I LIMITI DELL'APPROCCIO NHST E L'ALTERNATIVA BAYESIANA

MASSIMILIANO PASTORE

Università di Padova

Riassunto. Nella ricerca psicologica l'approccio NHST è la modalità più frequente per l'inferenza statistica. Di fatto tale metodo si riduce ad una decisione sulla significatività dei risultati basata solo sulla valutazione del *p-value*. In realtà il *p-value* non è una misura dell'evidenza statistica perché influenzato da elementi esterni quali la numerosità campionaria, dati che non vengono osservati e le intenzioni soggettive del ricercatore (Wagenmakers, 2007). In questo lavoro sono presentate alcune esemplificazioni che illustrano come interpretare il *p-value* ed i problemi ad esso connessi. Una soluzione alternativa viene proposta con l'approccio Bayesiano, presentando un'applicazione della modalità per il confronto tra ipotesi nel caso del *t-test*. Utilizzando il paradigma Bayesiano diventa possibile valutare con una probabilità l'evidenza dell'ipotesi nulla in relazione a possibili ipotesi alternative.

1. INTRODUZIONE

La sigla NHST (*Null Hypothesis Significance Testing*; Cohen, 1994) indica la modalità più frequente di condurre i test statistici nell'ambito della ricerca psicologica. Tale approccio deriva da una sorta di ibrido tra il metodo di Fisher (1935) e quello di Neyman-Pearson (1933). In sintesi il procedimento può essere 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*).

Sulla base di tale probabilità i ricercatori concludono se il risultato ottenuto sia o meno statisticamente significativo.

Il dibattito tra sostenitori e detrattori del metodo NHST è molto acceso da tempo nella letteratura psicologica (si veda ad esempio: Cohen, 1994; Cortina e Dunlap, 1997; Dixon, 2003; Frick, 1996; Gigerenzer, 1998; Hagen, 1997; Killeen, 2006; Loftus, 1996; Nickerson, 2000; Schmidt, 1996; Wainer, 1999; Wagenmakers, 2007). Oggi sono riconosciuti i limiti strutturali di questo approccio che si è ridotto ad una mera applicazione meccanica senza una accurata riflessione sul

senso dei risultati ottenuti (Ziliak e McCloskey, 2008). In particolare sono stati individuati i seguenti aspetti critici: [1] NHST tende a indurre confusione tra la probabilità dell'ipotesi condizionata ai dati (probabilità a posteriori) e probabilità dei dati condizionati all'ipotesi (verosimiglianza) (Cohen, 1994; Wagenmakers, 2007); [2] NHST viene erroneamente considerato un metodo per la verifica delle ipotesi. In realtà esso tiene conto solo di H_0 e permette solo la falsificazione di tale ipotesi senza che questo abbia relazione con la veridicità di H_1. A questo proposito, già nel 1999 l'APA Task Force on Statistical Inference esortava a non utilizzare l'espressione «*accept the null hypothesis*» (Wilkinson e la Task Force on Statistical Inference, 1999, p. 599); [3] il criterio $\alpha = 0.05$ è puramente arbitrario, lo stesso Fisher (1959) scriveva «*... no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he reject hypotheses ...*» (p. 42); [4] i test tradizionali tendono a sovrastimare l'evidenza contro H_0: infatti H_0 nei contesti reali non è mai esattamente vera e pertanto aumentando a dovere il numero di osservazioni è quasi sempre possibile rigettarla (Wagenmakers, 2007; Rouder, Speckman, Sun, Morey e Iverson, 2009); [5] l'ipotesi nulla legata ad un unico valore puntuale (H_0:$\theta = \theta_0$) senza opportuni accorgimenti porta a conclusioni distorte (Berger e Sellke, 1987; Sellke, Bayarri e Berger, 2001).

Nella sostanza, l'uso dell'approccio NHST si riduce alla mera considerazione del *p-value*, utilizzato spesso in maniera impropria.

Gli obiettivi del presente lavoro sono: 1) illustrare il significato del *p-value* ed i limiti ad esso connessi in relazione al suo utilizzo nell'approccio NHST; 2) presentare, con un esempio specifico legato al *t*-test, la soluzione proposta dall'alternativa Bayesiana ai test di ipotesi.

2. IL *P-VALUE*

Il *p-value* è una probabilità condizionata; geometricamente si può rappresentare come un'area definita da un valore calcolato t_0 di una certa statistica t e da una funzione di densità condizionata della stessa statistica $f(t|H_0)$. Immaginiamo di avere un campione di osservazioni sul quale abbiamo effettuato un test ad una coda ottenendo una statistica $t_0 = 1.96$, formalmente possiamo scrivere:

$$p\text{-}value = Prob(t \geq 1.96|H_0) = \int_{1.96}^{+\infty} f(t|H_0)dt$$

ovverosia, la probabilità di ottenere un valore di t maggiore o uguale a 1.96 condizionata al fatto che l'ipotesi H_0 sia vera. Geometricamente

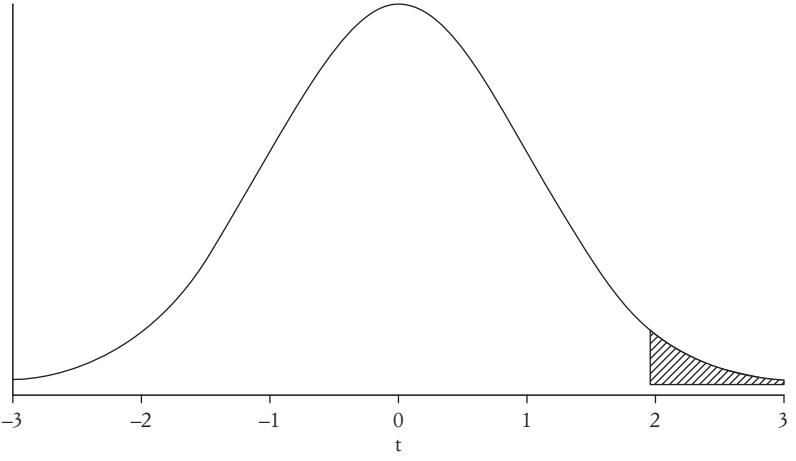

FIG. 1. Rappresentazione del *p-value* in un test ad una coda. La curva rappresenta la funzione di densità della statistica test t in H_0. L'area tratteggiata in nero è la probabilità di ottenere un valore di t maggiore o uguale a 1.96.

si tratta dell'area sottostante la curva di densità compresa tra $t_0 = 1.96$ e $+\infty$ (vedi fig. 1).

Se indichiamo con R il risultato di un esperimento, possiamo dire che i test statistici nella forma NHST stimano $Prob(R|H_0)$ ossia la probabilità di osservare il risultato R (o uno più estremo) se è vera H_0. Tale probabilità è diversa da $Prob(H_0|R)$ ossia la probabilità che sia vera l'ipotesi H_0 condizionata al risultato osservato R. In sostanza: $Prob(R|H_0)$ è il famigerato *p-value* mentre $Prob(H_0|R)$ è la probabilità a posteriori.

La relazione tra le due è definita dal teorema di Bayes:

$$P(H_0|R) = \frac{P(H_0)P(R|H_0)}{P(H_0)P(R|H_0) + P(H_1)P(R|H_1)}$$

in cui: $P(H_i)$ è la probabilità a priori di $H_i (i = 0,1)$, $P(R|H_i)$ è la probabilità del risultato R condizionato ad H_i, $P(H_0|R)$ è la probabilità a posteriori di H_0.

La confusione nell'interpretazione del *p-value* è spesso riscontrabile anche nella letteratura. Basta sfogliare gli articoli pubblicati per trovare scritture di questo tipo:

$p < 0.035$ (se p intende essere il *p-value* ottenuto dal software la scrittura corretta è $p = 0.035$), $p < 0.000$ (una probabilità non può essere minore di zero, più verosimilmente il software omette le cifre

TAB. 1. *Tabelle di contingenza fittizie. Tra parentesi sono riportate le frequenze percentuali, calcolate sul totale*

	[A]	
	a_1 a_2	totale
b_1	5(25) 4(20)	9(45)
b_2	2(10) 9(45)	11(55)
totale	7(35) 13(65)	20(100)

	[B]	
	a_1 a_2	totale
b_1	10(25) 8(20)	18(45)
b_2	4(10) 18(45)	22(55)
totale	14(35) 26(65)	40(100)

decimali dopo la terza), $p<0.0001$ (se il software statistico omette le cifre decimali dopo la terza significa che possiamo scrivere solo $p<0.001$ in quanto non conosciamo la quarta cifra decimale).

3. PROBLEMI LEGATI AL *P-VALUE*

In un recente lavoro Wagenmakers (2007) ha illustrato dettagliatamente quali possano essere i problemi legati al *p-value* individuandone tre in particolare: [1] p dipende dalla numerosità del campione; [2] p dipende da dati che non vengono mai osservati; [3] p dipende dalle intenzioni soggettive e spesso sconosciute del ricercatore.

Vediamo dei semplici esempi per illustrare questi tre problemi (per maggiori dettagli si veda Wagenmakers, 2007).

3.1. *Il p-value dipende dalla numerosità campionaria*

Consideriamo la tabella di contingenza 1[A] in cui sono riportate le frequenze rilevate in due variabili qualitative A e B. Immaginiamo di essere interessati a valutare se esista un'associazione statisticamente significativa tra A e B, il metodo più semplice consiste nell'eseguire un test χ^2. Si può facilmente verificare che in questa tabella il valore $\chi^2_{(1)} = 3.0392$ ha una probabilità associata di $p = 0.0813$ che ci porterebbe a propendere per la non associazione.

Proviamo però a raddoppiare le frequenze ottenendo la tabella 1[B]. Dalla lettura delle percentuali, riportate tra parentesi, risulta

TAB. 2. *Esempi di due differenti distribuzioni di probabilità f(x) e g(x)*

x	1	2	3	4	5	6	7	8	9	10
$f(x) \mid H_0$	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
$g(x) \mid H_0$	0.20	0.28	0.17	0.11	0.08	0.10	0.02	0.01	0.01	0.02

evidente che il rapporto di associazione tra le variabili non viene alterato. Ripetendo il test otteniamo però $\chi^2_{(1)} = 6.0784$, $p = 0.0137$ che ci porterebbe a rigettare l'ipotesi H_0 in favore di una associazione statisticamente significativa.

È facile dimostrare che tale effetto legato alla numerosità campionaria si può riscontrare in tutti i test statistici più noti: aumentando la numerosità campionaria il valore di p tende a diminuire.

3.2. *Il p-value dipende da dati che non vengono effettivamente osservati*

Consideriamo un'urna contenente N palle numerate da 1 a 10. Sia x la variabile casuale che esprime il valore di una palla estratta a caso la cui distribuzione di probabilità $f(x)$ è riportata in tabella 2 (prima riga). Per semplicità la nostra statistica sia $t(x) = x$.

Immaginiamo di avere estratto una palla con il numero 6 e di voler calcolare la probabilità di estrarre a caso una palla con un valore maggiore o uguale a 6. In base alla distribuzione definita $f(x)$ tale probabilità sarà data da $\sum_{x=6}^{10} f(x) = 0.5$.

Appare evidente che tale probabilità è condizionata al fatto che la vera distribuzione di probabilità sia $f(x)$. Se la distribuzione campionaria vera fosse $g(x)$ (riportata in tab. 2, seconda riga) si arriverebbe ad un risultato differente, infatti: $\sum_{x=6}^{10} g(x) = 0.16$.

Nella sostanza, pur non avendo osservato alcun valore superiore a 6, questi dati influenzano il valore di probabilità relativo alla statistica t. L'unico dato osservato è $x = 6$, che ha la stessa probabilità nelle due distribuzioni f e g.

Lo stesso processo avviene nella ricerca quando si esegue un test statistico su un campione di dati: il risultato ottenuto è condizionato alla distribuzione delle probabilità di riferimento (generalmente la distribuzione relativa ad H_0), se la vera distribuzione è diversa il risultato potrebbe cambiare.

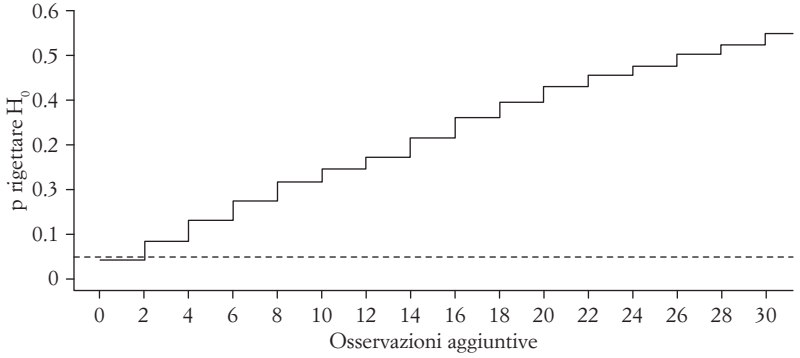

FIG. 2. Funzione cumulata della probabilità stimata di rigettare H_0 erroneamente in funzione del numero di osservazioni campionarie aggiunte.

3.3. *Il p-value dipende dalle intenzioni soggettive*

Immaginiamo un ricercatore interessato a studiare un certo fenomeno. Egli predispone un esperimento e lo esegue su 5 soggetti ottenendo un *p-value* pari a 0.09. Convinto della bontà della sua ipotesi alternativa, ritenendo che, con soli 5 soggetti, il test non sia sufficientemente potente, decide di continuare a reclutare soggetti, eseguendo ogni volta che ne aggiunge uno il test statistico. Anche se tale procedura può sembrare neutrale, accade che la probabilità di trovare un risultato statisticamente significativo aumenta, anche se H_0 è vera.

Proviamo a simulare questo processo con un semplice esperimento Monte Carlo. Ripetiamo per 1000 volte la seguente procedura:

– estraiamo due campioni casuali di $n=5$ osservazioni dalla stessa popolazione normale con media $\mu_1=\mu_2=0$;
– confrontiamo le medie dei due campioni con un *t*-test;
– se il risultato è statisticamente significativo interrompiamo il processo, altrimenti aggiungiamo un'osservazione per ciascuno dei due gruppi campionando sempre dalla stessa popolazione (con $\mu=0$) fino ad un massimo di 30 osservazioni aggiunte (15 per ogni campione).

Al termine della procedura contiamo quante volte il processo si è interrotto entro le 30 osservazioni.

In figura 2 è riportata la distribuzione cumulata della probabilità di arrestare il processo in quanto si è ottenuta una differenza statisticamente significativa tra le medie. Osserviamo che circa il 5% delle volte il processo si è arrestato subito (0 osservazioni aggiunte). Questo

dato è perfettamente coerente con la probabilità che viene assegnata all'errore di I tipo. Poiché i campioni sono estratti da popolazioni con la stessa media, quindi con H_0: $\mu_1 = \mu_2$ vera, ci aspettiamo di ottenere dei risultati significativi (per errore) il 5% delle volte.

Il problema sta nell'incremento di questa probabilità di errore in funzione dell'aggiunta di osservazioni. Dopo aver aggiunto solo sei osservazioni (tre per campione) la probabilità di rigettare H_0 è raddoppiata. In più, solo circa la metà delle volte (con esattezza 462 volte su 1000) abbiamo campionato fino a 30 osservazioni senza trovare un risultato significativo.

Nella sostanza, effettuare una procedura di campionamento soggetti non definita a priori, ma basata ad esempio solo su un criterio di arresto legato alla significatività statistica, rischia di avere un effetto dirompente sul *p-value*.

4. IL PARADOSSO DEL T-TEST

Un ulteriore punto critico nell'approccio NHST viene messo in evidenza da Rouder *et al.* (2009) in relazione al *t*-test.

Consideriamo il seguente esperimento: vengono presentate in sequenza delle coppie di stimoli raffiguranti figure geometriche tridimensionali. Nella metà delle coppie l'immagine è presentata in negativo (ossia con polarità invertita). Le coppie di immagini possono essere perfettamente identiche oppure in una delle due vi può essere un particolare modificato rispetto all'altra. Il compito dei soggetti consiste nell'individuare quando ci sia stato il cambiamento.

Dopo aver effettuato l'esperimento su un campione di 66 soggetti, ci chiediamo se i tempi di reazione (TR) risultino diversi nelle due condizioni di presentazione degli stimoli (immagine positiva *vs*. negativa). La differenza tra le medie nelle due condizioni è di 13.63 ms, da cui si ottiene una statistica $t_{(65)} = 2.05$ con $p = 0.04$. Sulla base di questo risultato saremmo portati a rigettare H_0 propendendo per una differenza statisticamente significativa tra i tempi di reazione nelle due condizioni.

Un modo per quantificare l'evidenza del risultato ottenuto consiste nel calcolare la verosimiglianza del valore ottenuto sotto varie ipotesi alternative. Nel caso di H_0 la verosimiglianza di $t = 2.05$ è data dalla densità della distribuzione t con 65 gradi di libertà e cioè circa 0.05. Supponiamo ora che nel tipo di esperimento svolto la reale differenza tra le medie sia di 30 ms. La verosimiglianza del valore osservato è data allora dalla densità t con 65 gradi di libertà e parametro di noncentralità $\delta_\mu \sqrt{n}/\sigma$, cioè circa 0.02. Il rapporto di verosimiglianza $0.05/0.02 = 2.5$ indica che H_0 è circa 2.5 volte più verosimile di H_1 nonostante il test sia risultato significativo e ci porti a rigettare H_0.

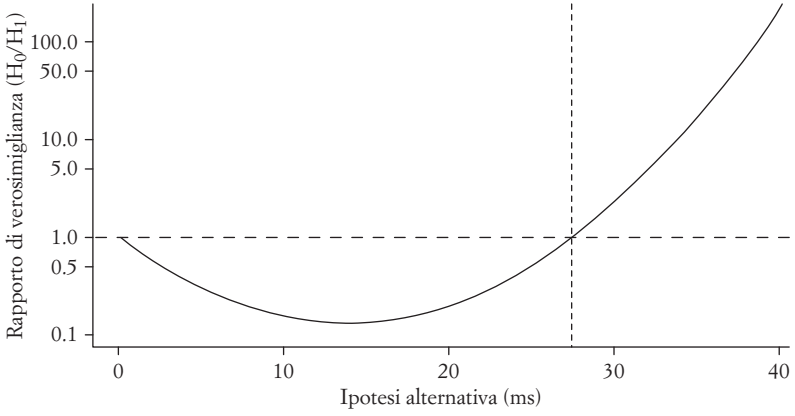

FIG. 3. Rapporto di verosimiglianza (H_0/H_1) in funzione dei valori dell'ipotesi alternativa relativo a t=2.05 ottenuto su un campione di 66 soggetti. In ordinata è stata utilizzata la scala logaritmica. La linea orizzontale tratteggiata nera indica la soglia 1, valori al di sotto di tale soglia indicano che H_1 ha una verosimiglianza superiore ad H_0, valori al di sopra indicano il viceversa.

In figura 3 abbiamo rappresentato il rapporto di verosimiglianza per valori dell'ipotesi alternativa fino a 40 ms. Quando il rapporto è inferiore ad 1 vuol dire che H_1 ha una verosimiglianza superiore ad H_0, viceversa H_0 ha un valore di verosimiglianza maggiore di H_1 quando il rapporto supera 1. Paradossalmente, si osserva che H_1 risulta essere più verosimile di H_0 per differenze inferiori a circa 27.5 ms. Per differenze maggiori accade il contrario.

5. TEST DI IPOTESI IN FORMA BAYESANA

Una delle caratteristiche più importanti nel test di ipotesi in forma Bayesiana è la sua natura comparativa: non è sufficiente che l'ipotesi nulla sia poco probabile per rigettarla in quanto i dati potrebbero essere ancora più improbabili sotto l'ipotesi alternativa.

Un metodo pratico per confrontare due ipotesi è dato dal rapporto tra le probabilità a posteriori:

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

in cui x indica i dati osservati, $f(x|H_i)(i=0,1)$ la funzione di verosimiglianza, $Pr(H_i)$ la probabilità a priori e $Pr(H_i|x)$ la probabilità a poste-

riori. Il rapporto $\frac{f(x|H_0)}{f(x|H_1)}$ viene chiamato *Bayes Factor* (BF). 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. Se le probabilità a priori delle due ipotesi sono uguali, esso indica direttamente il rapporto a posteriori, per mezzo del quale potremmo concludere 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 $1 - Pr(H_0|x) = 0.25$.

5.1. *Confronto tra ipotesi*

Supponiamo di avere osservato una variabile Y con distribuzione campionaria $f(y|\theta)$ e di voler testare le seguenti ipotesi:

$$H_0: \theta \in \Theta_0, H_1: \theta \in \Theta_1$$

in cui $\Theta_0 \cup \Theta_1 = \Theta$ è lo spazio parametrico, ossia l'insieme dei possibili valori del parametro θ. Assegnando una funzione di densità a priori per il parametro $g(\theta)$, allora il rapporto tra le probabilità a priori è dato da

$$\frac{\pi_0}{\pi_1} = \frac{Pr(\theta \in \Theta_0)}{Pr(\theta \in \Theta_1)} = \frac{\int_{\Theta_0} g(\theta) d\theta}{\int_{\Theta_1} g(\theta) d\theta}$$

in cui π_i $(i = 0,1)$ indica la probabilità a priori che il vero valore del parametro θ cada nell'insieme di valori Θ_i. Dopo avere osservato un campione di dati y è possibile ottenere il rapporto tra le probabilità a posteriori

$$\frac{p_0}{p_1} = \frac{Pr(\theta \in \Theta_0 | y)}{Pr(\theta \in \Theta_1 | y)} = \frac{\int_{\Theta_0} g(\theta | y) d\theta}{\int_{\Theta_1} g(\theta | y) d\theta}$$

in cui $g(\theta|y)$ è la densità a posteriori e $p_i (i = 0,1)$ indica la probabilità che il vero valore del parametro θ cada nell'insieme di valori Θ_i dopo avere osservato y.

Il BF è dato dal rapporto tra i due rapporti di probabilità a posteriori e a priori

$$BF = \frac{p_0/p_1}{\pi_0/\pi_1}$$

Utilizzando quest'ultima relazione è possibile scrivere la probabilità a posteriori di H_0 come funzione di BF e π_0

$$p_0 = \frac{\pi_0 BF}{\pi_0 BF + 1 - \pi_0}$$

Di seguito presentiamo un'applicazione del confronto tra ipotesi nel caso del *t*-test.

5.2. *t-test Bayesiano*

Riprendiamo in esame l'esperimento descritto prima con i TR in due condizioni sperimentali (immagini positive *vs.* negative), questa volta supponiamo di avere solamente 20 soggetti. Data la natura dei dati rilevati (variabili appaiate), la variabile che prendiamo in considerazione è data dalla differenza tra le coppie di TR rilevati nelle due condizioni per ciascun soggetto ($\mu = \mu_{pos} - \mu_{neg}$). Le ipotesi che mettiamo a confronto sono:

$$H_0: \mu = 0,\ H_1: \mu \neq 0$$

In altri termini: H_0 indica che non vi sono differenze tra i TR nelle due condizioni sperimentali, H_1 invece che tali differenze ci sono.

Dato il caso particolare in cui il valore atteso in H_0 è unico, possiamo assegnare a tale ipotesi una probabilità a priori di 0.5 ($\pi_0 = 0.5$). In relazione all'ipotesi alternativa le scelte possibili sono molte (si veda Rouder *et al.*, 2009). Per semplicità consideriamo, secondo quanto suggerito da Albert (2007, p. 167), una distribuzione a priori della differenza tra le medie con forma normale, media 0 e deviazione standard τ. Per la stima di quest'ultimo parametro utilizziamo il seguente criterio: fissiamo a 4τ il 95% del *range* di una distribuzione normale. In questo modo, supponendo che il *range* della differenza tra le medie del nostro esempio si collochi tra ±30 ms, avremo che $60 = 4\tau$ e, di conseguenza, $\tau = 15$. Infine, possiamo utilizzare la seguente relazione per la stima del BF:

$$\mathrm{BF} = \frac{\frac{\sqrt{n}}{\sigma} exp\left\{-\frac{n}{2\sigma^2}(\bar{y})^2\right\}}{(\sigma^2/n + \tau^2)^{-1/2} exp\left\{-\frac{1}{2(\sigma^2/n + \tau^2)}(\bar{y})^2\right\}}$$

TAB. 3. *Valori calcolati del BF e della probabilità a posteriori di $H_0(p_0)$ in funzione di 5 valori scelti di τ*

τ	*BF*	p_0
0.5	3.89	0.80
1.0	7.76	0.89
3.0	23.25	0.96
6.0	46.50	0.98
15.0	116.25	0.99

in cui N indica la numerosità campionaria, σ^2 la varianza della popolazione da cui sono campionati i valori e $\bar{y}$ la media osservata delle differenze.

Nei dati del nostro esperimento fittizio abbiamo ottenuto una differenza tra le medie $\bar{y}=0.356$, con stima della varianza della popolazione pari a $\hat{\sigma}^2 = 0.006$, e $t_{(19)}=1.987$, $p=0.061$. Un simile risultato secondo la logica NHST ci dice poco. A rigore dovremmo concludere che non possiamo rigettare l'ipotesi di uguaglianza delle medie nelle due condizioni, senza sapere però se tale risultato implichi in qualche modo che la differenza non esista realmente oppure se, data la scarsa numerosità del campione, la differenza non emerga per una inadeguata potenza del test.

In tabella 3 abbiamo riportato i valori ottenuti del BF e della probabilità a posteriori di $H_0(p_0)$ in funzione di cinque valori scelti per il parametro τ. In generale, infatti, data la difficoltà di definire dei valori esatti di τ, conviene piuttosto considerare un insieme di valori plausibili.

Dalla lettura della tabella possiamo trarre delle informazioni in più. I valori di *BF*, tutti superiori ad 1, fanno propendere per una maggiore evidenza di H_0 rispetto alle varie ipotesi alternative definite per diversi valori di τ. In più, dalle probabilità a posteriori di H_0, possiamo concludere che tale ipotesi risulta decisamente più probabile rispetto ad H_1.

Recentemente (Rouder *et al.*, 2009; Wetzels, Raaijmakers, Jakab e Wagenmakers, 2009) è stato riconsiderato il *t*-test in forma Bayesiana proponendo forme alternative a quella illustrata. In questi lavori emerge come l'approccio Bayesiano permetta un utile contributo alla valutazione in senso positivo dell'evidenza a favore di H_0 superando la logica puramente falsificazionista.

6. CONCLUSIONI

Nella ricerca psicologica (come in molte altre discipline) l'approccio NHST è il metodo più adottato per l'inferenza statistica. Di fatto

l'unica informazione che viene tenuta in conto è quella legata al *p-value*, il cui utilizzo è sostanzialmente contrapposto nelle originarie definizioni dei test secondo Fisher e Neyman-Pearson. Il primo infatti considerava *p* come una misura diretta dell'evidenza statistica, mentre secondo Neyman-Pearson andava fissato preventivamente un valore critico della statistica test relativo ad una soglia di errore stabilita (per una descrizione estesa si veda Berger, 2003). In realtà *p-value* non è una misura dell'evidenza statistica perché, come abbiamo visto, dipende da molti altri elementi (numerosità campionaria, dati non osservati, intenzioni soggettive).

Un altro elemento di debolezza dell'approccio NHST è che tiene conto solo di H_0 e può al massimo confutarla, mai verificarla.

Idealmente una procedura statistica dovrebbe dipendere solo dai dati osservati e fornire una misura dell'evidenza che tenga conto di tutte le ipotesi in gioco (e non solo H_0). Secondo Ziliak e McCloskey (2008), ormai l'approccio ha assunto lo status di culto piuttosto che di metodo scientifico vero e proprio: «*Statistical significance is not a scientific test. It is a philosophical, qualitative test. It does not ask how much. It asks whether. Existence, the question of whether, is interesting. But it is not scientific*» (p. 4).

Al contrario, l'approccio Bayesiano permette di superare gli elementi critici illustrati. In particolare sembra convincente l'idea comparativa che ne sta alla base: le ipotesi vengono coinvolte entrambe nella procedura di calcolo ed il risultato finale che si ottiene permette una valutazione dell'evidenza statistica delle stesse. Questo permette di dare la giusta importanza anche alla verifica dell'ipotesi nulla (Gallistel, 2009). Piuttosto, il problema cruciale nel paradigma Bayesiano è legato alla scelta delle distribuzioni a priori (*priors*). Una volta che queste sono note, infatti, l'inferenza procede in modo meccanico indipendentemente dal ricercatore. È su questo aspetto che si sono focalizzate storicamente le critiche al metodo Bayesiano fin dalle sue origini (si vedano ad esempio Boole, 1854; Venn, 1866; Chrystal, 1891). Per un uso appropriato delle tecniche basate sull'approccio Bayesiano, è importante che le *priors* siano scelte con criterio o si corre il rischio di produrre risultati distorti (Rouder *et al.*, 2009), ma è altresì vero che la letteratura ha proposto molte varianti, ad esempio basate su metodi definiti robusti (Insua e Ruggeri, 2000), che garantiscono una buona affidabilità dei risultati (Robert, 2001). In più, oggi vari software permettono facilmente di utilizzare tecniche simulative (es. Monte Carlo) di forte supporto a tale approccio (si vedano ad esempio Albert, 2007; Gelman, Carlin, Stern e Rubin, 2004).

In conclusione, ci sembra che l'approccio Bayesiano fornisca un utile contributo per superare la logica qualitativa legata alla significati-

vità statistica (NHST) introducendo elementi di significatività sostanziale che permettono di quantificare i risultati ottenuti in termini di evidenza statistica.

BIBLIOGRAFIA

ALBERT J. (2007). *Bayesian computation with R*. New York: Springer.

BERGER J.O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing?. *Statistical Science*, *18*, 1-32.

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.

BOOLE G. (1854). *A investigation of the laws of thought*. London: Walton and Maberly.

CHRYSTAL G. (1891). On some fundamental principles in the theory of probability. *Transactions of the Actuarial Society of Edinburgh*, *2*, 421-439.

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

CORTINA J.M., DUNLAP W.P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161-172.

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

FISHER R.A. (1935). *The design of experiments*. Edinburgh: Oliver & Boyd.

FISHER R.A. (1959). *Statistical methods and scientific research* (2nd ed.). New York: Hafner.

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

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

GELMAN A., CARLIN J.B., STERN H.S., RUBIN D.B. (2004). *Bayesian Data Analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

GIGERENZER G. (1998). We need statistical thinking, not statistical rituals. *Behavioral & Brain Sciences*, *21*, 199-200.

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

INSUA D.R., RUGGERI F. (eds.) (2000). *Robust Bayesian Analysis*. New York: Springer.

KILLEEN P.R. (2006). Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review*, *13*, 549-562.

LOFTUS G.R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161-171.

NEYMAN J., PEARSON E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transaction of the Royal Society: Series A*, 231, 289-337.

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

ROBERT C.P. (2001). *The Bayesian Choice* (2nd ed.). New York: Springer.

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.
SELLKE T., BAYARRI M.J., BERGER J.O. (2001). Calibration of p values for Testing Precise Null Hypotheses. *American Statistician*, *55*, 62-71.
VENN J. (1866). The logic of chance. London: Macmillan.
WAGENMAKERS E.J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779-804.
WAINER H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *4*, 212-213.
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*. Ann Arbor, MI: University of Michigan Press.

The weakness of NHST approach and the Bayesian alternative

Summary. The field of experimental psychology uses NHST as the main vehicle for statistical inference. This approach reduces to considering only p-value in the evaluation of statistical significance. Since p-value depends on sample size, data that were never observed, and subjective intentions it does not quantify statistical evidence (Wagenmakers, 2007). In this paper, correct p-value interpretation and p-value problems are presented. The Bayesian approach is here proposed to solve p-value problems in t-test context. In particular, the usage of Bayesian analysis allows to support null hypothesis evidence.

Keywords: Null hypothesis significance testing, Bayes' theorem, Bayes Factor, t-test, Posterior probability.

La corrispondenza va inviata a Massimiliano Pastore, Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università di Padova, via Venezia 8, 35100 Padova, e-mail: massimiliano.pastore@unipd.it