Vita di P: 16 anni di statistiche sul GIP*

Massimiliano Pastore[1], Massimo Nucci[2] e Andrea Bobbio[3]

[1]Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università di Padova.
[2]Dipartimento di Psicologia Generale, Università di Padova.
[3]Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata, Università di Padova.

Sommario In questo lavoro abbiamo considerato la diffusione e l'utilizzo della statistica inferenziale nel Giornale Italiano di Psicologia. Per mezzo di una procedura automatizzata sono stati selezionati tutti gli articoli pubblicati tra il 1997 ed il 2012 che presentavano risultati statistici espressi attraverso almeno un *p-value*. Le statistiche più utilizzate sono risultate essere F, t e χ^2 (circa il 70% del totale), e la loro frequenza d'uso è rimasta costante nel periodo di tempo esaminato. Ove è stato possibile, abbiamo ricalcolato i valori di p confrontandoli con quelli riportati dagli autori ed analizzato la distribuzione complessiva. Infine, adottando una procedura bayesiana, abbiamo esaminato le stime della dimensione degli effetti.

Parole chiave: *p-value*, *effect size*, norme APA, *Null Hypothesis Significance Testing*

Life of P: 16 years of statistics on the Italian Journal of Psychology.

Summary In this paper we studied the use of inferential statistics in the Italian Journal of Psychology. By means of an automated procedure, all articles published from 1997 to 2012 that presented statistical results along with at least one *p-value* were selected. The most commonly used statistics turned out to be F, t and χ^2 (about 70% of the total), and their frequency of use remained stable throughout the period here considered. Where possible, we recalculated the *p-values*, compared them with those reported by the authors, and analyzed the overall distribution. Finally, applying a Bayesian procedure, the effect size estimates were assessed.

La corrispondenza va inviata a Massimiliano Pastore, c/o Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Via Venezia, 8 I-35131 Padova (PD), Italy.
Email: massimiliano.pastore@unipd.it

* Giornale Italiano di Psicologia / a. XLII, 2015.

1 INTRODUZIONE

Qualità, efficacia, credibilità della ricerca sono argomenti di grande attualità in tutte le discipline scientifiche. Nella psicologia il dibattito legato alla serietà e al valore degli studi pubblicati è presente sin dalla fine degli anni Ottanta, grazie ad un famoso articolo di Jacob Cohen (*The Earth is round ($p < .05$);* 1988). Il principale tema di discussione è il paradigma NHST (*Null Hypothesis Significance Testing*), di fatto pressoché l'unica modalità di analizzare i dati nella ricerca in psicologia (si veda anche, ad esempio: Cohen, 1994; Kline, 2004; Wagenmakers, 2007; Johnson, 2013), nonché il riferimento per tutti o quasi gli insegnamenti di contenuto statistico nei corsi di laurea in Psicologia. L'attenzione è stata posta sui problemi legati alle errate interpretazioni di questo approccio (Bakker & Wicherts, 2011; Gigerenzer & Marewski, 2015; Ziliak & McCloskey, 2008), sull'influenza delle cosiddette *Questionable Research Practices* (John, Loewenstein, & Prelec, 2012), sulla credibilità (Shea, 2011), la replicabilità e la consistenza dei risultati pubblicati (Schimmack, 2012; Simmons, Nelson, & Simonsohn, 2011; Francis, 2013). Di recente, anche in Italia sono stati affrontati alcuni temi legati ai problemi dell'approccio NHST e alle possibili alternative adottabili (si veda ad es. Agnoli & Furlan, 2009; Balboni & Cubelli, 2009; Di Nuovo, 2009; Pastore, 2009; Altoé & Pastore, 2013; Perugini, 2014b).

Per un prodotto scientifico che includa analisi statistiche, le linee guida dell' *American Psychological Association*, a partire dall'edizione del 1994, indicano la necessità di riportare le seguenti informazioni: valore della statistica test, gradi di libertà, probabilità associata (*p-value*), incoraggiando inoltre a tenere in considerazione la potenza del test ('*take seriously the statistical power considerations associated with your tests of hypotheses*', p. 16) e la dimensione dell'effetto ('[*you are*] *encouraged to provide effect-size information*', p. 18). Nella quinta edizione del manuale APA (2001) si davano le seguenti ulteriori indicazioni: nel caso di valore statisticamente significativo, riportare la soglia, ovvero $p < .05$, $p < .01$ oppure $p < .001$, mentre nel caso di valore non significativo riportare la probabilità esatta (es. $p = .12$). Evidentemente consapevoli dei problemi legati alla significatività statistica, anche a seguito del lavoro dell'APA *Task Force on Statistical Inference* (1999), si introduceva la necessità di riportare la misura dell'*effect size* (anche nel caso di risultati non significativi) e si suggeriva l'uso combinato del *p-value* e dell'intervallo di confidenza. Ancora, nella sesta edizione (2010), a tutte le precedenti prescrizioni viene aggiunto il criterio di riportare sempre la probabilità esatta per valori di p non inferiori a .001 (es. $p = .006$) e di usare la soglia ($p < .001$) solo negli altri casi.

Nonostante queste indicazioni le principali debolezze del paradigma NHST rimangono presenti, ed anzi la constatazione della significatività statistica continua ad essere condizione spesso necessaria per la pubblicazione, comportando così una serie di distorsioni non solo nei risultati, ma anche nella modalità di fare ricerca (Maxwell, 2004; Young, Ioannidis, & Al-Ubaydi, 2008; Scargle, 2000; Schooler, 2011). A questo si aggiungano alcuni "errori" d'interpretazione dei risultati; tra tutti il più frequente e denunciato è l'assunzione del *p-value* come misura di evidenza statistica ed il suo uso per quantificare la forza dei risultati ottenuti (Cumming, 2008; Gelman, 2013a, 2013b;

Hubbard & Lindsay, 2008; Ioannidis, 2005; Johnson, 2013; Sterne & Smith, 2001; Wagenmakers, 2007).

In questo lavoro abbiamo considerato la diffusione, l'utilizzo e le modalità di comunicazione dei risultati ottenuti con la statistica inferenziale nella ricerca in psicologia condotta in Italia negli ultimi 16 anni. In particolare, abbiamo considerato le pubblicazioni del Giornale Italiano di Psicologia, in quanto rivista generalista che tradizionalmente presenta lavori facenti riferimento a tutte le anime della psicologia e dunque ragionevolmente rappresentativa del panorama italiano. L'obiettivo del lavoro non è dibattere sulla correttezza dei metodi impiegati e ancor meno esprimere una valutazione sulla bontà dei risultati presentati, questo per due ragioni: 1) nessuno può davvero credere di essere senza peccato e quindi ritenersi legittimato a scagliare pietre sul prossimo; 2) i contributi posti sotto esame appartengono a molti e diversi ambiti di ricerca: entrare nel merito di ognuno di essi presuppone competenze tali che chi scrive non sente di possedere. Piuttosto, ci siamo semplicemente proposti di scattare una fotografia di ciò che è stato fatto negli ultimi anni, compendiando in statistiche descrittive l'evoluzione e lo stato dell'arte, nella convinzione che ogni discussione ed ogni eventuale percorso di miglioramento debbano partire dalla conoscenza dei dati oggettivi. Per fare questo ci siamo concentrati dapprima sui valori di probabilità associati alle statistiche e successivamente, adottando una procedura di tipo bayesiano, abbiamo analizzato le dimensioni degli effetti. Tutte le analisi sono state effettuate con R (R Core Team, 2014).

2 METODO

Sono stati presi in considerazione tutti gli articoli pubblicati sul Giornale Italiano di Psicologia dal 1997 al 2012 per un totale di 820 lavori[2]. L'obiettivo iniziale è stato individuare tutte le occorrenze in cui si utilizzassero dei valori di probabilità per convalidare ipotesi statistiche, in altre parole individuare tutti i *p-value*. Gli articoli, originariamente in formato pdf, sono stati convertiti in file formato testo utilizzando il pacchetto `tm` (Feinerer, Hornik, & Meyer, 2008). Quindi, sono stati analizzati i file ricercando la stringa 'p' isolata nel testo oppure vicina ad operatori quali uguale ('='), maggiore ('>') o minore ('<')[3]; in altre parole abbiamo rintracciato, grazie ad un'apposita procedura automatizzata, tutti i contesti in cui la lettera fosse usata in modalità tale da indicare inequivocabilmente un *p-value*, ossia una probabilità associata ad un test statistico. Con questa procedura sono state identificate tutte le righe di testo degli articoli che ne contenevano almeno una e salvato tali righe in un *database* per le successive analisi. Come secondo passaggio abbiamo analizzato tutte le righe selezionate individuando i valori riportati di ciascuna p e, laddove fosse possibile, identificando in modo automatico la statistica test da cui tale probabilità era derivata ed altre informazioni ad essa connesse. In un terzo passaggio – non automatico – abbiamo verificato l'appropriatezza delle informazioni rilevate, isolato tutti casi in cui la gestione automatica si fosse rivelata difettosa o insufficiente, effettuando correzioni relative al valore di p, alla statistica test ed alle eventuali informazioni annesse, con particolare attenzione ai gradi di libertà.

[2] La scelta dell'intervallo 1997-2012 è legata alla disponibilità on-line degli articoli. Poiché il primo numero del GIP uscì nel 1974, si è passata in rassegna buona parte della metà più recente della sua storia.

[3] Abbiamo anche riscontrato la scrittura '$\leq$' oppure '$\geq$'; in entrambi i casi si è considerato l'operatore come minore o maggiore, senza tenere conto dell'uguale.

Complessivamente abbiamo identificato 342 articoli che contenevano almeno un valore di p, ovvero il 41.7% del totale dei lavori presi in considerazione. Laddove possibile, abbiamo ricalcolato il *p-value* così da poterlo studiare indipendentemente da quello riportato dagli autori. Quando sarà necessario distinguere, indicheremo pertanto con p_a i valori di probabilità riportati dagli autori e con p_r i valori di probabilità ricalcolati sulla base delle informazioni presenti negli articoli.

2.1 Corpus dei dati

Si sono esaminati 342 articoli (circa 21 per ogni anno) nei quali fosse riportato almeno un valore di p, rilevandone un totale di 4903. Il numero di p per articolo varia da un minimo di 1 ad un massimo di 144, con media pari a 14.34 (mediana 10, d.s. 15.2); di questi il 77.2% esprime solo un criterio di soglia (scritto con $p <$ oppure $p >$). In ogni caso abbiamo registrato la statistica test relativa a ciascuna probabilità e controllato se, insieme a quest'ultima, fossero presenti le informazioni per ricalcolare p (ad esempio i gradi di libertà, ove necessari). In questo modo è stato possibile calcolare le probabilità esatte associate alle statistiche test nel 60.2% di casi (2954 valori in tutto), ovvero nei casi in cui le statistiche test fossero le seguenti: χ^2, F, r, t, Z. I coefficienti di correlazione sono le statistiche che mancano più frequentemente delle informazioni necessarie per un'appropriata interpretazione: infatti, solo nel 13% circa dei casi sono presenti i gradi di libertà associati; viceversa, escludendo i valori Z (che non richiedono i gradi di libertà), le statistiche F sono quelle riportate nella maniera più appropriata (circa il 91% delle volte).

Per semplicità abbiamo scelto di classificare tutte le statistiche in base alla distribuzione utilizzata per il calcolo di p ovvero χ^2, F, t, Z. A queste categorie abbiamo poi aggiunto r (che meritava una categoria a parte anche se, per l'inferenza, utilizza la distribuzione t) ed una generica categoria 'altre' in cui collocare le statistiche poco frequenti o di cui non fosse chiara la natura.

2.2 Nuovo calcolo di p

Per disporre del quadro completo delle probabilità esatte associate alle statistiche rilevate le abbiamo calcolate (nel caso gli autori avessero riportato solo un criterio soglia) o ricalcolate (nel caso fosse già presente la probabilità esatta). È stata ritenuta *usabile* una statistica riportata in modo tale da consentire un nuovo calcolo di p chiaro ed univoco. Nel dettaglio, sono risultati usabili:

- I valori Z e con essi anche tutti i test che si basano sulla distribuzione normale (es. Wilcoxon, Mardia Kurtosis).
- Le classiche statistiche F, t, χ^2 se presentavano i gradi di libertà.
- I coefficienti di correlazione (r di Pearson o di Spearman) e i valori affini (quali ad esempio r^2), se presentavano i gradi di libertà o la numerosità campionaria di riferimento.
- Il numero o la proporzione di casi favorevoli utilizzati nel test binomiale, se era presente anche la numerosità campionaria totale.
- Gli Odd Ratio (OR) se era presente l'errore standard associato[4].

[4] Dati OR ed il suo errore standard, se(OR), è possibile trasformare l'Odd Ratio in un valore z con la formula $\Phi^{-1}\left(\frac{-\log(OR)}{se(OR)}^2\right)$ in cui Φ^{-1} è l'inversa della distribuzione normale cumulata (vedi Morris & Gardner, 1988).

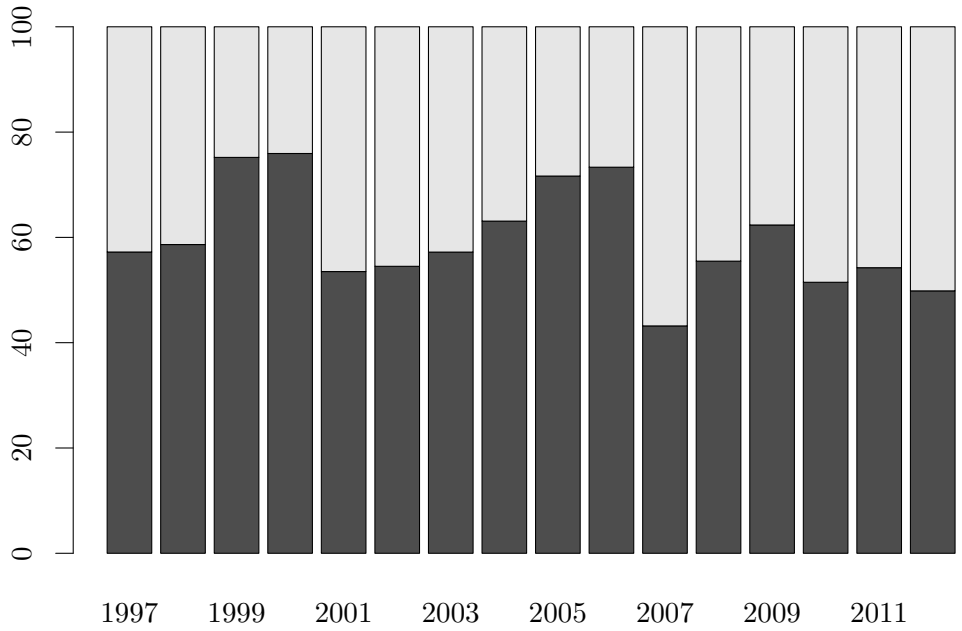

Figura 1. Frequenze percentuali per anno delle statistiche riportate negli articoli e classificate come usabili (porzione in nero) rispetto a quelle non usabili (porzione in grigio).

In relazione a t e Z abbiamo scelto di calcolare la probabilità assumendo che il test fosse bidirezionale, quindi più conservativo; controllando la corrispondenza con le p_a abbiamo trovato pochissimi casi discordanti da questa assunzione. Per quanto riguarda i valori di n riferiti alle correlazioni o al test binomiale o, più in generale, i gradi di libertà, li abbiamo registrati solo se presenti nel testo in prossimità della statistica. I casi in cui invece le statistiche non sono risultate usabili sono i seguenti:

- Tutti i coefficienti relativi a modelli di regressione semplice, o modelli di equazioni strutturali, in quanto mancanti quasi sempre dell'errore standard, o del relativo valore t, o ancora dei gradi di libertà.
- I test corretti con il metodo Bonferroni: negli unici casi in cui è presente un valore di una statistica mancano le specifiche ad essa associate.
- I test *post-hoc* (es. Duncan o Scheffé) poiché non abbiamo mai riscontrato i valori delle relative statistiche, ma solo i p.
- Più in generale, tutti i casi in cui non si sono riscontrate le informazioni sufficienti a calcolare la probabilità associata.

In figura 1 abbiamo rappresentato la percentuale di statistiche usabili (in nero) rispetto a quelle non usabili (in grigio) nei 16 anni considerati. La percentuale media di statistiche considerate *usabili* è del 59.8% (2954 statistiche su 4903), valore che sembra essere rimasto piuttosto costante nel corso del tempo.

2.3 Selezione dei *p*

In un secondo passaggio è stato estratto a caso un campione rappresentativo di articoli (pari a circa il 25% del totale). Per la selezione abbiamo proceduto come segue: 1) calcolato il numero di p per ciascun articolo e suddiviso tale variabile nelle seguenti 4 classi di frequenza: $(0, 5], (5, 10], (10, 20], (20, 200]$[5]; 2) calcolato la proporzione w_{ij} $(i = 1, \ldots, 16;\ j = 1, \ldots, 4)$ di lavori per anno nelle 4 classi; 3) estratto casualmente gli articoli usando w_{ij} come probabilità di estrazione associata a ciascuno di essi. Come risultato abbiamo ottenuto un campione stratificato per anno e per numero di p presentati la cui distribuzione congiunta rispetto a queste due variabili non risulta difforme da quella osservata empiricamente.

Gli 86 articoli così selezionati sono stati letti con l'obiettivo di individuare e classificare i p in relazione alla loro rilevanza, adottando uno schema simile a quello proposto da Simonsohn, Nelson, e Simmons (2014), in modo da poter compiere analisi più specifiche. In questo modo, sul totale di 1022 valori di p, abbiamo escluso i casi (233) non rilevanti, ovvero riferiti a statistiche non usabili o a risultati di letteratura. Nei restanti 832 casi abbiamo distinto tra valori direttamente associati alle ipotesi (circa il 38%), considerati rilevanti, e quelli relativi ad analisi di contorno, ad esempio di controllo, descrittive o esplorative. Quando faremo riferimento ai p rilevanti selezionati li indicheremo come p_s, in modo da distinguerli dai valori p_a (indicati dagli autori) e p_r (ricalcolati per il presente lavoro).

Seguendo il suggerimento di Gigerenzer e Marewski (2015) è stato possibile calcolare il rapporto: $f = n_t/n_r$, in cui n_t indica il numero totale di p riportati e n_r il numero di p rilevanti. Questo indice vale 1 quando vi è perfetta corrispondenza ed aumenta all'aumentare del numero di test condotti. Tolti 8 articoli di tipo prettamente esplorativo e quindi che non presentavano ipotesi specifiche, sui rimanenti 78 abbiamo un f medio di 2.6 (d.s. 1.83) con valori compresi tra 1 e 10.8.

3 RISULTATI

Una prima esplorazione per parole chiave sul totale degli 820 articoli pubblicati, ha avuto come obiettivo l'individuazione delle tecniche più citate indipendentemente dall'uso effettivo del *p-value*. In particolare, abbiamo identificato alcune macrocategorie all'interno delle quali riportare insiemi affini di parole chiave. Ad esempio la categoria `ANOVA` conteneva parole quali: analisi della varianza, Anova, Ancova, Mancova, Manova, t-test, z-test, test t. Le maggiori ricorrenze riguardano le categorie `misure di associazione` (con termini quali Correlazione, Cronbach, etc.) e `ANOVA` (entrambe trovate in più di 200 articoli), seguite dalle categorie `regressioni`, `test non parametrici`, `confronti multipli` e `analisi multivariate` (in circa 100 articoli). Sebbene questa prima ispezione sia interpretabile solo a livello generale-descrittivo, poiché il fatto che una parola chiave ricorra in un testo non implica necessariamente che la relativa tecnica statistica sia stata effettivamente utilizzata, appare piuttosto evidente che non compaiano riferimenti ad approcci alternativi a NHST. In particolare, parole chiave legate all'approccio bayesiano (Bayes factor, Bayesian Information Criterion, BIC, Bayesian) o all'uso di Effect size (d di Cohen, effect

[5] La parentesi tonda indica che il valore non è contenuto nell'intervallo pertanto $(0, 5]$ si intende un numero di p da 1 a 5 (compreso) e così via.

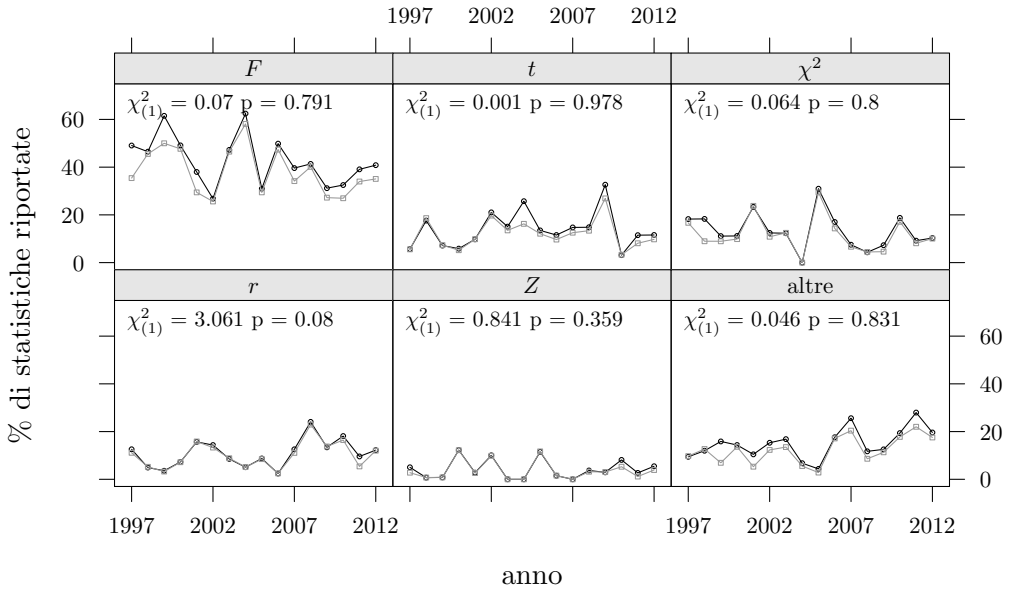

Figura 2. Frequenze percentuali delle statistiche per anno. Le linee nere sono riferite al numero complessivo, le linee grigie alle statistiche significative. I valori di χ^2 sono relativi ai Ljung-Box test condotti su ciascuna serie.

size, dimensione dell'effetto, Cohen's d, eta quadro[6]) sono state riscontrate solo in pochi articoli e per la maggior parte si tratta di citazioni bibliografiche.

3.1 Statistiche utilizzate

La statistica più frequente in assoluto è la F, presente in 228 lavori (pari al 66.7% dei 342 lavori considerati) e con 2005 valori riportati. Se consideriamo anche i 120 lavori (35.1%) in cui è stata usata la t (642 valori riportati), che è praticamente equivalente a F, si osserva che tali statistiche compaiono nel 76% di contributi pubblicati. Al terzo posto si trova il χ^2, presente in 101 lavori (29.5%), quindi, a seguire, i vari coefficienti di correlazione (presenti nel 27.2% di articoli) e Z (8.8%). La categoria 'altre' rappresenta complessivamente il 15% circa di casi.

In figura 2 sono rappresentate, in percentuale, le statistiche riportate per anno nelle 6 categorie considerate: il grado di impiego delle varie statistiche sembra essere costante nel tempo. Grazie al test Ljung-Box (Ljung & Box, 1978) abbiamo valutato se fosse possibile individuare la presenza di un trend. Nessuno dei 6 test, i cui risultati sono nei pannelli della figura 2, ha dato esito significativo. Confrontando linee nere e linee grigie, risulta evidente che l'88.2% dei risultati è statisticamente significativo, sostenendo la tendenza, già messa più volte in evidenza a livello internazionale, di omettere o trascurare i risultati nulli (si veda ad esempio: Fanelli, 2012; Francis, 2013; John et al., 2012; Simmons et al., 2011) puntando quasi esclusivamente su quelli significativi. Certamente è noto che un risultato non significativo non è indizio di validità dell'ipotesi nulla; tuttavia, per i detrattori della logica NHST è proprio questa dissimetria tra H_0 ed H_1 ad essere

[6] La ricerca di queste parole chiave è complicata dal fatto che le lettere greche (es. η^2, ω^2), nella conversione da file pdf a file testo, spesso sono rese non riconoscibili.

Operatore riportato	Valore di p riportato	Valore di p ricalcolato $\leq .05$	$> .05$	Totale
$<$	$\leq .05$	2036	59	2095
	$> .05$	2	20	22
$=$	$\leq .05$	432	18	450
	$> .05$	7	250	257
$>$	$\leq .05$	5	55	60
	$> .05$	1	19	20
	Totale	2483	421	2904

Tabella 1. Frequenza delle statistiche in funzione dell'operatore e valore riportato dagli autori (in riga) rispetto ai valori ricalcolati (in colonna). La non coincidenza con il numero di statistiche definite usabili (ovvero 2954) è dovuta al fatto che vi sono 47 casi in cui manca l'operatore e 3 casi in cui, pur essendo calcolabile, p non è stato riportato dagli autori.

causa di distorsioni nella ricerca scentifica, sia riguardo le ipotesi da verificare (le ricerche più ambiziose ma più azzardate sono trascurate) sia riguardo la ripetibilità dei risultati (aumentando la probabilità di pubblicare errori di I tipo).

3.2 Uso dell'operatore

Un primo aspetto cruciale nella descrizione del *p-value* riguarda l'operatore. Solo in 134 casi le probabilità riportate non erano associate ad un operatore: si tratta di risultati non significativi per cui gli autori al posto del valore o della soglia di p indicavano la sigla 'ns'. Nella maggioranza dei casi (4769) i p erano accompagnati da un operatore (minore, uguale oppure maggiore); si è allora cercato di analizzare l'uso che ne veniva fatto. Rammentiamo che, secondo le recenti norme APA (2010), la probabilità deve sempre essere riportata in modo esatto, quindi utilizzando l'operatore '=', per valori superiori a .001, mentre negli altri casi va riportata la dicitura $p < .001$.

Rispetto ai valori di probabilità riportati dagli autori (p_a) l'operatore minore ('<', il più usato) è stato impiegato in 3522 casi (circa il 72%), l'operatore uguale ('=') in 1120 casi e l'operatore maggiore ('>') in 127. Abbiamo suddiviso i valori di probabilità riportati (p_a) e ricalcolati (p_r) in due classi: valori minori o uguali a 0.05, che abbiamo considerato riferiti a risultati statisticamente significativi[7], e valori maggiori di 0.05, ovvero non significativi.

In tabella 1 sono state incrociate le frequenze delle due variabili così ricodificate in funzione dell'operatore. L'ultima riga della tabella indica, rispettivamente, il totale dei valori ricalcolati significativi (2483) e non significativi (421); nell'ultima colonna si possono calcolare, in relazione a ciascun operatore, le frequenze totali dei risultati significativi (2095 + 450 + 60 = 2605) e non significativi (22 + 257 + 20 = 299).

Ciascuna cella della tabella rimanda ad un determinato operatore e permette di individuare i valori riportati correttamente e le eventuali inesattezze. Nel caso in cui l'operatore associato alla probabilità fosse *minore*, abbiamo osservato 2036 casi in cui gli autori riportavano la dicitura '$p <$' correttamente, ovvero, il valore effettivo della probabilità (ricalcolato) era realmente minore di 0.05. In 59 casi, invece, il risultato che gli autori indicavano come minore di 0.05 in realtà non lo era. In 20 casi gli autori hanno utilizzato la scrittura 'minore di' associando comunque una

[7] L'inclusione dell'uguale è necessaria per far ricadere nella categoria 'significativi' tutti i casi in cui viene riportata la soglia e non il valore esatto.

probabilità esatta superiore alla soglia 0.05 (ad esempio: $p < 0.181$). Infine, in 2 casi, gli autori hanno scritto $p < 0.1$ e $p < 0.28$ ma in realtà entrambe le probabilità erano inferiori a 0.05. Quando è stato utilizzato l'operatore *uguale*, per 432 + 250 = 682 volte il valore indicato dagli autori era corretto, per 18 volte la probabilità era riportata come minore della soglia ed in realtà era superiore (ad esempio $p_a = .043$ ed invece $p_r = .165$); per 7 volte, al contrario, riportata come maggiore mentre invece era minore di 0.05. L'uso dell'operatore *maggiore* infine, risulta strettamente legato alla presentazione di test non significativi: in 19 casi il risultato riportato dagli autori era corretto, non significativo. In $5 + 55 + 1 = 61$ casi il senso dell'operatore era semplicemente invertito per un evidente errore di trascrizione; ad esempio, gli autori riportavano $p_a > 0$, ed invece $p_r < 0.001$, o ancora $p_a > 0.2$ con $p_r = 0.02$.

In sintesi, abbiamo rilevato 2737 (2036 + 432 + 250 + 19) valori riportati correttamente e coerenti con l'operatore associato, 25 (20 + 5) valori corretti ma con operatore invertito e 142 valori riportati non correttamente. Dunque gli errori sono 142 su 2904, ovvero pari al 5% dei p riportati.

3.3 Distribuzione di *p*

In figura 3 sono rappresentate le distribuzioni dei valori di p riportati dagli autori (p_a, primi due pannelli a sinistra) e ricalcolati (p_r, terzo pannello e p_s, quarto pannello). Per comodità di lettura, l'intervallo di rappresentazione è limitato tra 0 e 0.1. Nel primo pannello da sinistra (p_a) sono compresi tutti i valori riscontrati in tale intervallo, ovvero 4388. Nella maggior parte dei casi (circa il 77%) tali valori però non sono esatti, ma presentati come soglia (ad esempio $p < 0.05$), pertanto la raffigurazione risente dell'eccesso di questi valori. Per eliminare questo effetto sono

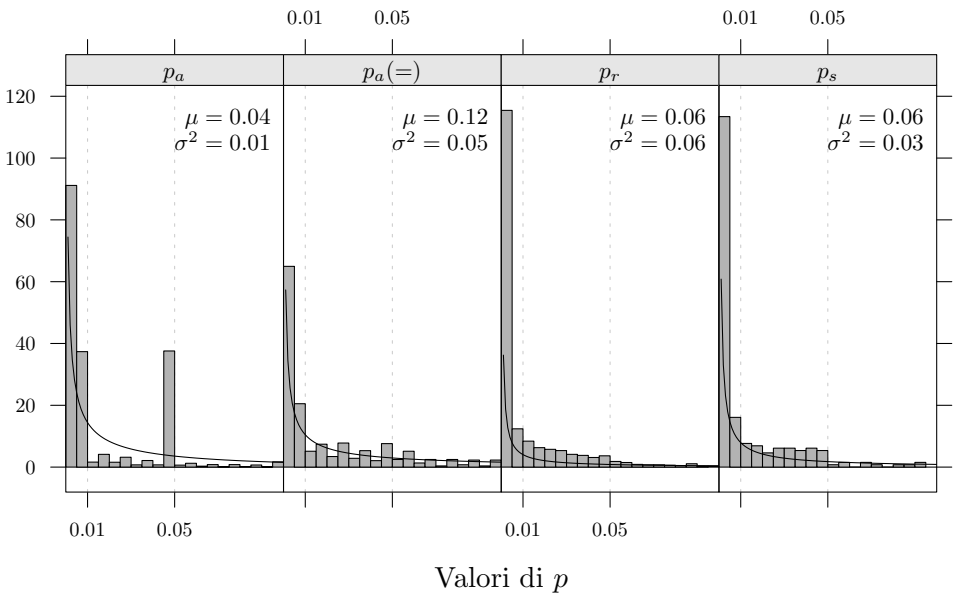

Figura 3. Distribuzioni dei valori p_a ($n = 4388$), p_a associati all'operatore uguale ($n = 774$), p_r ($n = 2612$) e p_s ($n = 242$); l'intervallo di rappresentazione è limitato tra 0 e 0.1.

stati considerati solo i casi in cui gli autori hanno riportato le probabilità esatte (in tutto, nello stesso intervallo, 774), rappresentate nel pannello indicato con $p_a(=)$. Infine, nel terzo e quarto pannello, si trovano solo i casi per cui è stato possibile ottenere il calcolo della probabilità esatta, sul totale di articoli (p_r) e sul campione di p selezionati (p_s, cfr. paragrafo 2.3), rispettivamente costituiti da 2612 e 242 valori.

Secondo quanto messo in evidenza più volte in letteratura (si veda ad es: Sellke, Bayarri, & Berger, 2001; Cumming, 2008; Simonsohn et al., 2014), in presenza di effetti forti i valori di p più bassi sono più frequenti di quelli più alti, pertanto la naturale distribuzione dei p in questi casi tende ad assumere una asimmetria positiva tanto più pronunciata quanto è maggiore l'effetto (Simonsohn et al., 2014). Evidentemente tutte le distribuzioni in figura 3 hanno tale forma; le curve nere sovraimposte sono state stimate utilizzando degli appositi modelli GAM (*Generalized Additive Models*; Rigby & Stasinopoulos, 2005) su tutti i valori di p, compresi ovviamente anche quelli al di fuori dell'intervallo 0-0.1. In sostanza, la distribuzione osservata dei *p-value* (sia riportati dagli autori sia ricalcolati), è coerente con il tipo di distribuzione attesa. In ogni pannello sono riportate anche media (μ) e varianza (σ^2) stimate dai parametri dei modelli GAM. Escludendo i casi in cui il p_a era espresso come soglia e non come valore esatto, la correlazione tra le 707 coppie di valori $p_a(=)$ e p_r è pari a 0.91. È interessante notare che le medie stimate dei p_r e dei p_s sono praticamente uguali ($\mu = 0.06$) con una minore variabilità nel caso dei p_s, i quali tendono a concentrarsi maggiormente verso i valori più bassi, come si nota confrontando le altezze delle curve stimate.

Sulla base delle osservazioni riportate da Masicampo e Lalande (2012) si è proceduto a un'analisi più dettagliata delle distribuzioni osservate, in particolare considerando i valori compresi tra 0.01 e 0.07 e utilizzando intervalli di ampiezza pari a 0.002; il risultato è rappresentato in figura

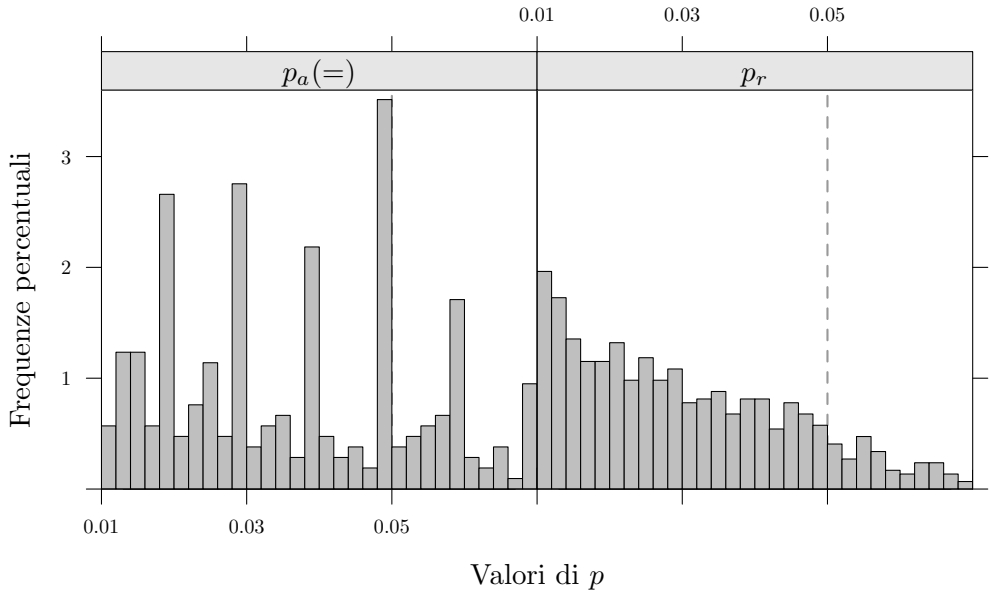

Figura 4. Frequenze percentuali dei p-value riportati dagli autori (p_a, n=334) e ricalcolati sul totale dei lavori (p_r, n=671) nell'intervallo tra .01 e .07. La linea tratteggiata indica il valore .05.

4. Nel pannello sinistro vi sono le frequenze percentuali dei valori di p_a (relative ai soli casi in cui fosse riportata la probabilità esatta), nel pannello destro quelli ricalcolati a partire dalle statistiche usabili, ossia p_r. Le altezze degli istogrammi indicano la percentuale di valori che cadono all'interno di un determinato intervallo. L'effetto che si osserva nel pannello sinistro replica quello descritto da Masicampo e Lalande (2012), ovvero la presenza di un numero particolarmente elevato di valori nelle classi immediatamente sotto le soglie .05, .04, .03, .02. Tale effetto scompare completamente nel pannello destro: probabilmente vi è una tendenza ad arrotondare i valori di p alla seconda cifra decimale per difetto.

3.4 Analisi della dimensione degli effetti

I lavori da noi considerati sono relativi ad una molteplicità di effetti rilevati in contesti di ricerca profondamente diversi tra loro e quindi non sarebbe adeguato effettuare una meta-analisi di tipo tradizionale. È invece possibile una stima quantitativa degli effetti riportati. Data la grande mole di articoli e l'effettiva difficoltà di individuare in forma automatica le informazioni utili per un preciso calcolo di tutti gli *effect size* (in particolare le numerosità campionarie) abbiamo deciso di considerare un sottoinsieme delle statistiche a nostra disposizione.

Solo r permette una determinazione univoca della dimensione campionaria a partire dai gradi di libertà, se questi sono presenti. In relazione a t, la numerosità campionaria complessiva è determinabile a patto di sapere se il valore sia riferito ad un confronto tra gruppi indipendenti o no; nel caso di gruppi indipendenti non si può però ricavare la numerosità di ciascuno di essi. Per i valori di F il calcolo è ancora più indeterminato, specialmente quando i gradi di libertà del numeratore sono maggiori di 1. Con il vincolo di avere a disposizione la numerosità campionaria, avevamo la necessità di scegliere l'indicatore di *effect size* applicabile nella maggioranza dei nostri casi.

La statistica t è convertibile in d di Cohen che a sua volta può essere trasformato in correlazione (Cohen, 1988). Dal momento però che si era in grado di stabilire automaticamente se le statistiche t derivassero da confronti tra campioni indipendenti o appaiati, e di conseguenza la numerosità associata ai gruppi, si è scelto di considerarle tutte come provenienti da confronti tra gruppi indipendenti con la stessa numerosità. In questo modo, ovvero collocandoci nella condizione ideale per il calcolo di d, abbiamo ottenuto delle stime per difetto, e pertanto più prudenti, degli *effect size*. Sulla base dei gradi di libertà (g) associati al valore della statistica, abbiamo dunque stimato la numerosità campionaria totale con la formula $n = g+2$ e utilizzato la formula di trasformazione in d di Cohen e successivamente in valori di correlazione[8]. La statistica F con un grado di libertà al numeratore è equivalente alla t, per cui abbiamo selezionato tutti i casi di questo genere e calcolato d e r anche per essi, sempre ipotizzando la condizione ideale di gruppi indipendenti di pari numerosità.

In base a questi vincoli abbiamo avuto a disposizione 406 valori t, 1044 F e 55 correlazioni, ovvero un campione complessivo di 1505 statistiche, sul totale delle 2954 individuate come usabili (pari a circa il 51%) e ricavato per tutte un valore di r. Come suggerito da Hedges e Olkin (1985) e Rosenthal (1991), esso è stato convertito in punteggi z normalizzati, utilizzando la formula di Fisher[9]

[8] Le formule utilizzate sono $d = t\sqrt{\frac{n}{(n/2)^2}}$ in cui n indica la numerosità totale e $r = \frac{d}{\sqrt{(d^2+4)}}$.

[9] I valori Z non sono stati aggiunti a questo campione in quanto non direttamente confrontabili con questi e non convertibili poiché mancanti della numerosità campionaria relativa.

$$z = .5 \times log\left[\frac{(1+r)}{(1-r)}\right]$$

In seconda battuta abbiamo considerato il campione di articoli selezionati ($n = 86$). In questo caso, avendo raccolto direttamente tutte le informazioni necessarie (laddove effettivamente presenti), è stato possibile ottenere delle stime più precise degli effect size. In particolare abbiamo identificato 54 valori di χ^2, 161 F, 23 t e 21 r. I χ^2 sono stati convertiti in r con la formula $\sqrt{\chi^2/n}$. Per i valori t e F abbiamo utilizzato le apposite formule di conversione in base al fatto che si trattasse di statistiche derivate da confronti tra i gruppi o entro i gruppi (per i dettagli si veda ad es. Lakens, 2013). In conclusione, per il campione selezionato, si è potuto disporre di 259 *effect size* di statistiche associate alle ipotesi rilevanti.

Sulla base delle soglie suggerite da Cohen (1988) per identificare effetti deboli, medi o forti, rispettivamente $r = [0.1, 0.3, 0.5]$ abbiamo identificato le corrispondenti soglie per i valori trasformati $z = [0.1, 0.31, 0.549]$. In tabella 2 sono riassunte le frequenze dei valori z all'interno delle classi risultanti da tali soglie sia per gli effetti stimati su tutti i lavori (prima riga) sia per quelli ottenuti negli 86 articoli selezionati. In entrambi i gruppi di valori risulta evidente un aumento della frequenza di effetti riportati all'aumentare della loro grandezza: nel campione totale il 38% dei valori cade nella categoria degli effetti forti, il 30% in quella tra gli effetti medi e forti; nel campione selezionato la percentuale di effetti forti sale al 45%. Oltre il 68% degli effetti riportati è superiore alla soglia di .31.

In figura 5 abbiamo rappresentato i valori di z in funzione della numerosità campionaria (per il campione totale: media = 140.82, mediana = 50, e dev. st. = 245.45; per il campione selezionato: media = 305.54, mediana = 87, e dev. st. = 707.85). I punti grigi sono gli effetti stimati sul campione totale, quelli neri gli effetti stimati sul campione selezionato. Le linee orizziontali indicano i valori delle soglie convenzionali di *effect size*, la linea verticale è posta al valore $n = 30$, limite che comprende il 30% dei lavori qui considerati. Si osserva che all'aumentare della numerosità campionaria la dimensione degli effetti si riduce considerevolmente (la correlazione è pari a -0.35 nel campione totale ed a -0.21 nel campione selezionato). Il 72% circa di studi con meno di 31 soggetti produce effetti forti, superiori a 0.55, mentre nessun lavoro basato su campioni con $n > 564$ supera tale soglia.

Al fine di dare un peso all'evidenza di questo risultato, sono stati eseguiti tre test t in forma bayesiana con il pacchetto `BayesFactor` (Morey & Rouder, 2014), prima sul campione totale e poi sul campione di effetti selezionati, definendo ogni volta come ipotesi $H_0 : \mu = z_i$, con $z_i = (0.1, 0.31, 0.549)$, calcolando il Bayes Factor (BF) e stimando la relativa distribuzione a posteriori del parametro μ su un campione di 30000 repliche[10]. In figura 6 sono rappresentate

[10] La distribuzione a posteriori di un parametro viene stimata con un processo di campionamento definito MCMC (*Markov chain Monte Carlo*; Gelfand & Smith, 1990; Geman & Geman, 1984; Morey, Rouder, Pratte, & Speckman, 2011) che prevede un numero molto elevato di repliche.

	≤ 0.1	(0.1,0.31]	(0.31,0.549]	> 0.549
campione totale	93	385	448	579
campione selezionato	7	68	68	116

Tabella 2. Frequenze dei valori z (*effect size*) stimati all'interno delle classi delimitate dalle soglie di Cohen: 0.1 (effetto piccolo), 0.31 (effetto medio), 0.549 (effetto grande). Il campione totale comprende 342 articoli, il campione selezionato 86.

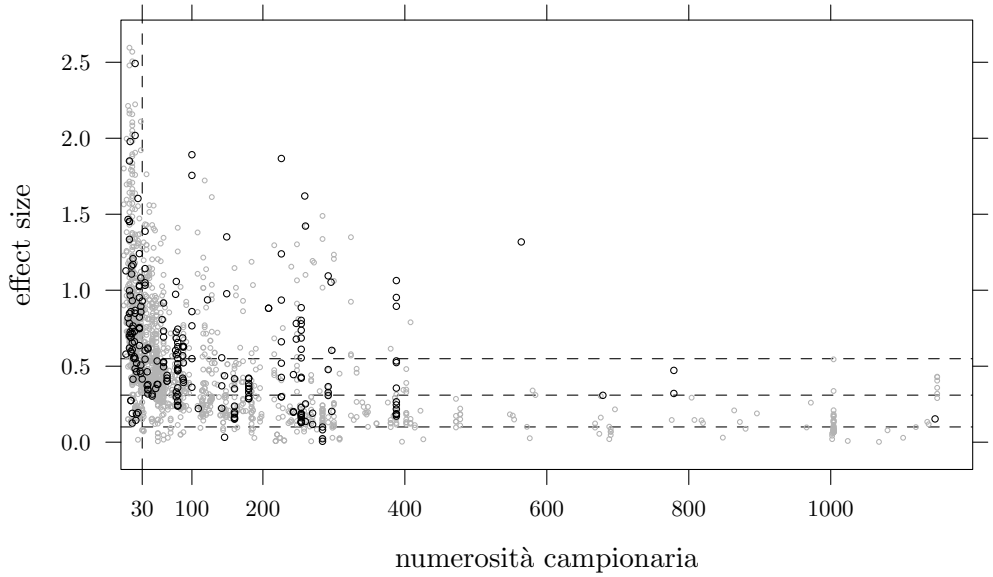

Figura 5. Dimensione degli effetti in funzione della numerosità campionaria. I punti grigi sono gli effetti stimati sul campione totale, quelli neri gli effetti stimati sul campione selezionato. Le linee orizzontali indicano le soglie convenzionali di *effect size*, la linea verticale è posta a $n = 30$.

le distribuzioni a posteriori del parametro μ, ottenute rispetto alla prima ipotesi a priori (H_0 : $\mu = 0.1$), sia nel campione totale (in grigio più chiaro) sia nel campione selezionato (in grigio scuro); le distribuzioni a posteriori rispetto alle altre ipotesi sono del tutto equivalenti e pertanto non si è ritenuto necessario rappresentarle. Tutti e tre gli scenari hanno prodotto le stesse stime a posteriori del parametro μ (le medie sono tutte di circa 0.54 nel campione totale e 0.6 nel campione selezionato) ed intervalli HDI (*Highest Density Interval*, HDI; Kruschke, 2013), ovvero il corrispondente bayesiano dell'intervallo di confidenza o, in altri termini, l'intervallo di valori più credibile, molto simili.

	H_0	$\hat{\mu}$	HDI	BF
campione totale	0.10	0.54	0.52 0.56	+100
	0.31	0.54	0.52 0.56	+100
	0.55	0.54	0.52 0.56	0.04
campione selezionato	0.10	0.60	0.54 0.66	+100
	0.31	0.60	0.54 0.66	+100
	0.55	0.60	0.54 0.66	0.35

Tabella 3. Statistiche relative alle tre ipotesi calcolate sulle distribuzioni a posteriori: media stimata ($\hat{\mu}$), intervallo HDI e BF; +100 indica che il valore calcolato è superiore a 100.

In tabella 3 sono riassunte le statistiche calcolate sulle distribuzioni a posteriori rispetto alle tre ipotesi e nei due diversi campioni. Osserviamo che i BF per le prime due ipotesi sono superiori a 100. Rammentando che il BF misura l'evidenza di un'ipotesi in funzione di un'altra, nel nostro caso indica il rapporto tra l'evidenza di $H_1 : \mu \neq z_i$ versus H_0. Pertanto nel primo e nel secondo

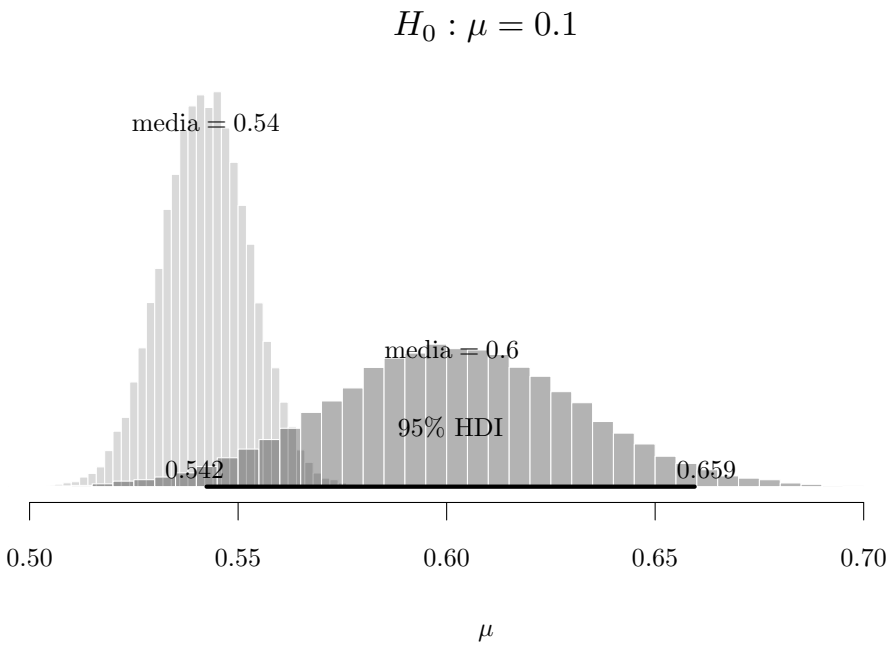

Figura 6. Distribuzioni a posteriori del parametro μ ottenute sulla base dell'ipotesi $H_0 : \mu = (0.1)$. Gli istogrammi più chiari sono relativi al campione totale mentre quelli più scuri al campione selezionato. HDI indica l'intervallo con la densità più alta (*Highest Density Interval;* Kruschke, 2013).

caso H_1 è oltre 100 volte più evidente di H_0, mentre nel terzo caso accade il contrario ovvero H_0 è $1/0.04 = 27$ e $1/0.35 = 3$ volte più evidente di H_1, rispettivamente. Detto in altri termini, lo scenario con $\mu = 0.55$ risulta, tra i tre considerati, quello più verosimile in base ai dati osservati e quindi la dimensione degli effetti riportata negli studi è stabilmente elevata.

4 CONCLUSIONI

Vi è ragione di credere che la logica sottostante alla NHST si avvii, seppure lentamente, ad essere superata. Un segnale di questo può essere individuato nel crescente interesse verso approcci alternativi alla tradizionale verifica di ipotesi. Una possibilità è l'uso della dimensione degli effetti (ad es. Fidler, Geoff, Mark, & Neil, 2004; Finch et al., 2004; Fidler, 2005; Cumming et al., 2007), che si propone di superare la semplice decisione dicotomica tra due alternative. Altra modalità è il confronto e la selezione tra modelli (es. Burnham & Anderson, 2004; Burnham, Anderson, & Huyvaert, 2011; Fox, 2008), il cui obiettivo è l'individuazione di quello che meglio descrive i dati osservati. Ancora, la sempre maggiore potenza di calcolo dei PC sta favorendo l'utilizzo di procedure bayesiane (es. Kruschke, 2010, 2011; Klugkist, van Wesel, & Bullens, 2011; Dienes, 2011), per mezzo delle quali è possibile quantificare la plausibilità di una o più ipotesi o modelli alternativi. Vero è, tuttavia, che anche gli approcci alternativi disponibili al momento non sono del tutti immuni da critiche o limiti (ad es., Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Perugini, 2014a; Simonsohn et al., 2014; Umiltà, 2014; Gigerenzer & Marewski, 2015).

A rigore, essendo legati al solo valore di p, i risultati del nostro lavoro non permettono di valutare questo cambio di paradigma nel panorama italiano più recente, anche se non vi è evidenza,

almeno nell'arco temporale qui considerato, di una inversione di tendenza. Dalla figura 2 emerge una sostanziale staticità sia nella quantità sia nelle tipologie delle statistiche usate per l'inferenza nei 16 anni presi in esame. F, t e χ^2 costituiscono da soli il 69% del totale delle statistiche impiegate dai ricercatori. Lo stesso si può dire per la percentuale di statistiche da noi classificate come *usabili*, rimasta pressoché stabile intorno al 60% (vedi fig. 1). Vale anche la pena ricordare che la stima del rapporto tra il numero di p presentati ed il numero di p legati alle ipotesi principali è tendenzialmente elevato (vedi paragrafo 2.3). Questo potrebbe indicare che un certo approccio esplorativo è sempre presente, anche ragionevolmente, nella grande maggioranza dei lavori esaminati.

Il presente studio si concentra sulle statistiche legate al *p-value* e pertanto sono state escluse tutte le analisi descrittive e gli altri casi in cui tale valore non fosse rilevante o presente (ad esempio gli indici di adattamento dei modelli di equazioni strutturali o i modelli di analisi fattoriale). Anche le tabelle di risultati sono state considerate solo nei casi in cui contenessero informazioni complete o facilmente recuperabili, e dunque sono stati esclusi molti casi in cui i valori di p non avessero una chiara controparte nel testo.

Inoltre, i principali risultati si riferiscono ai p ricalcolati (circa il 60% del totale), dunque non necessariamente rappresentativi di tutti i lavori. Ciò considerato, la correlazione tra i p riportati dagli autori (p_a) e quelli ricalcolati (p_r) è risultata alta, certamente un indice di buona accuratezza (ovviamente, nulla si può dire su quella porzione di lavori per cui non è stato possibile ricalcolare p). Dalla tabella 1 emerge come il numero di errori nella presentazione dei p si attesti attorno al 5% (evidentemente lo spirito NHST è più pervasivo di quanto non si creda). Infine, è necessario considerare che l'uso degli operatori – pur ammettendo una certa quantità di refusi – risulta incerto in una percentuale di casi non del tutto irrilevante. La stima degli *effect size* è stata possibile su un campione ancora più ridotto (circa il 51% delle statistiche *usabili* e circa il 32% del totale); pur attenendoci a principi di massima prudenza (ovvero considerare per la stima il valore più basso possibile), la distribuzione a posteriori degli *effect size* mostra dei valori medio e mediano decisamente alti, sia nel campione totale sia in quello selezionato. Inoltre, la relazione inversa tra le numerosità campionarie e le dimensioni degli effetti può suggerire la presenza di una sovrastima di questi ultimi soprattutto nei campioni piccoli ($n \leq 30$), che costituiscono comunque circa il 46% del totale degli articoli e circa il 38% del campione selezionato. Può essere di qualche utilità ricordare che al diminuire della dimensione campionaria aumentano le probabilità di commettere errori di I e II tipo (Cohen, 1988; Garthwaite, Jolliffe, & Jones, 2009).

Come detto all'inizio, l'obiettivo di questo lavoro non era valutare la qualità dei risultati pubblicati e neppure metterne in discussione la validità (si sa che dispensare buoni consigli significa precludersi la possibilità di dare il *cattivo esempio*), ma semplicemente descrivere i dati a disposizione e fornire degli spunti di riflessione, in linea con il processo di rinnovamento e ripensamento di alcuni paradigmi di fondo che sta investendo la ricerca scientifica psicologica a cui anche la comunità italiana non può e non deve sottrarsi (vedi ad esempio la sezione dedicata a questo tema in un recente numero di *Perspectives on Psychological Science*: Ledgerwood, 2014; Lakens & Evers, 2014; Sagarin, Ambler, & Lee, 2014; Stanley & Spence, 2014; Perugini, Gallucci, & Costantini, 2014; Braver, Thoemmes, & Rosenthal, 2014; Maner, 2014).

References

Agnoli, F., & Furlan, S. (2009). I cambiamenti nella verifica di ipotesi: statistiche migliori per decisioni migliori. *Giornale Italiano di Psicologia*, *36*(4), 849–882.

Altoé, G., & Pastore, M. (2013). L'effetto della numerosità sul significato di un risultato statisticamente significativo. *Giornale Italiano di Psicologia*, *40*(2), 367–376.

American Psychological Association. (1994). *Publication Manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

American Psychological Association. (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678.

Balboni, G., & Cubelli, R. (2009). Convergenza delle evidenze e molteplicità delle ipotesi: la verifica dell'ipotesi nulla nella ricerca psicologica. *Giornale Italiano di Psicologia*, *36*(4), 883–898.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously Cumulating Meta-Analysis and Replicability. *Perspectives on Psychological Science*, *9*(3), 333–342.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates, Hillsdale, NJ.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cumming, G. (2008). Replication and p Intervals: p Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives On Psychological Science*, *3*(4), 286–300.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., ... Wilson, S. (2007). Statistical Reform in Psychology: Is Anything Changing? *Psychological Science*, *18*(3), 230–232.

Di Nuovo, S. (2009). Cosa significa significativo? Ovvero: ci sono alternative all'ipotesi alternativa? *Giornale Italiano di Psicologia*, *36*(4), 899–911.

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science*, *6*(3), 274–290.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal Of Statistical Software*, *25*(5), 1–54.

Fidler, F. (2005). *From Statistical Significance to Effect Estimation: Statistical Reform in Psychology, Medicine and Ecology.* University of Melbourne, Department of History and Philosophy

of Science. Scaricabile da `http://books.google.it/books?id=yjiCNAAACAAJ`

Fidler, F., Geoff, C., Mark, B., & Neil, T. (2004). Statistical reform in medicine, psychology and ecology. *The Journal of Socio-Economics*, *33*(5), 615–630.

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., ... Goodman, O. (2004). Reform of statistical inference in psychology: The case of Memory & Cognition. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 312–324.

Fox, J. (2008). *Applied regression analysis and generalized linear models*. Sage Publications.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal Of Mathematical Psychology*, *57*(5), 153–169.

Garthwaite, P., Jolliffe, I., & Jones, B. (2009). *Statistical inference (Second Edition)*. Oxford University Press.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Gelman, A. (2013a). Interrogating p-values. *Journal Of Mathematical Psychology*, *57*(5), 188–189.

Gelman, A. (2013b). P Values and Statistical Practice. *Epidemiology*, *24*(1), 69–72.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate Science The Idol of a Universal Method for Scientific Inference. *Journal of Management*, *41*(2), 421–440.

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-analysis*. Academic Press.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164.

Hubbard, R., & Lindsay, R. M. (2008). Why P values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, *18*(1), 69–88.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), 696–701.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, *110*(48), 19313–19317.

Kline, R. B. (2004). *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. Washington, DC: American Psychological Association.

Klugkist, I., van Wesel, F., & Bullens, J. (2011). Do we know what we test and do we test what we want to know? *International Journal of Behavioral Development*, *35*(6), 550–560.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300.

Kruschke, J. K. (2011). *Doing Bayesian Data Analysis*. Academic Press, Burlington, MA.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*, 863. doi: 10.3389/fpsyg.2013.00863

Lakens, D., & Evers, E. R. K. (2014). Sailing From the Seas of Chaos Into the Corridor of Stability: Practical Recommendations to Increase the Informational Value of Studies. *Perspectives on Psychological Science*, *9*(3), 278–292.

Ledgerwood, A. (2014). Introduction to the Special Section on Advancing Our Methods and Practices. *Perspectives on Psychological Science*, *9*(3), 275–277.

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303.

Maner, J. K. (2014). Let's Put Our Money Where Our Mouth Is: If Authors Are to Change Their Ways, Reviewers (and Editors) Must Change With Them. *Perspectives on Psychological Science*, *9*(3), 343–351.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal Of Experimental Psychology*, *65*(11), 2271–2279.

Maxwell, S. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163.

Morey, R. D., & Rouder, J. N. (2014). BayesFactor: Computation of Bayes factors for common designs [Computer software manual]. (R package version 0.9.9)

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*(5), 368–378.

Morris, J., & Gardner, M. (1988). Calculating Confidence-Intervals For Relative Risks (Odds Ratios) And Standardized Ratios And Rates. *British Medical Journal*, *296*(6632), 1313–1316.

Pastore, M. (2009). I limiti dell'approccio NHST e l'alternativa Bayesiana. *Giornale Italiano di Psicologia*, *36*(4), 925–938.

Perugini, M. (2014a). Ciò che non uccide, rende più forte: La reazione della psicologia alla crisi di credibilità. *Giornale Italiano di Psicologia*, *41*(1), 127–138.

Perugini, M. (2014b). La crisi internazionale di credibilità della psicologia come un'opportunità di crescita: Problemi e possibili soluzioni. *Giornale Italiano di Psicologia*, *41*(1), 23–46.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard Power as a Protection Against Imprecise Power Estimates. *Perspectives on Psychological Science*, *9*(3), 319–332.

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Scaricabile da `http://www.r-project.org`

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape, (with discussion). *Applied Statistics*, *54*(3), 507–554.

Rosenthal, R. (1991). *Meta-Analytic Procedures for Social Research*. SAGE Publications.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An Ethical Approach to Peeking at Data. *Perspectives on Psychological Science*, *9*(3), 293–304.

Scargle, J. D. (2000). Publication Bias: The 'File-Drawer' Problem in Scientific Inference. *Journal of Scientific Exploration*, *14*(1), 91–106.

Schimmack, U. (2012). The Ironic Effect of Significant Results on the Credibility of Multiple-Study Articles. *Psychological Methods*, *17*(4), 551–566.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*(7335), 437.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p Values for Testing Precise Null Hypotheses. *The American Statistician*, *55*(1), 62–71.

Shea, C. (2011). Fraud Scandal Fuels Debate Over Practices of Social Psychology. *Chronicle of Higher Education*. Scaricabile da `http://chronicle.com/article/As-Dutch-Research-Scandal/129746`

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534.

Stanley, D. J., & Spence, J. R. (2014). Expectations for Replications: Are Yours Realistic? *Perspectives on Psychological Science*, *9*(3), 305–318.

Sterne, J. A., & Smith, G. D. (2001). Sifting the evidence—what's wrong with significance tests? *Physical Therapy*, *81*(8), 1464–1469.

Umiltà, C. (2014). Le frodi scientifiche si possono perpetrare anche con l'inferenza statistica. *Giornale Italiano di Psicologia*, *41*(1), 117–120.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals - Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydi, O. (2008). Why Current Publication Practices May Distort Science. *PLOS Medicine*, *5*(10), 1418–1422.

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance*. Ann Arbor, MI: University of Michigan Press.