L'EFFETTO DELLA NUMEROSITÀ SUL SIGNIFICATO DI UN RISULTATO STATISTICAMENTE SIGNIFICATIVO. UN COMMENTO SULL'ARTICOLO DI BENASSI *ET AL.*

GIANMARCO ALTOÈ[1] E MASSIMILIANO PASTORE[2]

[1] *Università di Cagliari,* [2] *Università di Padova*

Riassunto. L'interpretazione di un test statisticamente significativo può prescindere dalla numerosità campionaria sulla quale è stato condotto lo studio? In questo breve commento all'articolo di Benassi *et al.* (2013) cercheremo di rispondere a tale domanda utilizzando lo stesso caso di studio e lo stesso metodo (la *False Positive Report Probability*) proposti dagli autori. Dimostreremo che, a differenza di quanto sostenuto da Benassi *et al.*, l'interpretazione della significatività statistica non può prescindere dalla numerosità campionaria. I risultati ottenuti saranno discussi ponendo particolare enfasi sulla loro rilevanza applicativa.

1. INTRODUZIONE

Supponiamo di voler valutare la relazione tra due variabili quantitative Y e X e di avere a disposizione due studi, il primo condotto su un campione a bassa numerosità campionaria e il secondo su un campione a elevata numerosità campionaria. Supponiamo inoltre che le analisi statistiche condotte indichino la presenza di un risultato statisticamente significativo in entrambi gli studi. A quale risultato dobbiamo credere maggiormente? Al primo, al secondo, oppure i risultati ottenuti sono ugualmente validi? Sulla base delle considerazioni di Benassi, Casadio e Bolzani, i risultati sono ugualmente validi. In termini più generali, gli autori concludono che un risultato statisticamente significativo è ugualmente affidabile, indipendentemente dalla numerosità campionaria su cui è stato condotto uno studio.

L'obiettivo di questo breve commento è dimostrare che, contrariamente a quanto sostenuto da Benassi *et al.*, l'interpretazione della significatività statistica non può prescindere dalla numerosità campionaria. In particolare: 1) richiameremo la logica del *Null Hypothesis Significance Testing* (*NHST*, vedi Pastore, 2009) per la verifica di ipotesi, il concetto di significatività statistica che ne deriva, e alcuni aspetti critici legati al ruolo della numerosità campionaria; 2) presenteremo sinteticamente il criterio della *False Positive Report Probability*, ossia la probabilità che un test riporti un risultato falsamente po-

sitivo (*FPRP*; Wackholder, Chanok, Garcia-Closas, El ghormli e Rothman, 2004). Tale criterio è stato utilizzato da Benassi *et al.* per valutare la validità dei risultati statisticamente significativi; 3) partendo dal caso proposto dagli autori (valutazione della correlazione tra due variabili)[1] utilizzeremo il criterio *FPRP* per studiare il ruolo della numerosità campionaria nell'interpretazione di un risultato significativo; 4) illustreremo i risultati ottenuti confrontandoli con quelli di Benassi *et al.*, mettendone in evidenza la rilevanza a livello applicativo.

2. L'APPROCCIO NHST E IL CONCETTO DI SIGNIFICATIVITÀ STATISTICA: ALCUNI PUNTI CRITICI A LIVELLO INTERPRETATIVO

Nel loro articolo, Benassi *et al.* hanno considerato l'approccio *NHST* per la verifica di ipotesi. Tale approccio è il più utilizzato nella ricerca in psicologia ed è proprio al suo interno che viene definito il concetto di significatività statistica. In breve, l'approccio *NHST* può essere descritto come segue.

Il primo passo consiste nel definire un'ipotesi H_0 (detta ipotesi nulla) su un parametro incognito della popolazione (ad esempio il coefficiente di correlazione tra due variabili). In seguito viene calcolata una opportuna funzione dei dati campionari (x), detta statistica test $t(x)$, di cui è nota la distribuzione di densità di probabilità se vale l'ipotesi nulla $f(t(x)|H_0)$. Sulla base di queste assunzioni è possibile calcolare la probabilità di ottenere un risultato uguale o più estremo di quello osservato se vale H_0. Tale probabilità, solitamente chiamata *p-value*, è considerata una misura dell'evidenza contro l'ipotesi nulla. Minore è il *p-value* e maggiore sarà la forza dell'evidenza contro H_0. Nell'approccio *NHST* viene definito a priori un livello di significatività critico α; se *il p-value* è inferiore a tale soglia (in genere .05), l'ipotesi nulla viene rifiutata e il risultato del test viene considerato statisticamente significativo[2].

Dal punto di vista teorico, Cornfield (1966) propose il cosiddetto «Postulato-α» in riferimento all'effetto della numerosità campionaria. In sintesi, tale postulato afferma che un risultato statisticamente

[1] Nel lavoro originale gli autori hanno considerato il caso della regressione semplice. Senza perdita di generalità e per maggiore semplicità a livello interpretativo, in questo commento ci focalizzeremo sul caso del coefficiente di correlazione tra due variabili. È opportuno sottolineare che, dal punto di vista statistico e considerando lo scopo di questo *commentary*, i due casi possono essere considerati del tutto equivalenti.

[2] Per evitare ambiguità terminologiche, in questo commento utilizzeremo sempre il termine *p-value* con riferimento alla probabilità osservata, sulla base dei dati, di ottenere un risultato uguale o più estremo a quello osservato se vale H_0, e con il termine α il livello di significatività critico fissato a priori.

significativo esprime la stessa evidenza contro l'ipotesi nulla indipendentemente dalla numerosità campionaria su cui è stato condotto uno studio. Il Postulato-α, la cui validità è stata rifiutata dallo stesso Cornfield (1966) in favore dei test basati sul rapporto tra verosimiglianze, è stato implicitamente adottato da Benassi e collaboratori. Nel loro lavoro, gli autori infatti concludono che un risultato statisticamente significativo in campioni piccoli ha la stessa valenza di un risultato significativo in campioni grandi.

Attualmente la validità del Postulato-α è messa in discussione dalla maggior parte degli studiosi, e sembra ormai accertato che l'interpretazione di un risultato significativo non possa prescindere dal considerare la numerosità campionaria (Royall, 1986). In particolare, si sono storicamente imposte due posizioni apparentemente contraddittorie. Secondo alcuni autori (ad esempio Lindley e Scott, 1984) la forza dell'evidenza contro l'ipotesi nulla, sulla base di un risultato statisticamente significativo, è maggiore in campioni a bassa numerosità rispetto a campioni ad elevata numerosità, mentre secondo altri autori (ad esempio Peto *et al.*, 1976) un risultato statisticamente significativo indica una maggiore evidenza contro l'ipotesi nulla in campioni a elevata numerosità. In questo lavoro mostreremo, utilizzando come criterio la *FPRP*, che con un'adeguata attenzione a questioni sia teoriche che puramente terminologiche entrambe le posizioni sopra esposte possono essere ritenute corrette. Naturalmente, dimostrare la correttezza di entrambe le posizioni equivale a dimostrare l'invalidità del Postulato-α e quindi a sostenere che la plausibilità di un risultato significativo è legata alla numerosità campionaria.

3. LA FALSE POSITIVE REPORT PROBABILITY (FPRP)

La *False Positive Report Probability* (Wackholder *et al.*, 2004) è definita come la probabilità che l'ipotesi nulla sia vera dato un risultato statisticamente significativo, $Pr(H_0|risultato\ significativo)$. La formula per il calcolo della FPRP è:

$$FPRP = \frac{\pi_0 \alpha}{\pi_0 \alpha + (1 - \pi_0)(1 - \beta)} \tag{1}$$

dove π_0 e $(1 - \pi_0)$ sono le probabilità a priori che l'ipotesi nulla (H_0) e l'ipotesi alternativa (H_1) siano vere; α è la probabilità dell'errore di primo tipo (cioè rifiutare H_0 quando è vera) o, in altri termini, la probabilità che il test produca un falso positivo; $(1 - \beta)$ è la probabilità,

valutata ad un livello di significatività α, di rifiutare H_0 quando essa è falsa (la potenza del test), ossia la probabilità che il test produca un risultato correttamente positivo. Come si può notare, il calcolo della *FPRP* prevede di considerare sia l'ipotesi nulla che l'ipotesi alternativa. Quest'ultima viene infatti coinvolta nel calcolo della potenza del test.

Supponiamo ora, come proposto da Benassi *et al.*, di non avere informazioni a priori sulla plausibilità delle due ipotesi H_0 e H_1 o, equivalentemente, di non voler favorire nel calcolo di *FPRP* nessuna delle due ipotesi. In questo caso, poiché $\pi_0 = (1 - \pi_0) = .5$, la formula per la *FPRP* diventa:

$$FPRP = \frac{\alpha}{\alpha + (1 - \beta)} \tag{2}$$

Osservando tale formula è facile intuire la logica sottostante alla *FPRP*. Essa può essere vista come la proporzione di risultati falsamente significativi di un test sul totale dei risultati significativi prodotti. Tanto minore sarà tale proporzione, tanto maggiore sarà la correttezza e quindi la validità dei risultati significativi prodotti.

Nel caso specifico di un test basato sul NHST, in cui α viene fissato a priori (ad esempio .05), appare evidente che la FPRP varia al variare dalla potenza del test. In particolare, fissata una dimensione dell'effetto sotto l'ipotesi alternativa, è naturale aspettarsi che all'aumentare della numerosità campionaria (n) aumenti la potenza del test, e che quindi diminuisca la FPRP. La conclusione a cui sono giunti gli autori sembra invece suggerire che la probabilità che l'ipotesi nulla sia vera dato un risultato statisticamente significativo rimanga costante indipendentemente da n.

Nel prossimo paragrafo presenteremo due applicazioni (basate sul medesimo impianto sperimentale di Benassi *et al.*) per cercare di chiarire l'effetto della numerosità campionaria sulla validità di un risultato significativo.

4. L'INFLUENZA DELLA NUMEROSITÀ CAMPIONARIA VALUTATA ATTRAVERSO L'FPRP

In tabella 1 è presentata la *FPRP* calcolata attraverso la (2) nel caso del test bidirezionale sul coefficiente di correlazione tra due variabili. Per coerenza con gli autori, abbiamo considerato un effetto di correlazione sotto H_1 di media dimensione ($\rho = .30$; vedi Cohen, 1992), un α critico pari a .05 e le seguenti numerosità campionarie: 10, 20, 40, 80, 160. Ad esempio, nel primo caso ($n = 10$) basterà calcolare la potenza del test per un valore del coefficiente di correlazione pari a .30

TAB. 1. *Valori di FPRP nel caso del test sul coefficiente di correlazione al variare della numerosità campionaria fissati* $\alpha = .05$ *e* $\rho = .30$

n	10	20	40	80	160
FPRP	.300	.172	.097	.060	.049

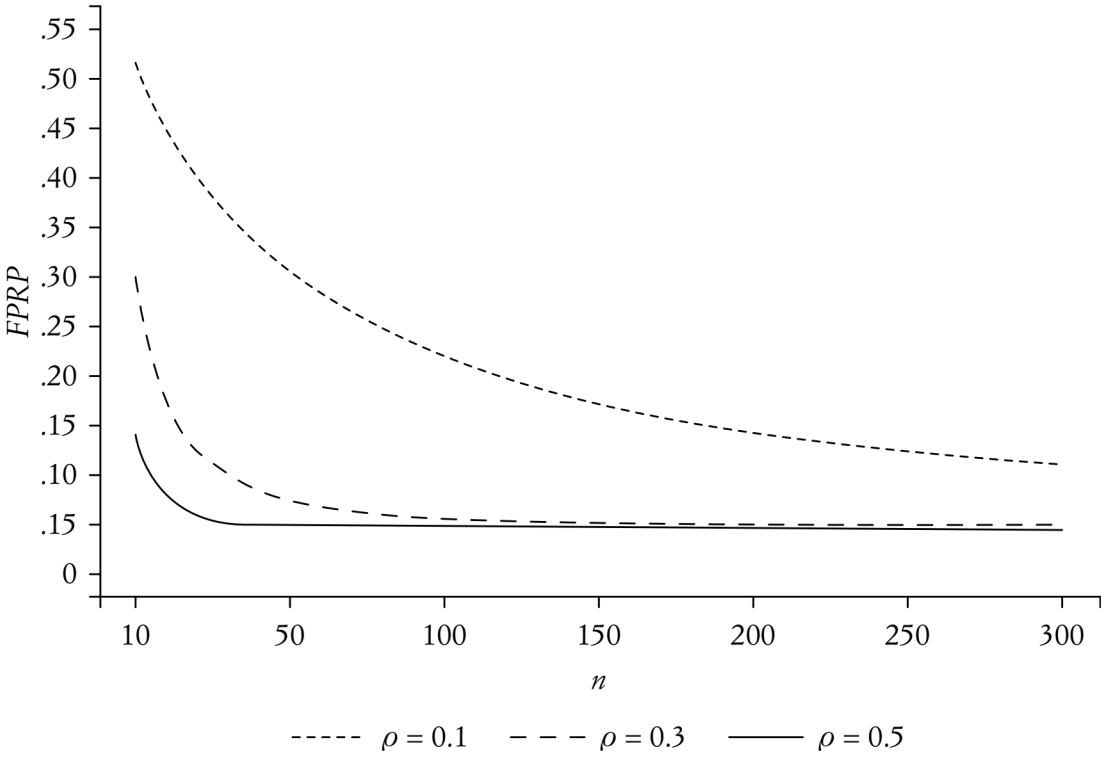

FIG. 1. Probabilità di riportare risultati falsamente positivi (FPRP) al variare della numerosità campionaria e della dimensione dell'effetto sotto H_1.

ed un livello critico di α pari a .05. Tale valore pari a .117, sostituito nella (2) al posto di $(1 - \beta)$ e con α pari a .05, produrrà un valore di *FPRP* = .300.

Come si può notare, la proporzione di falsi positivi (*FPRP*) decresce in maniera monotona all'aumentare della numerosità[3]. Si passa da una proporzione del 30% di falsi risultati significativi per $n = 10$ ad una proporzione di circa il 5% per $n = 160$.

Nella figura 1 abbiamo rappresentato l'andamento della *FPRP* al variare della numerosità campionaria per *n* compreso tra 10 e 300 e

[3] I risultati esposti in tabella 2 sono coerenti con quelli riportati dagli autori in tabella 6. Le differenze nei singoli valori possono essere dovute al fatto che, mentre noi abbiamo utilizzato la formula analitica per il calcolo del FPRP, gli autori hanno utilizzato una simulazione Monte Carlo. I risultati di entrambi gli approcci indicano chiaramente che all'aumentare della numerosità diminuisce la FPRP. Tuttavia, questa conclusione non viene riportata da Benassi *et al.*

considerando tre dimensioni di effetto (piccolo $\rho = .1$, medio $\rho = .3$ e grande $\rho = .5$, vedi Cohen, 1992) sotto H_1.

Dal grafico appare evidente che, all'aumentare della numerosità campionaria e fissata una dimensione dell'effetto sotto H_1, la probabilità di riportare risultati falsamente positivi diminuisce. Sulla base di questi risultati si può pertanto concludere che un risultato significativo è maggiormente affidabile, in termini di evidenza contro l'ipotesi nulla, se proviene da un campione ad elevata numerosità campionaria rispetto a un campione a ridotta numerosità campionaria. Tale conclusione è perfettamente coerente con quanto sostenuto da Peto *et al.* (1976).

Abbandoniamo ora per un attimo la logica *NHST*, che prevede la scelta dicotomica «risultato significativo *vs.* non significativo» sulla base di un α critico fissato a priori, e soffermiamoci sul *p-value* osservato nei dati a disposizione. Chiediamoci ora se, a parità di *p-value*, la forza dell'evidenza contro l'ipotesi nulla è maggiore in un campione piccolo rispetto ad un campione grande.

In questo caso la FPRP può essere vista come la probabilità che H_0 sia vera sulla base del *p-value*: $Pr(H_0|p)$. Ipotizzando un'uguale probabilità a priori per l'ipotesi nulla e l'ipotesi alternativa, la formula per il calcolo della FPRP diventa (Wakefield, 2007, p. 211):

$$FPRP = \frac{p}{p + (1 - \beta_{oss})} \tag{3}$$

dove p è il *p-value* osservato sulla base dei dati a disposizione e $(1 - \beta_{oss})$ è la potenza osservata del test per un livello di significatività pari a p. In particolare, fissato un valore di p e fissata una numerosità campionaria, è facile dimostrare che la dimensione dell'effetto sotto H_1 è immediatamente determinata. Per chiarire le idee, supponiamo di aver rilevato un *p-value* pari a .04 in un campione a numerosità 10. In questo caso, partendo dalla formula per la verifica di ipotesi sul coefficiente di correlazione, si ottiene dopo alcuni semplici passaggi matematici che il valore del coefficiente di correlazione osservato è pari a .655 e, conseguentemente, che la potenza del test calcolata per un livello di significatività di .04 è pari a .500[4]. Combinando assieme il *p-value* e la potenza osservata del test nella formula (3) si ottiene un valore di *FPRP* pari a .074. In altri termini, data una numerosità pari

[4] Si veda, per i dettagli tecnici, un qualsiasi manuale di psicometria o statistica.

TAB. 2. *Dimensione dell'effetto sotto H_1, potenza osservata e FPRP nel caso di verifica di ipotesi sul coefficiente di correlazione per n = 10 e n = 160 per un valore di p pari a .04.*

n	*p*	$\rho \mid H_1$	$(1 - \beta_{oss})$	*FPRP*
10	.04	.655	.500	.074
160	.04	.163	.500	.074

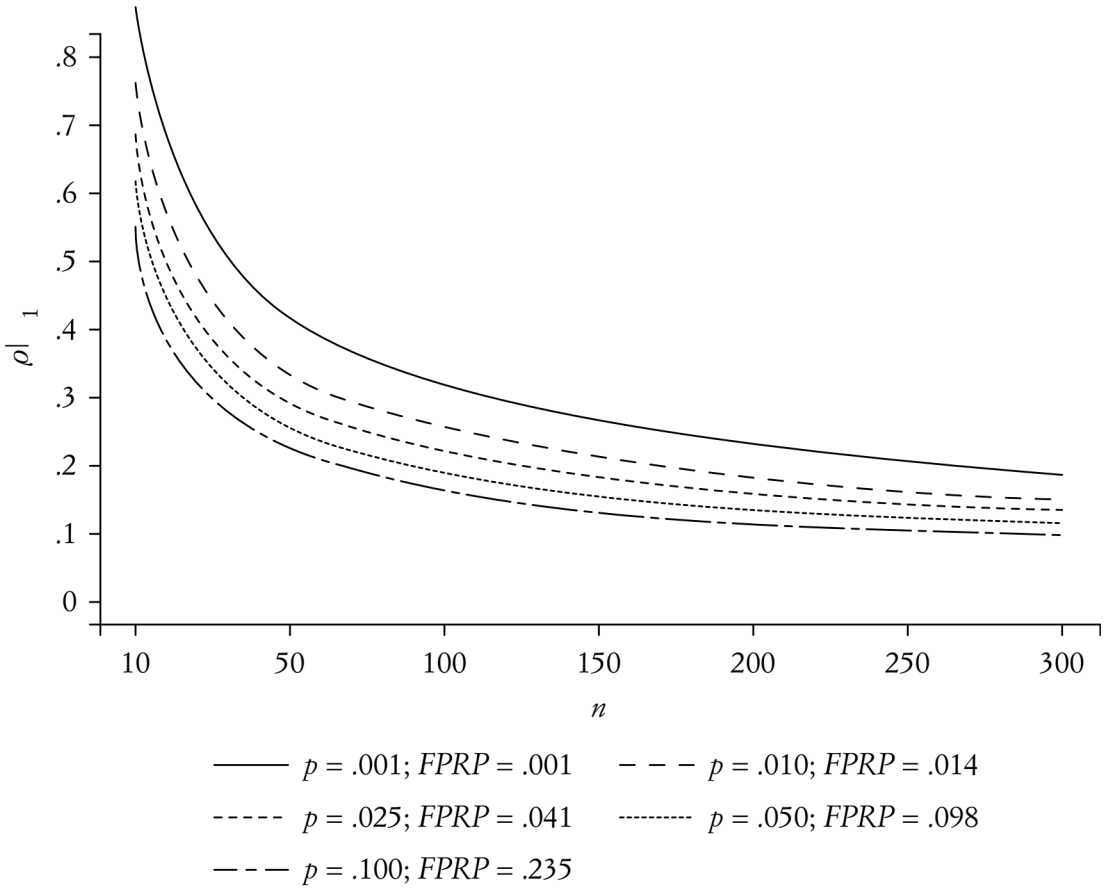

FIG. 2. Dimensione dell'effetto sotto H_1 al variare della numerosità campionaria e di diversi valori di *p-value.*

a 10 ed avendo osservato un *p-value* pari a .04, la probabilità che l'ipotesi nulla sia vera è del 7.4%.

Nella tabella 2 sono riportati tutti gli elementi che permettono di calcolare la *FPRP* (e la *FPRP* stessa) nel caso del test sul coefficiente di correlazione per due numerosità scelte a titolo esemplificativo ($n = 10$ e $n = 160$).

Come si può osservare, in entrambi i casi ($n = 10$ e $n = 160$) la *FPRP* è costante, ma la dimensione dell'effetto sotto H_1 è superiore per il campione a numerosità inferiore. A parità di *p-value*, il campione più piccolo sembra quindi fornire una maggiore evidenza contro l'ipotesi nulla.

Nella figura 2 viene rappresentato l'andamento della dimensione dell'effetto sotto H_1 al variare della numerosità campionaria, per n compreso tra 10 e 300 e considerando i seguenti valori di *p-value*: .001, .01, .025, .05, .10.

Osservando i risultati in figura 2 si può notare che, fissato un *p-value*, la probabilità di riportare risultati falsamente positivi non varia al variare della numerosità. Quello che cambia è però la dimensione dell'effetto sotto H_1, che diminuisce all'aumentare della numerosità. Fissato ad esempio un *p-value* pari a .05, la dimensione dell'effetto testata sotto H_1 risulta essere $\rho = .63$ (effetto grande) per $n = 10$ e $\rho = .11$ (effetto piccolo) per $n = 300$. In conclusione, coerentemente con quanto sostenuto da Lindley e Scott (1984) e Wagenmakers (2007), a parità di significatività osservata (ovvero di *p-value*), campioni a ridotta numerosità sembrano fornire maggiore evidenza contro l'ipotesi nulla rispetto a campioni ad elevata numerosità.

5. CONCLUSIONI

I risultati ottenuti dimostrano complessivamente che l'interpretazione di un risultato statisticamente significativo, e quindi della forza dell'evidenza contro l'ipotesi nulla, dipende dalla numerosità campionaria. In particolare, se si utilizza un criterio di decisione basato su un livello di significatività critico fissato a priori (α), la forza dell'evidenza contro l'ipotesi nulla è maggiore in campioni ad elevata numerosità campionaria, mentre se ci si focalizza sul *p-value* osservato sulla base dei dati rilevati, la forza dell'evidenza contro l'ipotesi nulla è maggiore in campioni a ridotta numerosità campionaria. In nessuno di questi casi tuttavia si può affermare che la forza dell'evidenza contro l'ipotesi nulla è indipendente dalla numerosità campionaria.

Il fatto che l'interpretazione dei risultati statisticamente significativi dipenda dalla numerosità campionaria può essere visto come un punto debole del *NHST*. In particolare, a livello applicativo è importante che il ricercatore tenga conto del ruolo della numerosità campionaria nel presentare e commentare i risultati dei test di significatività. Ad esempio, nel caso della valutazione di un trattamento psicologico va considerato che, a parità di *p-value*, un campione a bassa numerosità evidenzia una maggiore dimensione dell'effetto studiato rispetto a un campione ad elevata numerosità. In accordo quindi con quanto proposto dalle recenti linee guida dell'*American Psychological Association* (2010), i risultati di un test di significatività devono essere accompagnati da ulteriori indicazioni (ad esempio la dimensione dell'effetto) in modo da fornire un'informazione maggiormente completa e corretta, e limitando al minimo possibili interpretazioni errate.

BIBLIOGRAFIA

American Psychological Association (2010). *Publication manual of the American Psychological Association (6th ed.)*. Washington, D.C.: APA.

Benassi M., Casadio M., Bolzani R. (2013). Validità statistica dei risultati in esperimenti a bassa numerosità campionaria. Simulazione di un test parametrico. *Giornale Italiano di Psicologia*, *40*, 353-366.

Cohen J. (1992) A power primer. *Psychological Bulletin*, *112*, 155-159.

Cornfield J. (1966). Sequential trials, sequential analysis and the likelihood principle. *The American Statistician*, *20*, 18-23.

Lindley D.V., Scott W.F. (1984). *New Cambridge Statistical Tables*. Cambridge: Cambridge University Press.

Pastore M. (2009). I limiti dell'approccio NHST e l'alternativa Bayesiana. *Giornale Italiano di Psicologia*, *36*, 925-938.

Peto R., Pike M.C., Armitage P., Breslow N.E., Cox D.R., Howard S. V., Mantel N., Mcpherson K., Peto J., Smith P.G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient, I: Introduction and design. *British Journal of Cancer*, *34*, 585-612.

Royall R.M. (1986). The effect of sample size on the meaning of significance tests. *The American Statistician*, *40*, 313-315.

Wacholder S., Chanok S., Garcia-Closas M., El Ghormli L., Rothman N. (2004) Accessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, *96*, 434-442.

Wagenmakers E.J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779-804.

Wakefield J. (2007). A bayesian measure of the probability of false discovery in genetic epidemiology studies. *American Journal of Human Genetics*, *81*, 208-227.

The effect of sample size on the meaning of a statistically significant result

Summary. Can a statistically significant test be interpreted regardless of the sample size used in a particular study? In this brief commentary on Benassi *et al.* (2013), we seek to answer this question using the same case study and method (i.e., *False Positive Report Probability*) proposed by the authors. We will demonstrate that, differently from Benassi et al., the interpretation of statistical significance is strongly related to sample size. The results are discussed with a special emphasis on their applied relevance.

Keywords: Null Hypothesis Significance Testing, p-value, critical α level, False Positive Report Probability, sample size.

La corrispondenza va inviata a Gianmarco Altoè, Dipartimento di Pedagogia, Psicologia e Filosofia, Università di Cagliari, Via Is Mirronis 1, 0123 Cagliari. E-mail: gianma.alto@gmail.com