Bayes Factor e p-value: così vicini, così lontani

Massimiliano Pastore[1] e Gianmarco Altoé[2]

[1]Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università di Padova.
[2]Dipartimento di Psicologia, Università di Cagliari.

Sommario Negli ultimi anni l'approccio inferenziale più utilizzato nella ricerca psicologica, il NHST (Null Hypothesis Significance Testing), è stato al centro di profonde critiche accompagnate da un crescente interesse verso approcci alternativi. Tra questi, l'approccio Bayesiano è uno dei più rilevanti. Sebbene molti studi abbiano evidenziato le differenze tra i due approcci, poca attenzione è stata rivolta ai loro punti in comune. Il presente studio si propone di analizzare, sia dal punto di vista matematico che applicativo, la relazione tra i due approcci. A titolo esemplificativo è stato considerato il caso del t-test per campioni indipendenti. In particolare, il p-value associato al NHST e il Bayes factor (BF) di derivazione bayesiana sono stati confrontati, attraverso uno studio Monte Carlo, per diversi livelli di numerosità campionaria e di dimensione dell'effetto. I risultati indicano che, sebbene la relazione tra p-value e BF sia sostanzialmente deterministica, le interpretazioni dei due indicatori possono portare a decisioni diverse. Le maggiori discrepanze si hanno nei casi di bassa ed elevata numerosità campionaria e in quelli di ridotta dimensione dell'effetto.

Bayes Factor and *p-value*: so close, so far.

Summary. In recent years, the inferential approach most widely used in psychological research, i.e. Null Hypothesis Significance Testing (NHST), has been subject to profound criticism accompanied by a renewed interest in alternative approaches. Among these, the Bayesian inference approach is one of the most relevant. Despite several studies have shown differences between the two approaches, little attention has been devoted to their commonalities. The current study aims to analyze the relation between the two approaches from both a mathematical and an applied perspective. As an illustrative example, we considered the independent samples t-test. In particular, the p-value associated to NHST and the Bayes Factor (BF) derived from Bayesian approach were compared on different levels of sample size and effect size via Monte Carlo study. Results indicate that, even if the relation between p-value and BF is substantially deterministic, the interpretation of the two indices can lead to different decisions. Discrepancies occur especially in cases of small and large sample size as well as low effect size.

Parole chiave: *p-value*, Bayes Factor, posterior probability, *t*-test, Monte Carlo method

1 INTRODUZIONE

Nella ricerca psicologica, la valutazione della significatività statistica dei risultati ottenuti si basa quasi esclusivamente sull'approccio noto come *Null hypothesis significance testing* (NHST; Cohen, 1994). Il procedimento alla base del NHST può essere sinteticamente descritto come segue: a partire da x, un campione di osservazioni, si calcola una determinata statistica test $t = t(x)$; conoscendo $f(t|H_0)$, ossia la distribuzione campionaria di t sotto l'ipotesi H_0 è possibile determinare la probabilità di ottenere un risultato uguale o più estremo rispetto a quello osservato empiricamente (probabilità indicata come *p-value*). Se tale probabilità risulta inferiore ad una soglia critica definita a priori (solitamente 0.05), il ricercatore conclude rigettando l'ipotesi H_0. In questo caso si può parlare di risultato statisticamente significativo.

La criticità dell'approccio NHST è stata ampiamente dibattuta a livello internazionale (si veda ad esempio Cohen, 1994; Dixon, 2003; Frick, 1996; Hagen, 1997; Nickerson, 2000; Schmidt,

1996; Wagenmakers, 2007; Ziliak e McCloskey, 2008) ed anche nel panorama italiano (Agnoli e Furlan, 2009; Balboni e Cubelli, 2009; Di Nuovo, 2009; Pastore, 2009). Ció nonostante, una recente rassegna relativa alla letteratura psicologica (Bakker e Wicherts, 2011) ha rilevato circa il 18% di risultati statistici riportati incorrettamente, come a dire che ancora oggi l'approccio NHST risulta poco compreso.

Un primo problema strettamente legato a tale approccio è l'impossibilità di valutare l'evidenza di H_0 sulla base dei valori del *p-value* (si veda ad esempio Berger e Sellke, 1987; Wagenmakers, 2007). Tale impossibilità viene riconosciuta nel 1999 anche dall'APA Task Force on Statistical Inference nell'invito a non utilizzare l'espressione "*accept the null hypothesis*" (Wilkinson & the Task Force on Statistical Inference, 1999, p. 599). Nei classici paradigmi di inferenza statistica H_0 è l'ipotesi di invarianza o di assenza di effetti ma vi sono casi molto importanti in cui sarebbe utile poterne valutare l'evidenza (si pensi ad esempio alla valutazione dell'inefficacia di un trattamento o all'invarianza di un tratto latente rispetto a caratteristiche socio-demografiche; Gallistel, 2009).

Al tempo stesso, l'utilizzo del *p-value* non permette neppure di trarre conclusioni sulla plausibilità dell'ipotesi alternativa H_1 e questo per il semplice fatto che negli algoritmi utilizzati per il calcolo tale ipotesi non viene mai direttamente considerata. Si pensi al caso elementare del test t per il confronto tra le medie di due campioni indipendenti in cui $H_0 : \mu_1 = \mu_2$ esprime l'ipotesi che le medie delle popolazioni da cui i campioni provengono siano uguali contro $H_1 : \mu_1 \neq \mu_2$ che esprime il contrario. Questo test si basa sul calcolo della seguente statistica:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} \tag{1}$$

in cui $\bar{x}_1$ e $\bar{x}_2$ sono le medie campionarie e $\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}$ la stima dell'errore standard della differenza tra le medie campionarie. Una volta calcolato il valore di t ciò che consente di ottenere il *p-value* ad esso associato è la conoscenza della distribuzione campionaria di t nel caso in cui sia vera H_0. Si noti bene che il *p-value* è funzione di t condizionatamente ad H_0; l'ipotesi alternativa H_1 non ha alcun ruolo nel calcolo di tale probabilità, il che ha del paradossale visto che spesso è H_1 l'ipotesi che guida l'interesse di ricerca.

Un terzo elemento critico riguarda l'incoerenza nelle decisioni cui si arriva in base al comportamento del *p-value*. Possiamo dire che un approccio è coerente se, data una ipotesi vera, aumentando la numerosità campionaria, aumenta anche l'evidenza in favore dell'ipotesi (Rouder, Speckman, Sun, Morey e Iverson, 2009). Si può facilmente verificare che, fissato un livello critico a priori $\alpha = 0.05$, aumentando la numerosità del campione il *p-value* non si comporta sempre in maniera ottimale. Se H_0 è falsa il *p-value* tende a zero e quindi diminuisce l'evidenza a favore di H_0, il che è un buon risultato anche se, lo ricordiamo, questo non permette di valutare l'evidenza a favore dell'ipotesi alternativa. Se però H_0 è vera, la probabilità di rigettarla per errore rimane costante al valore fissato $\alpha = 0.05$, ossia è indipendente dalla numerosità campionaria, e quindi non aumenta l'evidenza a favore di H_0. Ancora più paradossale è il caso in cui H_0 è quasi vera ovvero quando, ad esempio, la differenza tra le medie delle popolazioni sia minima (e quindi non sostanziale; Agnoli e Furlan, 2008). In questo caso l'aumento di n comporta un abbassamento del *p-value* che tende a zero e con ciò aumenta l'evidenza contro H_0. In altri termini, ancora in relazione al test t citato sopra, si può osservare che differenze minime tra le medie vere delle popolazioni μ_1 e μ_2 con una *appropriata numerosità statistica* portano certamente al rifiuto di H_0[1].

In sintesi, l'aspetto più paradossale del classico approccio NHST è che non permette la valutazione dell'evidenza statistica di H_0 né tantomeno dell'ipotesi alternativa. Anche alcune proposte alternative al *p-value* quali, ad esempio, l'uso degli intervalli di confidenza (IC) e la probabilità di replicazione (p_{rep}; Killen, 2005) non permettono di superare questo problema. Per quanto riguarda gli IC, dato che dipendono dalle stesse statistiche da cui si deriva il *p-value* soffrono degli stessi limiti e pertanto possono essere utilizzati con profitto nella descrizione dei dati ma non per l'inferenza sugli stessi (Rouder e Morey, 2005). Il p_{rep} che inizialmente sembrava promettente, si è però rivelato del tutto simile al *p-value* e con la stessa debolezza nella valutazione dell'evidenza di H_0 (Wagenmakers e Grunwald, 2006; Iverson, Lee, Zhang e Wagenmakers, 2009).

[1] La criticità appena descritta sarà evidente nei risultati della simulazione descritta in seguito.

Recentemente, il dibattito critico verso NHST è stato accompagnato da numerose proposte alternative legate all'approccio Bayesiano (Dienes, 2011; Kruschke, 2010; Masson, 2011; Rouder, et al. 2009; Wagenmakers, 2007). Tale approccio si differenzia fortemente da NHST in quanto basato su valutazioni comparative (l'ipotesi A è più plausibile dell'ipotesi B) piuttosto che su una decisione secca dicotomica (il risultato è significativo o meno).

Obiettivo del presente lavoro è mettere a confronto il *p-value* derivato dal classico NHST con il *Bayes Factor* (BF) e la probabilità a posteriori di H_0 (ricavabile tramite esso), derivati dall'impostazione bayesiana. Per il confronto abbiamo scelto il test t per campioni indipendenti data la sua notorietà e semplicità d'uso. In particolare abbiamo cercato di confrontare i due metodi in relazione al controllo dell'errore di I tipo e della potenza. Pur essendo ben consapevoli che tale confronto possa sembrare improprio visto che la filosofia dell'approccio Bayesiano esce da questo schema, ci è sembrata utile una comparazione del genere in quanto è più probabile che l'eventuale lettore abbia una maggiore familiarità con l'approccio NHST e pertanto possa apprezzare le differenze che emergono tra i due approcci già a questo livello.

Pertanto, dapprima illustreremo brevemente il significato del *p-value*, del BF e della probabilità a posteriori in relazione al test t. Successivamente descriveremo il disegno sperimentale e presenteremo i risultati della simulazione Monte Carlo che abbiamo realizzato.

2 *P-VALUE* E *BAYES FACTOR*

2.1 Il *p-value*

Il *p-value* è una probabilità condizionata: data una generica statistica test t la cui funzione di densità sotto H_0 sia $f(t|H_0)$ e sia t_0 un valore rilevato empiricamente della stessa statistica allora

$$p\text{-}value = P(|t| \geq t_0|H_0)$$

ovverosia, la probabilità di ottenere un valore assoluto di t maggiore o uguale a t_0 condizionata al fatto che l'ipotesi H_0 sia vera. Detto in altri termini, si tratta della probabilità di ottenere un valore della statistica test più estremo di quello empiricamente rilevato (t_0) a condizione che H_0 sia vera.

In questo lavoro abbiamo preso come esempio il test t per campioni indipendenti e considerato il *p-value* associato a tale test per ipotesi alternativa bidirezionale. Poste quindi le ipotesi $H_0 : \mu_1 = \mu_2$ e $H_1 : \mu_1 \neq \mu_2$ e calcolata la statistica test t_0 con la (1), possiamo ricavare il valore p

$$p = \int_{-\infty}^{-t_0} f(t|H_0)dt + \int_{t_0}^{+\infty} f(t|H_0)dt \tag{2}$$

ossia, in pratica, sommando le aree delle code della funzione di densità $f(t|H_0)$. A questo punto, secondo il classico schema NHST prenderemo una decisione sulle ipotesi in base al valore ottenuto di p, favorendo H_1 se $p < \alpha$, favorendo H_0 se $p \geq \alpha$ e dove α indica il livello di significatività scelto, che nella prassi più comune è fissato a 0.05.

2.2 *Bayes Factor* e probabilità a posteriori

Il *Bayes Factor* (BF) viene definito come probabilità di un risultato sotto un'ipotesi relativamente ad un altra ipotesi, formalmente:

$$BF = \frac{f(x|H_0)}{f(x|H_1)}$$

in cui $f(x|H_0)$ rappresenta la verosimiglianza del risultato x nel caso in cui sia vera H_0 e $f(x|H_1)$ la verosimiglianza dello stesso risultato x nel caso in cui sia vera H_1. Diversamente dal *p-value* vediamo subito che nel calcolo del BF vengono tenute in conto entrambe le ipotesi e pertanto l'interpretazione del risultato ottenuto può essere fatta in funzione dell'evidenza di una ipotesi

rispetto all'altra. Quando tale rapporto supera il valore 1 vuol dire che i dati sono più verosimili sotto l'ipotesi H_0 che sotto H_1. Ad esempio $BF = 3$ vuol dire che il risultato osservato è tre volte più verosimile se è vera H_0 rispetto ad H_1.

Nel sistema di analisi Bayesiano si cerca di stabilire la plausibilità delle ipotesi in esame, tradotte nei valori dei parametri che le caratterizzano, prima di osservare dei nuovi dati. Questa credibilità dei valori dei parametri viene chiamata distribuzione a priori. Nella modellazione Bayesiana, la credibilità corrisponde alla probabilità in quanto coincidenti da un punto di vista matematico (Kruschke, 2011). Detto in altri termini, le probabilità a priori sono le probabilità che vengono assegnate in base alle credenze o alle conoscenze del ricercatore relative a studi precedenti o evidenze sperimentali. In questo sistema, il BF diventa un peso, basato sui dati osservati, per modificare il rapporto tra le probabilità a priori ed ottenere il rapporto tra le probabilità a posteriori nel seguente modo:

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{f(x|H_0)}{f(x|H_1)} \frac{P(H_0)}{P(H_1)} = BF \frac{P(H_0)}{P(H_1)}$$

Se le probabilità a priori delle due ipotesi sono uguali, BF indica direttamente il rapporto a posteriori. In questo caso quindi $BF = 3$ vorrebbe dire che, dopo aver rilevato un campione x, l'ipotesi H_0 è tre volte più probabile di H_1. In altri termini possiamo dire che la probabilità a posteriori di H_0 è $3/4 = 0.75$ mentre la probabilità a posteriori di H_1 è $1 - P(H_0|x) = 0.25$.

Nel caso del test t, a differenza dell'approccio NHST che, come visto, non tiene conto dell'ipotesi alternativa, per poter calcolare il BF e la probabilità a posteriori è necessario definire una distribuzione a priori del parametro oggetto del test sotto H_1. Ad esempio, se poniamo $\mu = \mu_1 - \mu_2$ potremmo indicare:

$$\mu \sim \mathcal{N}(0, \sigma^2)$$

ovvero che la distribuzione di μ (differenza tra le medie) è normale con media 0 e varianza σ^2. Questa formulazione presenta però il problema della scelta di quest'ultimo valore σ^2 il quale può avere effetto sul BF (Rouder et al., 2009). Una possibile soluzione può essere quella di porre $\sigma^2 = \infty$, ovverosia adottare una distribuzione a priori non informativa (*Noninformative Prior Distribution*; Robert, 2001).

Una migliore soluzione tuttavia, consiste nel riformulare le ipotesi in termini di rapporto tra media (o, come nel nostro caso, tra differenza delle medie) e deviazione standard $\delta = \mu/\sigma$ per cui si ha $H_0 : \delta = 0$ e come distribuzione a priori per δ nell'ipotesi alternativa viene utilizzata la Cauchy (Jeffreys, 1961). Questa distribuzione dipende da due parametri e, nel caso particolare da noi considerato in cui tali parametri sono 0 e 1, essa diventa l'equivalente della distribuzione t con un grado di libertà:

$$\delta \sim Cauchy(0, 1) = t(1)$$

Sulla base di quest'ultima formulazione Rouder et al. (2009) hanno proposto il metodo per il calcolo del BF nel test t. Inoltre, per il calcolo delle probabilità a posteriori l'algoritmo, descritto in dettaglio da Wetzels, Raaijmakers, Jakab, e Wagenmakers (2009), utilizza il metodo MCMC (*Markov Chain Monte Carlo*; Gamerman e Lopes, 2006).

3 SIMULAZIONE E DISEGNO SPERIMENTALE

Il programma per la simulazione Monte Carlo è stato scritto ad hoc in ambiente R (R Development Core Team, 2010), le distribuzioni a posteriori sono state ottenute con il programma Winbugs (Lunn, Thomas, Best, & Spiegelhalter, 2000). Questo esperimento prende spunto da un esempio illustrato da Berger e Sellke (1987) e ne replica i risultati essenziali.

Per la generazione dei dati abbiamo fatto variare la numerosità campionaria (n) e la differenza tra le medie (espressa come dimensione dell'effetto, d). In relazione a n abbiamo considerato quattro casi rappresentativi delle situazioni più comuni nella ricerca psicologica, a partire da campioni piccoli di 15 unità statistiche ciascuno, fino a campioni grandi con 300 unità statistiche.

Per la manipolazione della dimensione dell'effetto abbiamo scelto 6 valori del d di Cohen (1988) a partire da 0 che indica l'effetto nullo e quindi ipotesi H_0 vera. Oltre ai tradizionali valori indicati da Cohen come soglie per definire l'effetto piccolo (.2), medio (.5) e grande (.8) abbiamo considerato anche dei valori prossimi a zero (.05 e .1) a rappresentare casi in cui la differenza tra le medie delle popolazioni, pur essendo diversa da zero, non può considerarsi sostanzialmente rilevante (ipotesi nulla quasi-vera o non completamente vera; Meehl, 1997).

Per ciascuna combinazione di numerosità ed effect size abbiamo generato due insiemi di dati con distribuzione normale e varianza unitaria e su questi effettuato un test t per campioni indipendenti in cui le ipotesi contrapposte erano

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Sugli stessi dati abbiamo poi calcolato il BF utilizzando la procedura descritta da Wetzels et al. (2009) che implementa il test descritto nella formulazione di Rouder et al. (2009). Per le distribuzioni a priori abbiamo utilizzato sia la Cauchy che la Normale Standard ($\mu = 0$, $\sigma = 1$). Le probabilità a priori delle ipotesi H_0 e H_1 sono state, per semplicità, poste uguali a 0.5.

In sintesi, il disegno sperimentale è stato strutturato in base ai seguenti fattori:

1. numerosità campionaria a 4 livelli (n = 15, 50, 100, 300)
2. dimensione dell'effetto a 6 livelli (d = 0, .05, .1, .2, .5, .8)
3. tipo di distribuzione a priori del parametro δ (Cauchy e Normale)

Sulla base delle 24 condizioni derivate dall'incrocio $n \times d$ sono state generate 1000 coppie di dati con distribuzione normale e varianza unitaria. Per ciascuna delle 24000 coppie così generate abbiamo:

- effettuato un t-test per campioni indipendenti e calcolato il relativo *p-value*
- calcolato il BF e la relativa probabilità a posteriori di H_0 utilizzando una distribuzione a priori di tipo Cauchy (BF_C)
- calcolato il BF e la relativa probabilità a posteriori di H_0 utilizzando una distribuzione a priori normale (BF_N)

4 RISULTATI

In questa sezione presentiamo i risultati della simulazione suddivisi in quattro sezioni. Nella prima abbiamo valutato la relazione diretta tra il *p-value* e BF. Nella seconda sezione abbiamo convertito i valori di BF ottenuti con la simulazione in probabilità a posteriori. Questa trasformazione ha reso più facile l'analisi della relazione tra NHST e approccio Bayesiano in quanto basata sul confronto tra probabilità. Nella terza sezione abbiamo riportato il confronto in relazione al controllo dell'errore di I tipo e della potenza. Infine, nella quarta sezione abbiamo analizzato la relazione funzionale tra il *p-value* e la probabilità a posteriori.

4.1 Relazione tra *p-value* e BF

Come prima analisi abbiamo considerato la semplice relazione osservata tra i valori di p e quelli del BF. In relazione ai valori di ottenuti per BF_C e BF_N abbiamo riscontrato una lieve differenza in termini numerici nel senso che i valori ottenuti adottando come distribuzione a priori la Cauchy (media 3.58, ds 3.93) sono superiori a quelli ottenuti con la normale (media 2.82, ds 3.08) in tutte le condizioni di numerosità e dimensione dell'effetto. La correlazione tra BF_C e BF_N è comunque risultata pari ad uno pertanto per il momento presenteremo solo i risultati relativi alla distribuzione Cauchy. Ricordiamo comunque che l'uso della distribuzione Cauchy (Jeffreys, 1961) viene considerato migliore della normale in questo contesto, ovvero per modellare la distribuzione a priori del parametro di effect size δ (Rouder, et. al 2009).

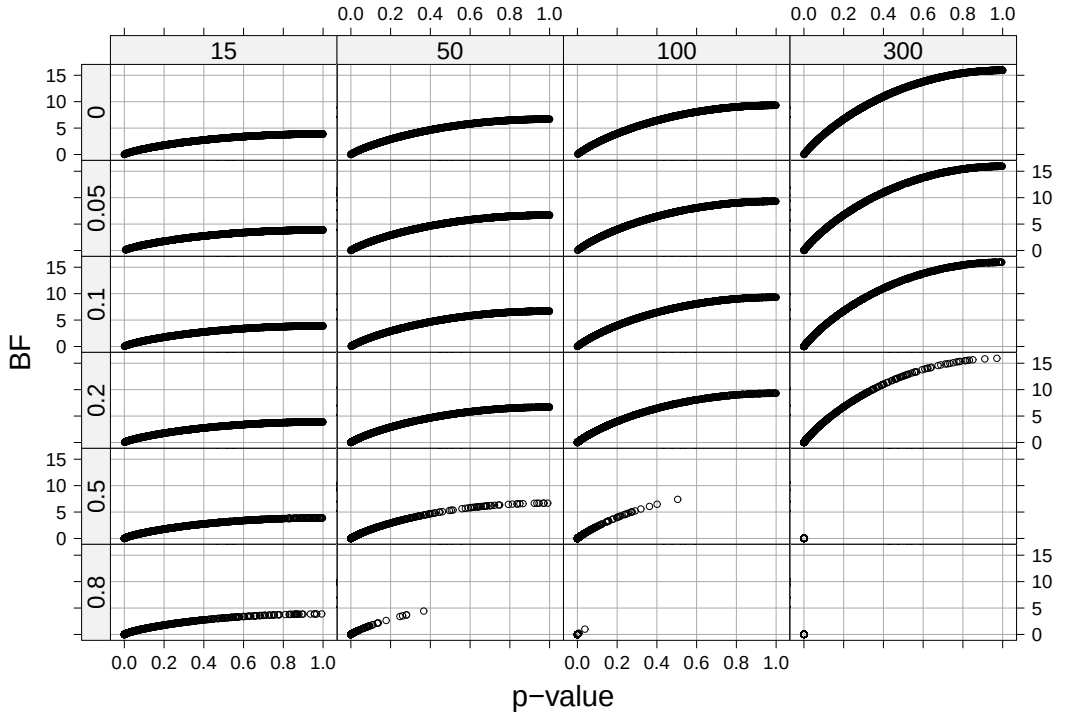

Figura 1. Relazione tra *p-value* e *Bayes Factor* calcolato con distribuzione a priori Cauchy, in funzione della dimensione dell'effetto (d, rappresentato dai grafici in riga) e della numerosità campionaria (n, rappresentata dai grafici in colonna).

In figura 1 sono rappresentati i valori del BF ottenuti con distribuzione a priori Cauchy in funzione dei valori di p per tutte le 24 combinazioni di n e d ognuna delle quali dà luogo ad un diverso grafico, ciascuno composto da 1000 punti. Sulle righe possiamo leggere i grafici in relazione ai diversi valori di dimensione dell'effetto: nella prima riga in alto $d = 0$ (effetto nullo), nell'ultima in basso $d = 0.8$ (effetto forte). Sulle colonne invece possiamo leggere i grafici in funzione della numerosità campionaria: nella prima colonna a sinistra $n = 15$ (campione piccolo), nell'ultima a destra $n = 300$ (campione grande).

La relazione tra le due grandezze appare evidente in tutti gli incroci. Inoltre osserviamo che tale relazione dipende strettamente dall'interazione tra i fattori n e d nel senso che l'aumento della numerosità campionaria implica un aumento dei valori di BF quando le dimensioni dell'effetto sono basse mentre per valori alti di d l'aumento di n porta i valori di p e di BF ad essere sempre più prossimi a zero.

4.2 Relazione tra *p-value* e probabilità a posteriori

Per una lettura più efficace abbiamo confrontato i *p-value* con $P(H_0|x)$, ossia le probabilità a posteriori di H_0 ottenuta per via Bayesiana[2]. In figura 2 sono rappresentate le probabilità a posteriori di H_0 in funzione dei valori di p per tutte le 24 combinazioni di n e d ognuna delle quali da luogo ad un diverso grafico composto da 1000 coppie di valori. Anche in questo caso emerge una relazione evidente.

Il confronto tra il *p-value* e $P(H_0|x)$ permette di valutare la concordanza dei due approcci in relazione alla decisione su H_0. Secondo il metodo NHST, H_0 non viene rigettata quando $p \geq .05$, secondo il metodo Bayesiano, si considera più evidente H_0 quando $P(H_0|x) \geq .5$

[2] Dato che H_0 e H_1 avevano la stessa probabilità a priori (ovvero 0.5), la probabilità a posteriori è direttamente calcolabile dal BF (vedi sez. 2.2).

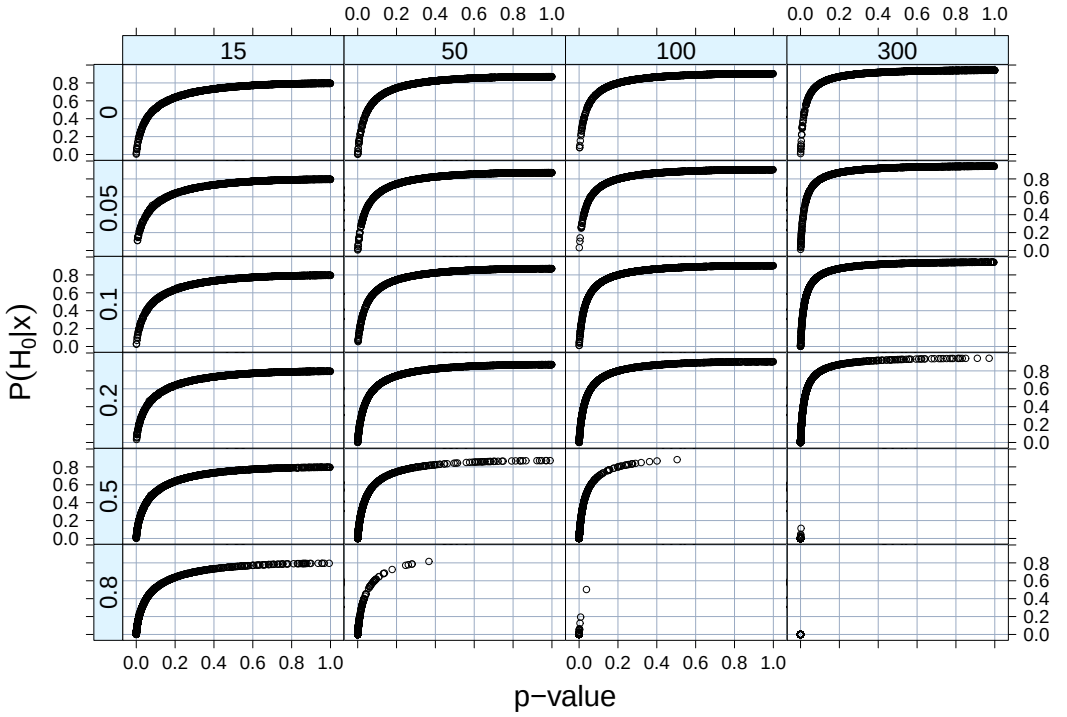

Figura 2. Relazione tra *p-value* e probabilità a posteriori di H_0 calcolate sulla base della distribuzione a priori Cauchy, in funzione della dimensione dell'effetto (d, rappresentato dai grafici in riga) e della numerosità campionaria (n, rappresentata dai grafici in colonna).

La tabella 1 confronta la percentuale di volte in cui il *p-value* e probabilità a posteriori cadono congiuntamente in determinati intervalli, sempre in funzione di n e d. Per quanto riguarda il *p-value* abbiamo considerato i valori per cui si ottiene la significatività statistica ($[p-] : p < 0.05$) rispetto a quelli per cui non si ottiene ($[p+]$). Anche i valori delle probabilità a posteriori (indicati in tabella con pH_0) sono stati suddivisi in due classi: valori per cui H_1 è più evidente di H_0 ($[pH_0^-] : pH_0 < 0.5$) e valori per cui H_0 è più evidente di H_1 ($[pH_0^+] : pH_0 \geq 0.5$). Pertanto: la colonna indicata con $[p-,pH_0^-]$ riporta la percentuale di casi in cui entrambi gli approcci porterebbero a rigettare H_0, mentre la colonna $[p+,pH_0^+]$ la percentuale di casi in cui si otterrebbe l'esito opposto (non viene rigettata H_0). Sommando i valori di queste colonne otteniamo la percentuale di casi in cui i due metodi porterebbero alla stessa decisione (accordo).

Le due colonne centrali riportano invece la percentuale di casi per cui si otterrebbero esiti diversi (disaccordo) ovvero un *p-value* non significativo ma H_1 più evidente di H_0 ($[p+,pH_0^-]$) e viceversa ($[p-,pH_0^+]$).

Un primo risultato interessante si osserva se consideriamo le percentuali medie di accordo in funzione di n, ossia: 93.2 per $n = 15$, 99.6 per $n = 50$, 98.2 per $n = 100$ e 95.2 per $n = 300$. Il massimo grado di concordanza tra le due procedure si ha con una numerosità medio-bassa, l'aumento o la diminuzione nella dimensione del campione comportano una maggiore diversità di conclusioni. In relazione a d, abbiamo le seguenti percentuali medie: 98 ($d = 0$), 97.3 ($d = 0.05$), 96 ($d = 0.1$), 94.4 ($d = 0.2$), 96.4 ($d = 0.5$) e 97.3 ($d = 0.8$). In pratica l'accordo massimo si ha per valori più estremi di d mentre per valori intermedi il grado di accordo tra i due metodi si riduce.

Leggendo i valori riportati nella tabella 1 notiamo che con i gruppi più piccoli ($n = 15$) l'aumento della dimensione dell'effetto comporta un aumento dei casi per cui NHST porta ad un risultato non significativo mentre con la probabilità a posteriori H_0 risulta meno evidente di H_1. Non osserviamo mai il caso opposto, ovvero un *p-value* significativo ma con H_0 più evidente di H_1.

n	d	$[p-,pH_0^-]$	$[p+,pH_0^-]$	$[p-,pH_0^+]$	$[p+,pH_0^+]$	n	d	$[p-,pH_0^-]$	$[p+,pH_0^-]$	$[p-,pH_0^+]$	$[p+,pH_0^+]$
15	0.00	5.2	3.7	0.0	91.1	100	0.00	3.3	0.0	1.2	95.5
	0.05	4.4	4.9	0.0	90.7		0.05	3.7	0.0	1.5	94.8
	0.10	5.0	4.7	0.0	90.3		0.10	8.5	0.0	2.3	89.2
	0.20	8.1	5.3	0.0	86.6		0.20	27.4	0.0	3.9	68.7
	0.50	24.5	11.4	0.0	64.1		0.50	91.3	0.0	1.7	7.0
	0.80	54.5	10.7	0.0	34.8		0.80	99.9	0.0	0.1	0.0
50	0.00	5.1	0.3	0.0	94.6	300	0.00	2.8	0.0	2.9	94.3
	0.05	6.5	0.2	0.0	93.3		0.05	5.5	0.0	4.2	90.3
	0.10	7.4	0.2	0.0	92.4		0.10	13.6	0.0	8.9	77.5
	0.20	18.8	0.4	0.0	80.8		0.20	57.5	0.0	12.6	29.9
	0.50	71.2	1.2	0.0	27.6		0.50	100.0	0.0	0.0	0.0
	0.80	97.1	0.0	0.0	2.9		0.80	100.0	0.0	0.0	0.0

Tabella 1. Percentuale di *p-values* e probabilità a posteriori che cadono in specifici intervalli in funzione della dimensione campionaria (n) e della dimensione dell'effetto (d). $p-$ indica $p < 0.05$, $p+$ indica $p \geq 0.05$. pH_0^- indica una probabilità a posteriori minore a 0.5 (i.e. H_1 è più evidente di H_0), pH_0^+ indica una probabilità a posteriori maggiore o uguale a 0.5 (i.e. H_0 è più evidente di H_1).

Tralasciando il caso di $n = 50$ in cui osserviamo pochissimi casi discordanti, vediamo che nei gruppi con dimensioni più ampie si invertono i risultati relativi ai casi discordanti; nel senso che non si osservano mai casi in cui il *p-value* non è significativo ma H_0 è meno evidente di H_1. Sia per $n = 100$ che per $n = 300$, all'aumentare di d tra 0 e 0.2 rileviamo una diminuzione dei casi con un *p-value* significativo e H_0 più evidente di H_1. Per dimensioni dell'effetto medio-grandi ($d = 0.5, 0.8$) i due metodi tendono a convergere verso le stesse conclusioni.

4.3 Controllo dell'errore di I tipo e potenza

Nella tabella 2 abbiamo voluto mettere a confronto i tre test considerati (t test tradizionale, t test bayesiano con distribuzione a priori di Cauchy, e con distribuzione a priori normale) utilizzando uno schema ben noto nell'approccio NHST, ovvero la comparazione in termini di controllo dell'errore di I tipo e potenza del test.

In tabella 2 sono riportate le percentuali di volte in cui è stata rigettata H_0 sulla base rispettivamente di p-value, probabilità a posteriori con distribuzione a priori Cauchy (p_C) e con distribuzione a priori normale (p_N) in funzione della dimensione dell'effetto (d) e della numerosità campionaria (n).

La prima riga della tabella, in cui l'ipotesi H_0 è vera, ci permette di rilevare una prima interessante differenza. NHST risulta indipendente dalla numerosità del campione, infatti la percentuale di casi rigettati (per errore) rimane fissa intorno al 5% per tutte le dimensioni. Con i metodi Bayesiani invece osserviamo che un aumento della numerosità campionaria induce una diminuzione nella quantità di decisioni errate. Nel caso di effetti medio-grandi ($d = 0.5, 0.8$) i due approcci tendono

	p-value				p_C				p_N			
d	15	50	100	300	15	50	100	300	15	50	100	300
0	5.2	5.1	4.5	5.7	8.9	5.4	3.3	2.8	11.9	7.6	4.2	2.8
0.05	4.4	6.5	5.2	9.7	9.3	6.7	3.7	5.5	14.2	8.5	4.6	5.8
0.1	5.0	7.4	10.8	22.5	9.7	7.6	8.5	13.6	14.1	10.1	10.7	14.7
0.2	8.1	18.8	31.3	70.1	13.4	19.2	27.4	57.5	17.1	21.2	29.7	59.7
0.5	24.5	71.2	93.0	100.0	35.9	72.4	91.3	100.0	41.7	76.6	92.6	100.0
0.8	54.5	97.1	100.0	100.0	65.2	97.1	99.9	100.0	72.2	97.7	100.0	100.0

Tabella 2. Percentuale di volte in cui è stata rigettata H_0 sulla base rispettivamente di p-value, probabilità a posteriori con distribuzione a priori Cauchy (p_C) e con distribuzione a priori normale (p_N) in funzione della dimensione dell'effetto (d) e della numerosità campionaria (n).

a convergere verso gli stessi risultati con un livello di potenza lievemente superiore per i metodi Bayesiani, chiaramente evidente nel campione più piccolo. Per gli effetti piccoli ($d = 0.05, 0.1, 0.2$) vediamo un'altra interessante diversità di risultati. NHST infatti tende ad aumentare notevolmente la potenza del test con l'aumento della numerosità campionaria. Se da un certo punto di vista questo risultato è corretto (se H_0 è falsa ci si deve aspettare che il test porti a rigettarla), dall'altro appare evidente che tende a sovrastimare l'evidenza contro H_0. Da questo punto di vista l'approccio Bayesiano sembra più prudente (Weetzel, Matzke, Lee, Rouder, Iverson, Wagenmakers, 2011), ricordiamo infatti che tale approccio non porta a decisioni rigidamente dicotomiche (come invece NHST) ma piuttosto a valutazioni comparative tra ipotesi (Dienes, 2011).

4.4 Relazione funzionale tra *p-value* e probabilità a posteriori

Un importante risultato confermato dal nostro esperimento consiste nel fatto che la relazione tra il *p-value* e $P(H_0|x)$ può essere definita da una precisa relazione funzionale. In pratica questo vuol dire che, dato un valore di p ottenuto con un test t (per campioni indipendenti) è possibile ottenere facilmente, senza bisogno di una vera e propria analisi Bayesiana, il corrispondente valore di probabilità a posteriori ed il relativo BF.

Una buona approssimazione analitica della relazione funzionale può essere espressa con la seguente formula (Berger e Sellke, 1987):

$$P(H_0|x) = \frac{1}{1 + \frac{1}{\sqrt{1+n}} e^{\frac{[\Phi^{-1}(p)]^2}{2(1+\frac{1}{n})}}} \tag{3}$$

in cui n è la numerosità di ciascun campione, e $\Phi^{-1}(p)$ rappresenta l'inversa della funzione di ripartizione della normale $\Phi(z)$.

In figura 3 abbiamo rappresentato le curve definite dalla equazione 3 in funzione del *p-value* (in ascissa) e delle quattro numerosità campionarie considerate (linee nere). In grigio abbiamo raffigurato le coppie di valori (p e $P(H_0|x)$) ottenute dalla simulazione nella condizione di distribuzione a priori Cauchy. Emerge chiaramente che i valori dati dalla equazione 3 approssimano molto bene i risultati empirici.

5 CONCLUSIONI

In questo lavoro abbiamo messo a confronto il classico *p-value* dell'approccio NHST con il BF e la probabilità a posteriori di derivazione Bayesiana in relazione al test t per campioni indipendenti. Sebbene i risultati presentati siano ovviamente circoscritti a tale condizione ci sembra di poter comunque trarre dal presente confronto alcuni interessanti spunti di riflessione.

In primis, sotto specifiche condizioni i due approcci portano agli stessi risultati (a tale proposito si puó vedere anche Kruschke, 2011). Il caso evidente è quando $n = 50$ quasi ad indicare in questo valore una sorta di numerosità ideale. In questa condizione abbiamo infatti rilevato che solo 23 volte su 6000 casi (ossia circa lo 0.4%) i due approcci hanno portato a conclusioni differenti.

La relazione tra il *p-value* e BF (e la probabilità a posteriori) da un punto di vista matematico risulta talmente stretta da poter essere approssimata con una precisa relazione funzionale (vedi la (3) e fig. 3). Questa similitudine però, non implica che NHST permetta di arrivare alle stesse conclusioni dell'approccio Bayesiano (Kruschke, 2011).

Come già detto, NHST non consente di verificare alcuna ipotesi, né H_0 né l'ipotesi alternativa, e di conseguenza neppure quantificare l'evidenza dei risultati. Le conclusioni che si ottengono con questo approccio sono qualitative e non quantitative (Cohen, 1988; Ziliak e McCloskey, 2008). Anche l'*American Psychological Association* ha riconosciuto questo limite quando nel 2001 (APA, 2001) ha esortato gli autori a riportare sui lavori indicazioni relative alla dimensione degli effetti, sebbene ancora oggi tali indici vengano poco utilizzati o, al più riportati senza alcuna spiegazione

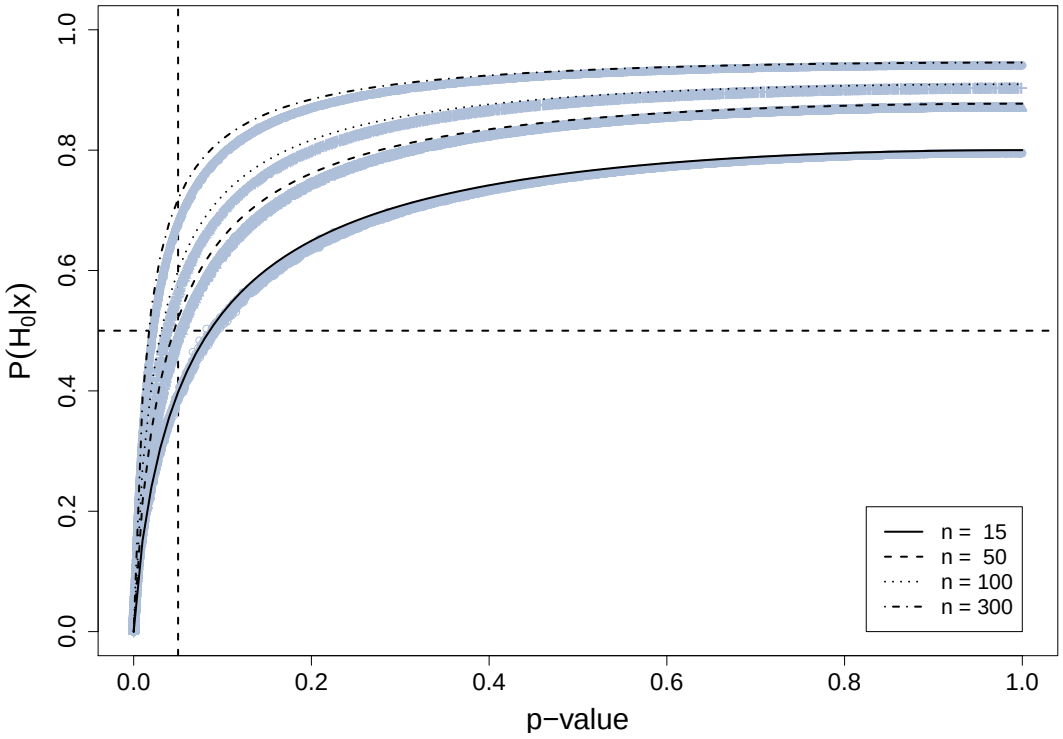

Figura 3. Relazione tra *p-value* e probabilità a posteriori. I valori grigi sono relativi ai dati ottenuti nella simulazione Monte Carlo nella condizione di distribuzione a priori Cauchy. Le linee sono state ottenute con la (3).

(Cumming, Fidler, Leonard, Kalinowski, Christiansen, Kleinig, Lo, McMenamin e Wilson, 2007; Fidler, Cumming, Thomason, Pannuzzo, Smith, Fyffe, Edmonds, Harrington e Schmitt, 2005; Finch, Cumming, Williams, Palmer, Griffith, Alders, Anderson, e Goodman, 2004).

NHST si limita a focalizzare la sua attenzione su un unico valore θ del parametro legato ad H_0, ossia quando tale ipotesi risulta vera o, in altri termini, quando la dimensione dell'effetto è nulla ($d = 0$)[3]. In questa specifica condizione la probabilità di errore (fissa a 0.05) risulta indipendente dalla numerosità campionaria. Particolarmente critico però ci sembra infatti il caso in cui la dimensione dell'effetto risulta bassa o quasi nulla: in queste situazioni infatti NHST tende a sovrastimare l'evidenza contro H_0 tanto più quanto si aumenta la dimensione campionaria (si veda tab. 2). E dato che, nei contesti empirici, H_0 non è mai esattamente vera, ma al più quasi vera (Wagenmakers, 2007) basterebbe avere un campione sufficientemente numeroso per rigettarla sempre.

Nel paradigma Bayesiano invece possiamo distinguere due approcci (Kruschke, 2011): uno chiamato *Bayesian model comparisons* ed uno chiamato *Bayesian parameter estimation*. Nel primo caso si mettono a confronto modelli alternativi e si cerca di valutare quale possa essere il più plausibile sulla base dei dati osservati. Per esempio un modello potrebbe essere quello in cui è vera H_0 (nel caso da noi considerato queso equivale a dire che $\delta = 0$) mentre quello alternativo contempla tutti i valori per cui H_0 è falsa (ovvero $\delta \neq 0$). Nel secondo caso viene definito un insieme di valori possibili del parametro oggetto di studio e si cerca di stimare la credibilità relativa di tali valori. Per esempio, sempre in riferimento al caso del test t, piuttosto che focalizzarsi su un unico valore del parametro δ, definire un insieme di valori tra loro equivalenti nella pratica (*Region of practical equivalence*, ROPE; Kruschke, 2011), e per cui H_0 risulta sostanzialmente vera. In quest'ottica si cerca di individuare una soglia nella dimensione dell'effetto al di sotto della quale lo stesso si considera irrilevante (si veda ad esempio la definizione di Chambless e Hollon, 1998, a

[3] Nel caso esaminato del test t questo equivale a dire $\mu_1 - \mu_2 = \mu = \theta = 0$.

proposito della significatività clinica). Inoltre, i confini della ROPE possono essere definiti volta per volta senza dover ricorrere ad un unico criterio universale come invece fa NHST rispetto alla soglia critica α. In questo modo l'approccio risulta essere più coerente in quanto l'aumento della numerosità campionaria risulta associato all'aumento dell'evidenza a favore dell'ipotesi vera.

Anche dal punto di vista decisionale le differenze sono molto marcate: mentre NHST si basa su conclusioni dicotomiche e qualitative, l'approccio Bayesiano si mantiene sempre su una logica comparativa che ha l'obiettivo di favorire l'ipotesi più probabile. Va detto che sono state proposte delle soglie per l'interpretazione del BF (Kass e Raftery, 1995; Raftery, 1995) ma ci sembra che queste soglie rispondano più ad una logica NHST che a quella Bayesiana. Nello specifico, il metodo chiamato *Bayesian model comparison* ha come obiettivo la valutazione di quale sia il più credibile tra vari modelli alternativi, mentre il metodo chiamato *Bayesian parameter estimation* ha come obiettivo la stima della relativa credibilità di un insieme di valori di uno o più parametri (Kruschke, 2011).

Una ulteriore lettura dei risultati del nostro esperimento è che, in pratica, BF e/o la probabilità a posteriori permettono di sintetizzare in un unico valore di evidenza anche la dimensione dell'effetto superando il limite intrinseco di NHST. Dalla lettura dei risultati della nostra simulazione (vedi tabelle 1 e 2) emerge chiaramente come il *p-value* risulti sensibile alla dimensione dell'effetto ma sappiamo che esso non contiene informazioni in proposito. Nello specifico, se avessimo a disposizione solo il *p-value* senza avere altre informazioni (per esempio sulla numerosità campionaria) sarebbe difficile prendere una posizione sull'evidenza delle ipotesi in gioco. Viceversa, sempre sulla base dei risultati della nostra simulazione, possiamo notare come il BF risulti, da questo punto di vista, più coerente e maggiormente in grado di supportare l'evidenza delle stesse.

Nella formula 3 possiamo osservare che nella relazione tra il *p-value* e $P(H_0|x)$ entra in gioco solo la numerosità campionaria (n) e non la dimensione dell'effetto. In altri termini, osservando la figura 1 non riusciamo a distinguere tra loro le curve relative ai diversi valori di d. Lo stesso ragionamento è valido nella relazione tra il *p-value* e BF dato che potremmo riscrivere la 3 inserendo quest'ultimo al posto di $P(H_0|x)$.

In sintesi, ci sembra di poter concludere che l'utilizzo del BF possa essere utile, in aggiunta ai classici *p-value* e dimensione dell'effetto per avere un quadro più completo dei dati analizzati e, soprattutto, per giungere a decisioni migliori. Sebbene nel caso da noi considerato sia estremamente semplice ottenere il BF (e da questo la probabilità a posteriori) a partire dal *p-value*, vi sono esempi di come ottenere le stesse informazioni a partire dall'ANOVA (Wagenmakers, 2007; Masson, 2011) senza bisogno di sofisticate elaborazioni. In più, bisogna considerare che vi sono casi caratterizzati da elevata complessità (ad esempio quando vi sono molti parametri o funzioni di densità non definite completamente per via analitica) in cui NHST non è utilizzabile in quanto non risulta nota la distribuzione dei parametri sotto H_0. Questi casi sono invece affrontabili con metodi Bayesiani dato che è possibile definire distribuzioni a priori non informative ed ottenere delle distribuzioni a posteriori che vengono aggiornate sulla base dei dati osservati. A questo si aggiunga che negli ultimi tempi anche i vari software statistici si stanno notevolmente arricchendo della possibilità di utilizzare tali modelli di analisi (si veda in tal senso l'aumento negli ultimi anni del numero di librerie dedicate ad essi in ambiente R; R Development Core Team, 2010).

Riferimenti bibliografici

AGNOLI, F., FURLAN, S. (2008). La differenza che fa la differenza: dalla significatività statistica alla significatività pratica. *Psicologia Clinica dello Sviluppo*, 12, 211-245.

AGNOLI, F., FURLAN, S. (2009). I cambiamenti nella verifica di ipotesi: Statistiche migliori per decisioni migliori. *Giornale italiano di Psicologia*, 36, 849-882.

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

BAKKER, M., WICHERTS, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods* , 43, 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, 883-898.

BERGER, J. O., SELLKE, T. (1987). Testing a Point Null Hypothesis: The Irreconciliability of P Values and Evidence. *Journal of the American Statistical Association*, 82, 112-122.

CHAMBLESS, D. L., HOLLON, S. D. (1998). Defining Empirically Supported Therapies. *Journal of Consulting and Clinical Psychology*, 66, 7-18.

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

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

CUMMING, G., FIDLER, F., LEONARD, M., KALINOWSKI, P., CHRISTIANSEN, A., KLEINIG, A., LO, J., MCMENAMIN, N., WILSON, S. (2007). Statistical Reform in Psychology. Is Anything Changing? *Psychological Science*, 18, 230-232.

DI NUOVO, S. (2009). Cosa significa significativo? Ovvero: ci sono alternative all'ipotesi alternativa? *Giornale italiano di Psicologia*, 36, 899-911.

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

DIXON, P. (2003). The p value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology*, 57, 189-202.

FIDLER, F., CUMMING, G., THOMASON, N., PANNUZZO, D., SMITH, J., FYFFE, P., EDMONDS, H., HARRINGTON, C., SCHMITT, R. (2005). Toward Improved Statistical Reporting in the Journal of Consulting and Clinical Psychology. *Journal of Consulting and Clinical Psychology*, 73, 136-143.

FINCH, S., CUMMING, G., WILLIAMS, J., PALMER, L., GRIFFITH, E., ALDERS, C., ANDERSON, J., GOODMAN, O. (2004). Reform of statistical inference in psychology: The case of Memory & Cognition. *Behavior Research Methods, Instruments, & Computers*, 36, 312-324.

FRICK, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, 1, 379-390.

GAMERMAN, D., LOPES, H. F. (2006). *Markov chain Monte Carlo: Stocastic simulation for Bayesian inference*. Boca Raton, FL: Chapman & Hall/CRC.

GALLISTEL, C. R. (2009). The Importance of Proving the Null. *Psichological Review*, 116, 439-453.

HAGEN, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, 52, 15-24.

JEFFREYS, H. (1961). *Theory of Probability*. New York: Oxford University Press.

KASS, R. E., RAFTERY, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90, 773-795.

KILLEEN, P. (2005). An alternative to null-hypothesis significance testing. *Psychological Science*, 16, 345-353.

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

KRUSCHKE, J. K. (2011). Bayesian Assesment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives in Psychological Science*, 6, 299-312.

IVERSON, G. J., LEE, M. D., ZHANG, S., WAGENMAKERS, E. J. (2009). p_{rep}: An agony in five Fits. *Journal of Mathematical Psychology*, 53, 195.202.

LUNN, D. J., THOMAS, A., BEST, N., SPIEGELHALTER, D. (2000). WinBUGS - a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics & Computing*, 10, 325-337.

MEEHL, P. E. (1997). The Problem Is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What If There Were No Significance Tests?* (pp. 395-425) Mahwah, NJ : Erlbaum, 1997.

NICKERSON, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241-301.

MASSON, M. E. J. (2011). A tutorial on a practical bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, 43, 679-690.

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

RAFTERY, A. E. (1995). Bayesian model selection in social research. In P.V. Marsden (ED.), Sociological Methodology, 1995 (pp. 111-196),

R DEVELOPMENT CORE TEAM (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

ROBERT, C. P. (2001). *The Bayesian Choice (Second Edition)*. New York: Springer.

ROUDER, J. N., MOREY, R. D. (2005). Relational and arenational confidence intervals: A comment on Fidelr et al. (2004). *Psychological Science*, 16, 77-79.

ROUDER, J. N., SPECKMAN, P. L., SUN, D., MOREY, R. D., IVERSON, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16, 225-237.

SCHMIDT, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1, 115-129.

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

WAGENMAKERS, E. J., GRUNWALD, P. (2006). A Bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychological Science*, 17, 641-642.

WEETZEL, R., MATZKE, D., LEE, M. D., ROUDER, J. N., IVERSON, G. J., WAGENMAKERS E. J. (2011). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. *Perspectives on Psychological Science*, 6, 291-298.

WETZELS, R., RAAIJMAKERS, J. G. W., JAKAB, E., WAGENMAKERS, E. J. (2009). How to Quantify Support For and Against the Null Hypothesis: A Flexible WinBUGS Implementation of a Default Bayesian t-test. *Psychonomic Bulletin & Review*, 16, 752-760.

WILKINSON, L., & THE TASK FORCE ON STATISTICAL INFERENCE (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

ZILIAK, S. T., MCCLOSKEY, D. N. (2008). *The Cult of Statistical Significance*. University of Michigan Press.