	n.s.	n.s.	n.s.
Coppia 32	T021 - T031	0.1872	n.s.	n.s.	n.s.
Coppia 33	T012 - T032	0.2070	n.s.	n.s.	n.s.
Coppia 34	T022 - T033	0.3151	n.s.	n.s.	n.s.
Coppia 35	T022 - T032	0.3773	n.s.	n.s.	n.s.
Coppia 36	T013 - T023	0.9592	n.s.	n.s.	n.s.

Tabella 2: Confronto tra i risultati ottenuti dai tre tipi di controllo considerati: α (.05) non corretto, metodo Bonferroni e False Discovery Rate. Il simbolo (**) indica che il test è statisticamente significativo e che l'ipotesi H^0 corrispondente viene rigettata. Le probabilità associate ai confronti sono disposte in ordine crescente.

In Tabella 2 sono riportati i risultati ottenuti con questa procedura confrontati con quelli ricavati con gli altri due metodi. Nella prima colonna (α non corretto) si considerano significativi i test con probabilità associata inferiore a .05. Nella seconda colonna (Bonferroni) sono riportati i risultati ottenuti con l'applicazione del controllo Bonferroni. Secondo tale metodo, si ottiene come probabilità soglia $\alpha_B = \frac{\alpha_{FW}}{36} = \frac{.05}{36} = .001$; ne consegue che risultano significativi i confronti con probabilità inferiore a .001. Infine, nella terza colonna (FDR), sono indicati i confronti che risultano statisticamente significativi con il controllo FDR. Si nota che il numero di confronti significativi di questa ultima colonna (26) è superiore a quello individuato con il metodo di Bonferroni (22) e comunque inferiore al numero individuato con i semplici confronti a coppie (27), come era preventivabile in virtù del fatto che gli altri due metodi scelti sono, rispettivamente, il più ed il meno conservativo.

Sulla base di questo unico campione, non si possono trarre conclusioni generali sui rapporti tra i tre metodi, tuttavia, abbiamo voluto effettuare un analisi Bootstrap sui nostri dati così da valutare la consistenza della relazione d'ordine osservata.

5 Bootstrap

Il Bootstrap è un metodo che consente di studiare le proprietà rilevanti di una qualche statistica T a partire da un campione di dati empirico. A differenza del classico metodo Monte Carlo, in cui i dati vengono generati a partire da un modello teorico imposto, il Bootstrap utilizza il campione come se fosse la popolazione (Gentle, 2002; Hinkley, 1988). Più precisamente, il metodo consiste nell'estrarre dai dati osservati una serie di campioni con ripetizione e, su questi, calcolare ogni volta la statistica T. In questo modo, si può ottenere una distribuzione campionaria di T e calcolare, ad esempio, l'intervallo di confidenza ed il bias (lo scarto tra la statistica osservata nel campione e la media della distribuzione ottenuta).
A partire dai nostri dati, abbiamo effettuato una analisi con metodo Bootstrap bilanciato, così come suggerito da Davison, Hinkley e Schechtman (1986), utilizzando uno degli algoritmi proposti da Gleason (1988) con una routine appositamente scritta in ambiente R (R Development Core Team, 2003).
Abbiamo eseguito 1000 replicazioni e considerato come statistiche oggetto T il numero di confronti significativi ottenuti con i tre metodi in ciascuno dei campioni così prodotti. Siano di conseguenza ${}_{.05}T$ il numero ottenuto nei confronti semplici con alpha non corretto, ${}_{B}T$ il numero ottenuto con controllo Bonferroni, e ${}_{FDR}T$ il numero ottenuto con controllo FDR. Per ciascuna replicazione (j), ottenuta dal campionamento con ripetizione dai dati originari, abbiamo calcolato le tre statistiche T ed ottenuto le rispettive stime ${}_{.05}T^*_j$, ${}_{B}T^*_j$, ${}_{FDR}T^*_j$.
In figura 1 abbiamo riportato le distribuzioni di frequenza delle tre statistiche nelle 1000 replicazioni. Mantenendo fisso il valore di α a .05, abbiamo ottenuto in media 28.979 confronti significativi, con un errore standard pari a 1.414 ed un bias di 1.979. È opportuno ricordare che, in questo caso, la probabilità associata all'errore di I tipo è pari a .84. Con il controllo Bonferroni il numero medio di confronti significativi si abbassa a 23.105, con un errore standard di 1.733 ed un bias di 1.105. Utilizzando, infine, il confronto FDR sono risultati significativi in media 25.936 confronti, con un errore standard di 1.71 ed un bias di -0.074. In Tabella 3 abbiamo riassunto questi risultati calcolando anche l'intervallo di confidenza delle medie considerando i percentili delle tre funzioni di distribuzione ottenute nelle 1000 replicazioni, rispettivamente ${}_{.05}T^*$, ${}_{B}T^*$, ${}_{FDR}T^*$.

	${}_{.05}T^*$	${}_{B}T^*$	${}_{FDR}T^*$
media	28.979	23.105	25.936
int. conf. 95%	26 - 32	20 - 27	23 - 29
err. st.	1.414	1.733	1.71
bias	1.979	1.105	-0.074

Tabella 3: Risultati dell'analisi con metodo bootsrap bilanciato, 1000 replicazioni.

Come si può osservare, il valore medio di ${}_{FDR}T^*$ risulta più vicino a quello effettivamente ottenuto nel nostro campione rispetto agli altri due metodi considerati e si colloca in posizione intermedia, in linea con quanto era lecito prevedere.

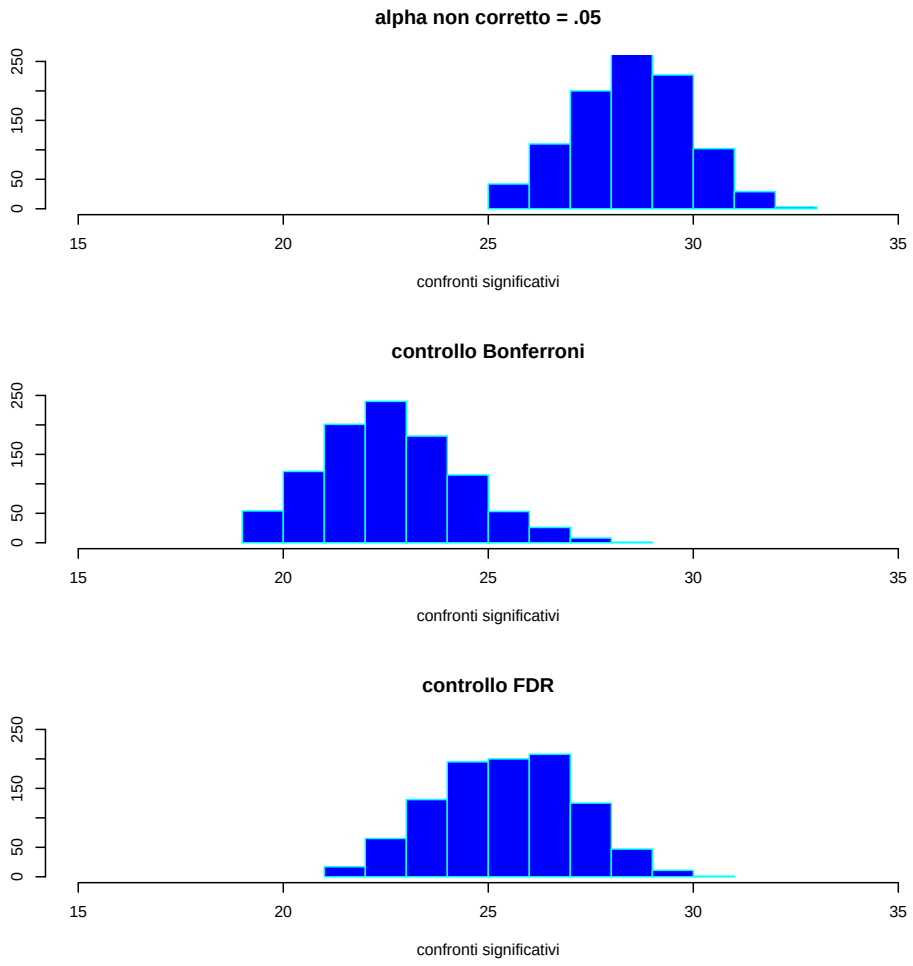

Figura 1: Frequenze dei confronti risultati significativi su 1000 replicazioni considerando i confronti semplici con alfa non corretto (.05), con la correzione di Bonferroni e con controllo FDR.

6 Conclusioni

Nel presente lavoro abbiamo presentato una possibile applicazione di FDR nel contesto della cronometria mentale. Rispetto al metodo che abbiamo chiamato α non corretto, FDR introduce un controllo sull'errore di I tipo altrimenti non considerato. Nello stesso tempo, è noto dalla letteratura che FDR è più potente rispetto al metodo Bonferroni, in quanto riduce la possibilità di commettere errori di II tipo (Benjamini & Hochberg, 1995; Keselman et al., 1999a). In particolare FDR risulta tanto più potente quante più risultano essere le ipotesi H^0 false (Benjamini & Hochberg, 1995, Williams et al., 1999).
Va sottolineato che il metodo può essere vantaggioso per chi fa ricerca in psicologia per almeno tre motivi. Innanzitutto, può essere utilizzato con un'ampia tipologia di statistiche, quali medie, correlazioni, proporzioni, et cetera (es., Keselman et al., 2002), tutte di largo impiego in ambito nella ricerca psicologica. Inoltre, FDR non fa particolari assunzioni circa la distribuzione della statistica considerata, un aspetto cruciale, ma talvolta ingiustamente trascurato nella pratica della ricerca. Frequentemente infatti, in psicologia, le variabili osservate non hanno le caratteristiche distributive che consentono un uso ottimale dei test classici. Ancora, FDR è un metodo estremamente flessibile perchè non necessita di particolari assunzioni. Una possibile dimostrazione di questa proprietà è rappresentata proprio dal fatto che FDR può essere applicato liberamente in un contesto come quello da noi esaminato, in cui le variabili misurate (tempi di reazione) sono tra loro fortemente correlate e potenzialmente non normali. Infine, è importante sottolineare come FDR sia un metodo particolarmente semplice da applicare (si vedano in proposito gli esempi riportati da Thissen, Steinberg & Kuang, 2002).
L'applicazione proposta ha cercato di rendere espliciti i vantaggi cha abbiamo appena discusso. Con l'analisi Bootstrap, è stato possibile definire, per ciascuno dei criteri considerati, un intervallo di confidenza relativo al numero di confronti significativi ed osservare così che quello associato a FDR risulta intermedio agli altri due, come era lecito attendersi.
FDR è stato applicato con successo in numerosi ambiti scientifici, come la biologia molecolare (Vogel, Berzuini, Bashton, Gough, & Teichmann, 2004), la genetica (Reiner, Yekutieli, & Benjamini, 2003), le scienze naturali (Basford & Tukey, 1997; Garcia, 2003; Ottaviani, Ji, & Pastore, 2003), le scienze economiche (Schaffer & Green, 1998), e le scienze mediche (Kanas, Salnitskiy, Grund, Gushin, Weiss, Kozerenko, Sled & Marmar, 2000; Vedantham, Brunet, Boyer, Weiss, Metzler, & Marmar, 2001). Recentemente, FDR ha destato grande interesse nel settore delle neuroscienze, in particolare per quanto riguarda studi neuropsicologici (es., Dohmes, Zwitserlood, & Bölte, in stampa) e soprattutto studi sui correlati neurali delle funzioni cognitive attraverso l'impiego di metodi di neuroimmagine (es., Friston, Glaser, Henson, Kiebel, Phillips & Ashburner, 2002; Nichols & Hayasaka, 2003, Logan & Rowe, 2004; Turkheimer, Smith, & Schmidt, 2001). Genovese, Lazar e Nichols (2002), per esempio, hanno dimostrato come tale metodo risulti essere particolarmente appropriato nel contesto delle analisi relative a dati ottenuti attraverso tecniche di neuroimmagine per determinare quali unità nelle mappe cerebrali vengano attivate durante l'esecuzione di un determinato compito sperimentale.
Alla luce di queste considerazioni, crediamo che FDR possa rappresentare un utile strumento in tutti quei casi, frequenti in psicologia, in cui si renda necessario eseguire molti confronti statistici.

Riferimenti bibliografici

[1] BASFORD, K. E., TUKEY, J. W. (1997). Graphical profiles as an aid to understanding plant breeding experiments. *Journal of Statistical Planning and Inference*, **57**, 93-107.

[2] BENJAMINI, Y., HOCHBERG, Y. (1995). Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistic Society B*, **57**, 289–300.

[3] BENJAMINI, Y., YEKUTIELI, D. (2001). The control of the False Discovery Rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1188.

[4] BONFERRONI, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.

[5] CRAMER, S. G., SWANSON, M. R. (1973). An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association*, **68**, 66-74.

[6] DAVISON, A. C., HINKLEY, D. V., SCHECHTMAN, E. (1986). Efficient boostrap simulation. *Biometrika*, **73**, 555-566.

[7] DOHMES, P., ZWITSERLOOD, P., BÖLTE, J. (in stampa). The impact of semantic transparency of morphologically complex words on picture naming. *Brain and Language*.

[8] EINOT, I., GABRIEL, K. R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*, **70**, 574-583.

[9] FISHER, R. A. (1935). *The Design of Experiments*. London: Oliver and Boyd.

[10] FRISTON, K. J., GLASER, D. E., HENSON, R. N. A., KIEBEL, S., PHILLIPS C., ASHBURNER, J. (2002). Classical and bayesian inference in neuroimaging: Applications. *NeuroImage*, **16**, 484-512.

[11] GARCIA, L. V. (2003). Controlling the false discovery rate in ecological research. *Trends in Ecology and Evolution*, **18**, 553-554.

[12] GENOVESE, C. R., WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistic Society B*, **64**, 499-517.

[13] GENOVESE, C. R., LAZAR, N. A., NICHOLS, T. (2002). Thresholding of statistical maps in functional neuroimaging using the False Discovery Rate. *NeuroImage*, **15**, 870–878.

[14] GENTLE, J. E. (2002). *Elements of computational statistics*. Heidelberg: Springer Verlag.

[15] GLEASON, J. R. (1988). Algorithms for balanced bootstrap simulations. *American Statistician*, **42**, 263-266.

[16] HINKLEY, D. V. (1988). Bootstrap methods. *Journal of the Royal Statistic Society B*, **50**, 321-337.

[17] KANAS, N., SALNITSKIY, V., GRUND, E. M., GUSHIN, V., WEISS, D. S., KOZERENKO, O., SLED, A., MARMAR, C. R. (2000). Social and cultural issues during shuttle/mir space missions. *Acta Astronautica*, **47**, 647-655.

[18] KEPPEL, G. (1991). *Design and analysis. A researcher's handbook.* Upper Saddle River, NJ: Prentice Hall.

[19] KESELMAN, H. J. (1982). Multiple comparisons for repeated measures means. *Multivariate Behavioral Research*, **17**, 87-92.

[20] KESELMAN, H. J. & KESELMAN, J. C. (1988). Repeated Measures Multiple Comparison Procedures: Effects of Violating Multisample Sphericity in Unbalanced Designs. *Journal of Educational Statistics*, **13**, 215-226.

[21] KESELMAN, H. J. & ROGAN, J. C. (1978). A Comparison of the Modified-Tukey and Scheffé Methods of Multiple Comparisons for Pairwise Contrast. *Journal of the American Statistical Association*, **73**, 47-52.

[22] KESELMAN, H. J., CRIBBIE, R., HOLLAND, B. (1999a). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise/comparisonwise Type I error control. *Psychological Methods*, **4**, 58-69.

[23] KESELMAN, H. J., CRIBBIE, R., HOLLAND, B. (2002). Controlling the rate of Type I error over a large set of statistical tests. *British Journal of Mathematical and Statistical Psychology*, **55**, 27-39.

[24] KESELMAN, H. J., KESELMAN, J. C., SHAFFER, J. P. (1991). Multiple pairwise comparisons of repeated measures means under violations of multisample sphericity. *Psychological Bulletin*, **110**, 162-170.

[25] KESELMAN, H. J., ALGINA, J., KOWALCHUK, R. K., WOLFINGER, R. D. A. (1999b). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology*, **52**, 63-78.

[26] KEULS, M. (1952). The use of Studentized range in connection with an analysis of variance. *Euphytica*, **1**, 112-122.

[27] LOGAN, B. R., ROWE, D. B. (2004). An evaluation of thresholding techniques in fMRI analysis. *NeuroImage*, **22**, 95-108.

[28] LUCE, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

[29] MAXWELL, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. *Journal of Educational Statistics*, **13**, 269-287.

[30] MILLER, R. G. (1977). Developments in multiple comparisons 1966-1976. *Journal of the American Statistical Association*, **72**, 779-788.

[31] NEWMAN, D. (1939). The distribution of the range in samples from the normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika*, **31**, 20-30.

[32] NICHOLS, T., HAYASAKA, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research*, **12**, 419-446.

[33] OTTAVIANI, D., JI, L., PASTORE, G. (2003). A multidimensional approach to understanding agro-ecosystems. A case study in Hubei Province, China. *Agricultural Systems*, **76**, 207-225.

[34] R DEVELOPMENT CORE TEAM (2003). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

[35] RAMSEY, P. H. (1978). Power Differences Between Pairwise Multiple Comparisons. *Journal of the American Statistical Association*, **73**, 479-485.

[36] RAMSEY, P. H. (1981). Power of univariate pairwise multiple comparison procedures. *Psychological Bulletin*, **90**, 352-366.

[37] REINER, A., YEKUTIELI, D., BENJAMINI, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368-375.

[38] SHAFFER, J. P. (1995). Multiple hypothesis testing: A review. *Annual Review of Psychology*, **46**, 561-584.

[39] SHAFFER, C. M., GREEN, P. E. (1998). Cluster based market segmentation: Some further comparisons of alternative approaches. *Journal of the Market Research Society*, **35**, 155-163.

[40] SCHEFFÉ, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, **40**, 87-104.

[41] THISSEN, D., STEINBERG, L., KUANG, D. (2002). Quick and easy implementation of the Benjamini - Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, **27**, 77–83.

[42] TUKEY, J. W. (1951). Quick and dirty methods in statistics. Part II. Simple analysis for standard designs. *Proceedings of the Fifth Annual Convention of the American Society for Quality Control*, 189-197.

[43] TUKEY, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

[44] TURKHEIMER, F. E., SMITH, C. B., SCHMIDT, K. (2001). Estimation of the number of true null hypotheses in multivariate analysis of neuroimaging data. *NeuroImage*, **13**, 920-930.

[45] VAN ZANDT, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin and Review*, **7**, 424-465.

[46] VEDANTHAM, K., BRUNET, A., BOYER, R., WEISS, D. S., METZLER, T. J., MARMAR, C. R. (2001). Posttraumatic stress disorder, trauma exposure, and the current health of Canadian bus drivers. *Canadian Journal of Psychiatry*, **46**, 149-155.

[47] VOGEL, C., BERZUINI, C., BASHTON, M., GOUGH, J., TEICHMANN, S. A. (2004). Supra-domains: Evolutionary units larger than single protein domains. *Journal of Molecular Biology*, **336**, 809-823.

[48] WESTFALL, P. H., TOBIAS, R. D., ROM, D., WOLFINGER, R. D., HOCHBERG, Y. (1999). *Multiple comparisons and multiple tests*. Cary, NC: SAS Institute, Inc.

[49] WILLIAMS, V. S. L., JONES, L. V., TUKEY, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, **24**, 42-69.