Valutazione della potenza nei confronti multipli: cinque metodi a confronto.

Massimo Nucci[1], Massimiliano Pastore[2], Giovanni Galfano[3]

[1] Dipartimento di Psicologia Generale, Università di Padova, via Venezia, 8, I-35100 Padova, Italy. Email: massimo.nucci@unipd.it

[2] Dipartimento di Psicologia, Università di Cagliari, via Is Mirrionis - Loc. Sa Duchessa, I-09100 Cagliari, Italy. Email: massimiliano.pastore@unica.it

[3] Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università di Padova, via Venezia, 8, I-35100 Padova, Italy. Email: giovanni.galfano@unipd.it

Sommario Il lavoro propone un paragone delle potenze, in un contesto di confronti multipli, fra tre metodi classici (Scheffé, Tukey, Bonferroni) e due metodi basati sul False Discovery Rate (FDR). Con una simulazione Monte Carlo sono stati effettuati tutti i confronti a coppie possibili tra k (k = 3,4,5,6) campioni di numerosità 20, manipolando il valore dell'effect size. I metodi sono stati confrontati sulla base di tre definizioni di potenza ammissibili in vari contesti di ricerca psicologica. I risultati hanno evidenziato una migliore prestazione dei metodi basati sul FDR rispetto a quelli classici.

Abstract. In this paper, we present a comparison, in terms of power, among three classic procedures (Scheffé, Tukey, Bonferroni) and two methods based on the False Discovery Rate (FDR) in the context of multiple comparison. A Monte Carlo simulation is illustrated, in which we performed all the possible pairwise comparisons among k (k = 3,4,5,6) samples with fixed numerosity (20), by manipulating effect size. Three different definitions of power are considered, based on their possible application in psychological research. The results show that FDR-based procedures offer a better performance compared to the classical methods.

Introduzione

Nella ricerca psicologica è spesso necessario eseguire molti confronti statistici. Tuttavia, al crescere del numero di test eseguiti aumenta anche la probabilità di commettere errori di I tipo, cioè di rifiutare un'ipotesi nulla che invece sia vera. Per affrontare questo problema sono disponibili specifiche procedure: metodi post hoc, confronti multipli, contrasti ecc. Molte procedure però comportano un'elevata perdita di potenza, con una riduzione della capacità del test di respingere correttamente un'ipotesi nulla falsa (Ramsey, 2002). Recentemente, alcuni autori (ad esempio Benjamini e Hochberg, 1995) hanno sviluppato nuovi metodi che hanno affiancato quelli più tradizionali (Pastore, Nucci e Galfano, 2005). Più volte è stato evidenziato come si ponga più attenzione all'errore di I tipo che alla potenza di un test (Cohen, 1988; Keppel, 1991; Bachmann, Luccio e Salvadori, 2005). Sembra sempre più necessario, dunque, non solo conoscere le condizioni d'applicabilità dei test, ma anche saper scegliere i metodi che offrono una maggiore potenza. In questa sede, ci proponiamo di paragonare le potenze di cinque metodi per i confronti multipli. Per mezzo di una simulazione Monte Carlo, sono stati considerati tre metodi tradizionali, Scheffè (1953, SH), Tukey (1953, TK), Bonferroni (1936, B), e due metodi più recenti: la prima originale procedura basata sul False Discovery Rate (FDR) (Benjamini e Hochberg 1995, BH) ed una più recente combinazione del FDR con le tecniche di ricampionamento (Reiner, Yekutieli e Benjamini, 2003, RYB). Il confronto è stato effettuato in funzione del numero di medie e dell'effect size calcolato. I risultati sono stati esaminati in riferimento a tre diverse definizioni di potenza ammissibili nel contesto dei confronti multipli: *per-power*, *any-power*, e *all-power* (Ramsey, 1978). La potenza *all-power* si basa sul corretto rifiuto di tutte le ipotesi nulle quando le medie confrontate appartengano a popolazioni distinte. Questo approccio è certamente da preferire quando il mancato rifiuto di un'ipotesi nulla non vera ha delle conseguenze particolarmente gravi (ad esempio la mancata diagnosi precoce di un disturbo o il mancato riconoscimento di non idoneità psicofisica a lavori rischiosi). Tuttavia, si utilizza più spesso la definizione di potenza *any-power*, basata sul corretto rifiuto di almeno una ipotesi nulla tra quelle false. Tale approccio, pur con un maggior rischio di errori di I tipo, offre una maggiore capacità discriminativa, particolarmente utile nelle ricerche esplorative. La potenza

per-power è una soluzione intermedia tra le due precedenti e può essere utile in contesti meta-analitici (Westfall e Young, 1993). La scelta di prendere in considerazione tre diverse definizioni di potenza è giustificata sia per illustrare in maniera più dettagliata le possibili variazioni nella prestazione dei metodi esaminati, sia per dare una panoramica d'applicazione esaustiva nei diversi contesti della ricerca psicologica.

Descrizione delle procedure

Ogni procedura è stata impiegata per testare m ipotesi, $\{H_1, ..., H_m\}$, per ciascuna delle quali viene effettuato un test e calcolata una probabilità p_h (con $h = 1, ..., m$). I vari metodi sono stati utilizzati per determinare le rispettive probabilità aggiustate secondo le procedure descritte brevemente di seguito. Mentre i metodi SH e TK richiedono di effettuare l'ANOVA, perchè l'aggiustamento dipende dalla stima della varianza d'errore, per gli altri metodi sono stati utilizzati dei semplici t-test.

Metodo SH. La probabilità aggiustata è calcolata utilizzando la seguente statistica:

$$\hat{F} = \frac{n(\sum_i c_i \bar{X}_i)^2}{(k-1)MS_{err}(\sum_i c_i^2)}$$

in cui n è la numerosità di ciascun gruppo, $\bar{X}_i$ la media del gruppo i-mo, k il numero di gruppi, MS_{err} la varianza d'errore, ricavata dall'ANOVA, e c_i dei pesi scelti in modo che $\sum c_i = 0$. La probabilità associata alla statistica $\hat{F}$ è data dalla distribuzione F con gradi di libertà $(k-1)$ e quelli relativi a MS_{err}.

Metodo TK. La probabilità aggiustata è calcolata utilizzando la seguente statistica basata sulla Studentized Range Distribution:

$$q_{ij} = \frac{\bar{X}_{(i)} - \bar{X}_{(j)}}{\sqrt{\frac{MS_{err}}{n}}};\ i > j$$

in cui i e j indicano il rango occupato nell'ordinamento crescente delle medie, $\bar{X}_i$ è la media campionaria, MS_{err} la varianza dell'errore ottenuta dall'ANOVA e n la numerosità campionaria di ciascun gruppo.

Metodo B. La probabilità viene aggiustata semplicemente moltiplicando il valore osservato (p_h) per il numero di confronti effettuati (m).

$$\widetilde{p}_h = \min(p_h m, 1)$$

Metodo BH. Siano $p_{(1)} \leq ... \leq p_{(m)}$ le probabilità osservate, disposte in ordine crescente; le probabilità aggiustate saranno date dalla relazione:

$$\widetilde{p}_h = \min\{p_{(j)} \frac{m}{j} : h \leq j\};\ j \in \{1, ..., m\}$$

Metodo RYB. Nella prima fase, i dati sono ricampionati per un numero B di volte e dei valori di probabilità sono stimati sulla base delle statistiche ottenute dal ricampionamento. Per la h-ma ipotesi, con una statistica osservata pari a t_h, la probabilità stimata sarà data dalla seguente formula:

$$p_h^{est} = \frac{\#\{t_{ij}^* : |t_{ij}^*| \geq |t_h|\}}{mB}$$

in cui $i = 1, ..., m$, con m il numero di ipotesi, $j = 1, ..., N$, con B il numero di ricampionamenti, e t_{ij}^* indica la statistica ottenuta con ricampionamento. Nella seconda fase, le probabilità ottenute sono aggiustate con il metodo BH descritto sopra.

Schema sperimentale

Il programma per la simulazione Monte Carlo è stato scritto in ambiente R (R Development Core Team, 2005) ed ha struttura affine allo schema sperimentale descritto da Ramsey (1978). Sono state considerate delle popolazioni normali con varianze fissate ad 1 e medie ($\mu_i; i = 1, ..., k$) scelte in modo da riprodurre degli specifici valori di effect size secondo la f di Cohen (1988, p. 275) con:

$$f = \sigma_m / \sigma$$

in cui valgono le seguenti relazioni:

$$\sigma_m^2 = \sum_{i=1}^{k} (\mu_i - \mu)^2 / k \qquad \text{e} \qquad \mu = \sum_{i=1}^{k} \mu_i / k$$

e dove f indica l'effect size, μ la media generale delle medie e σ_m^2 la varianza relativa alle medie delle popolazioni. Si può notare che, con questa definizione, a parità di effect size è possibile individuare combinazioni praticamente infinite di valori medi. Ad esempio, ponendo $k = 3$ e $\sigma = 1$ è facile verificare che con i valori $\mu_i = (-1, 0, 1)$ si ottiene lo stesso valore di f che con $\mu_i = (0, 0, 1.732)$. David, Lachenbruch e Brandis (1972) hanno dimostrato che comunque, per ciascun valore specificato di f, le configurazioni possibili hanno un ambito di variazione massimo ed uno minimo. L'ambito di variazione massimo è dato da

$$\mu_1 = -(k/2)^{1/2} f; \qquad \mu_2 = ... = \mu_{k-1} = 0; \qquad \mu_k = (k/2)^{1/2} f$$

L'ambito di variazione minimo, se k è pari, è dato da

$$\mu_1 = ... = \mu_{k/2} = -f; \qquad \mu_{(k/2)+1} = ... = \mu_k = f$$

se k è dispari è dato da

$$\mu_1 = ... = \mu_{(k+1)/2} = -[(k-1)/(k+1)]^{1/2} f; \qquad \mu_{(k+3)/2} = ... = \mu_k = [(k-1)/(k+1)]^{1/2} f$$

Le condizioni sperimentali sono state scelte ponendo il numero di medie $k = 3, 4, 5, 6$ e l'effect size $f = 0, 0.1, ..., 1.2$. La numerosità dei campioni generati è stata fissata a $n = 20$. Per ciascuna combinazione di k e f sono state effettuate 1000 repliche per la configurazione ambito massimo e 1000 per la configurazione ambito minimo. Sono stati dunque simulati $1000 \times 4 \times 13 \times 2 = 104000$ esperimenti. Su ciascuno dei campioni generati sono state messe a confronto le 5 procedure descritte sopra. Soltanto per la procedura RYB, ogni campione generato è stato ulteriormente ricampionato 1000 volte.

Stima della potenza e della probabilità di errore di I tipo

La potenza è stata stimata considerando le condizioni con $f > 0$, adottando le tre seguenti definizioni:

- *per-power*: probabilità di rigettare una generica ipotesi H_0 falsa; viene stimata contando il numero di ipotesi H_0 rigettate sul totale di ipotesi false.

- *any-power*: probabilità di rigettare almeno una ipotesi H_0 falsa per ciascun valore di f; viene stimata contando quante volte viene rigettata almeno una ipotesi falsa per ciascuna delle 1000 repliche dell'esperimento.
- *all-power* = probabilità di rigettare correttamente tutte le ipotesi H_0 false per ciascun valore di f; viene stimata contando quante volte vengono rigettate tutte le ipotesi false per ciascuna delle 1000 repliche dell'esperimento.

La probabilità di errore di I tipo è stata stimata empiricamente considerando le condizioni in cui $f = 0$, cioè i casi in cui non esiste differenza tra le medie delle popolazioni. Anche per l'errore di I tipo abbiamo adottato le stesse definizioni e lo stesso metodo di stima della probabilità, considerando che, in questi casi, le ipotesi sono vere.

Risultati

In Figura 1 abbiamo riportato le curve di potenza stimate, ottenute come media tra i valori di ambito massimo e minimo, per medie. Nel grafico è riportata anche la curva relativa alle probabilità non aggiustate (rawp) per avere un ulteriore termine di paragone. Come era lecito prevedere, con sole 3 medie i vari metodi non presentano evidenti differenze tra loro, in particolare quando si adotta una definizione di tipo *any-power*. Con le altre due definizioni, i metodi migliori risultano essere quelli basati sul FDR (BH e RYB), tra loro praticamente equivalenti, mentre il peggiore è il metodo SH.

In Figura 2 sono riportate le curve di potenza stimate, ottenute come media tra i valori di ambito massimo e minimo, per $k = 6$ medie. Le differenze tra i metodi sono più evidenti, soprattutto per le definizioni *per-power* e *all-power*. In entrambi i casi si osserva la stessa relazione d'ordine, con i metodi BH e RYB risultati più potenti, seguiti da TK, B e, per ultimo, SH. Adottando la definizione *any-power* invece, il metodo che si rivela più potente, anche se di poco, è TK. Le curve di potenza stimate relative a $k = 4$ e $k = 5$ medie si collocano in una posizione intermedia rispetto ai valori estremi illustrati nelle Figure 1 e 2 e non necessitano dunque di ulteriori commenti.

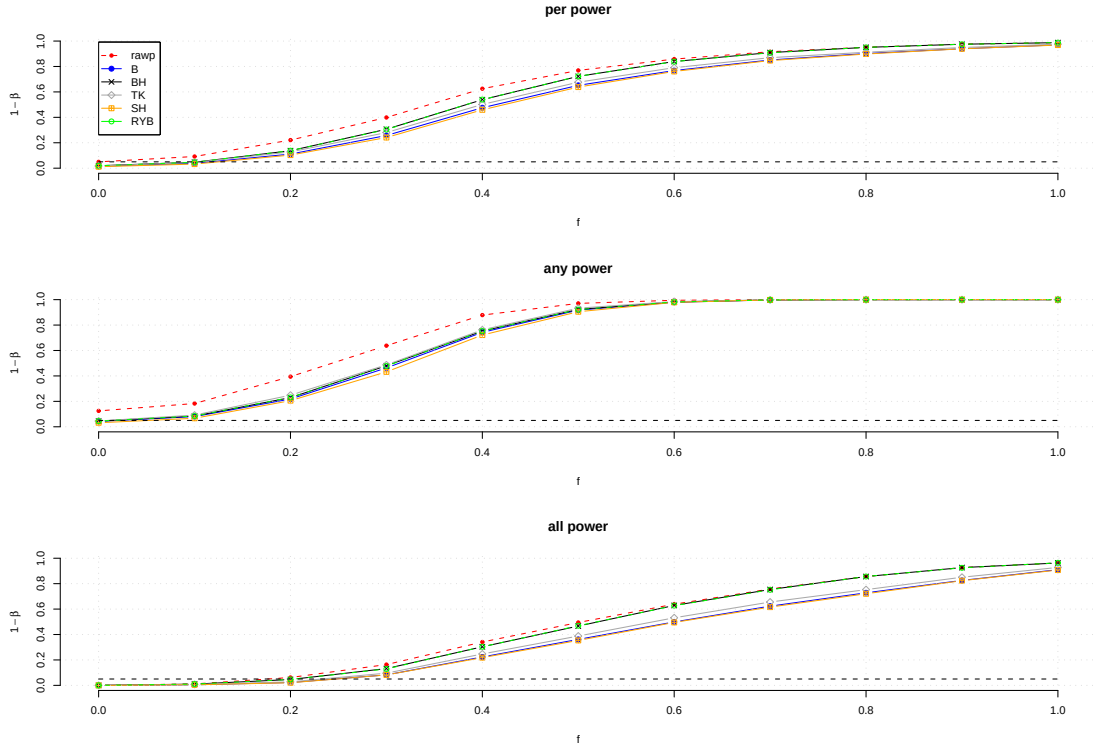

Figura 1. Curve di potenza stimate con $k = 3$ medie. Le curve sono ottenute come media tra le configurazioni di variazione massima e minima. La linea tratteggiata indica la soglia 0.05. In rosso le probabilità non aggiustate.

Conclusioni

Abbiamo messo confrontato le potenze di 5 diversi metodi per effettuare i confronti multipli. Le differenti definizioni di potenza adottate nel calcolo dei risultati mostrano come criteri più restrittivi comportino una minore potenza espressa. In particolare, la definizione *all-power* risulta essere meno sensibile delle altre.

La differenza tra i 5 metodi impiegati è modesta in contesti dove il numero di confronti è relativamente basso ($k = 3$), mentre cresce con il numero di medie. In genere, la relazione d'ordine tra le potenze dei metodi è stabile, risultando quasi ovunque valido l'ordine decrescente BH, RYB, TK, B, SH (con i primi due metodi pressoché di uguale potenza). Tutti i metodi utilizzati risultano adeguati al controllo dell'errore di primo tipo, mantenendo il valore sotto la soglia 0.05. In definitiva emerge chiaramente come i metodi basati sull'approccio FDR (BH e RYB) siano glo-

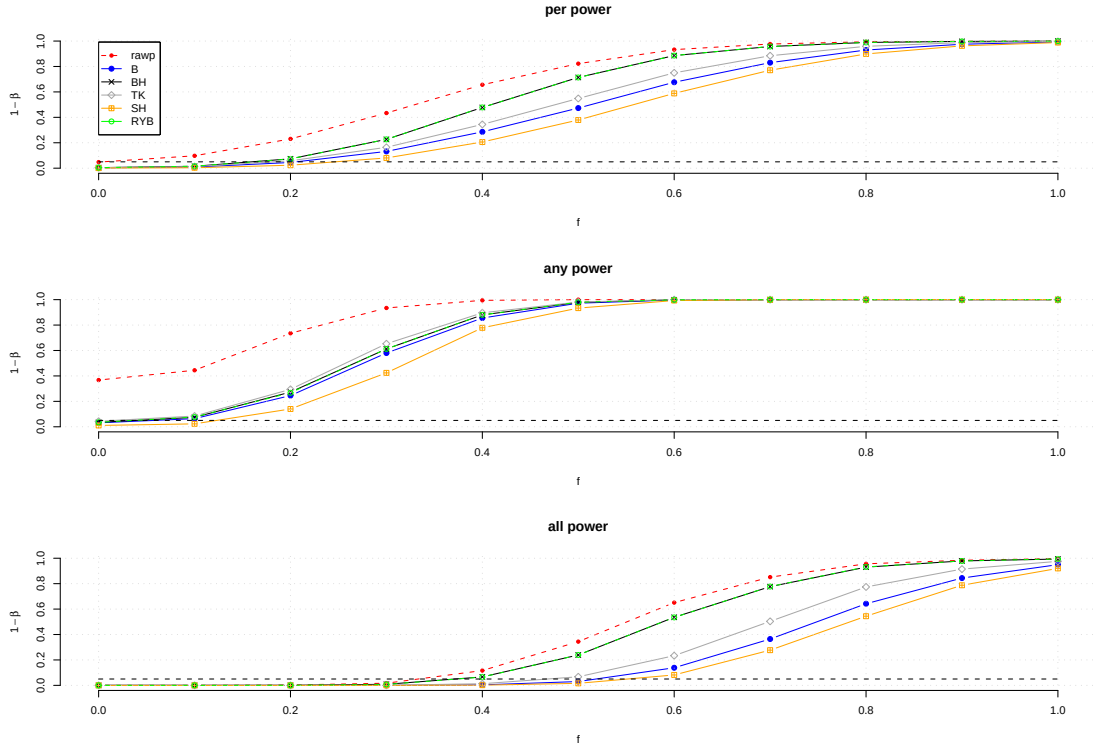

Figura 2. Curve di potenza stimate con $k = 6$ medie. Le curve sono ottenute come media tra le configurazioni di variazione massima e minima. La linea tratteggiata indica la soglia 0.05. In rosso le probabilità non aggiustate.

balmente più potenti rispetto ai metodi tradizionali considerati (SH, TK, e B) e tale differenza diviene importante al crescere del numero di medie confrontate.

Riferimenti bibliografici

BACHMANN, C., LUCCIO, R., SALVADORI, E. (2005). *La verifica della significatività dell'ipotesi nulla in psicologia.* Florence University Press, Firenze.

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.

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.

COHEN, J. (1988). *Statistical power analysis for the behavioral sciences*, II ed., Erlbaum, Hillsdale, NJ.

DAVID, H. A., LACHENBRUCH, P. A., BRANDIS, H. P. (1972). The power function of range and Studentized range test in normal samples. *Biometrika*, 59, 161-168.

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

PASTORE, M., NUCCI, M., GALFANO, G. (2005). False Discovery Rate: Applicazione di un metodo alternativo per i confronti multipli con misure ripetute. *Giornale Italiano di Psicologia*, 32, 639-652.

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

RAMSEY, P. H. (1978). Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association*, *363*, 479–485.

RAMSEY, P. H. (2002). Comparison of closed testing procedures for pairwise testing of means. *Psychological Methods*, 7, 504-523.

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

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

TUKEY, J. W. (1953). *The problem of multiple comparisons.* Princeton University Press, Princeton, NJ.

WESTFALL, P. H., & YOUNG, S. S. (1993). *Resampling-based multiple testing*, Wiley, NY.