Franca Agnoli, Patrizio Tressoldi
Dipartimento di Psicologia dello Sviluppo e della Socializzazione - Università di Padova
Dipartimento di Psicologia Generale - Università di Padova

Riportare i risultati delle ricerche sperimentali: come integrare le informazioni fornite dai test statistici

aggiungere titolo abstract

The purpose of this work is to alert the child experimental psychology research community in Italy that reporting the results of Null Hypothesis Significance Testing is insufficient evidence for conclusions about the experimental effects under study. First, we show some common biases of researchers related to the interpretation of Type 1 errors. Second, we identify some misunderstandings in the conclusions that researchers commonly draw from insufficient information. In accordance with the recommendations of the fifth edition of the APA Manual (2001), we indicate how *a*) the presentation of confidence intervals and *b*) various indices of Effect Size address the problems indicated above. Finally, we discuss what "statistical significance" means and compare it to the meaning of practical and/or clinical significance.

Fraintendimenti ed errori nella presentazione dei risultati

Nella presentazione dei risultati delle ricerche sperimentali si verificano comunemente alcuni fraintendimenti ed errori che vanno seriamente a compromettere l'interpretazione e la lettura della ricerca presa in analisi.

Scopo del presente lavoro è illustrare questi errori e, al contempo, offrire una serie di suggerimenti per migliorare questa pratica molto diffusa ma poco appropriata.

La sezione riguardante la presentazione dei risultati molto spesso non è altro che un elenco di analisi statistiche condotte senza che i dati vengano descritti, senza che gli aspetti più importanti della ricerca vengano evidenziati e senza che i dati vengano impiegati per fornire una base logica alle conclusioni da trarre.

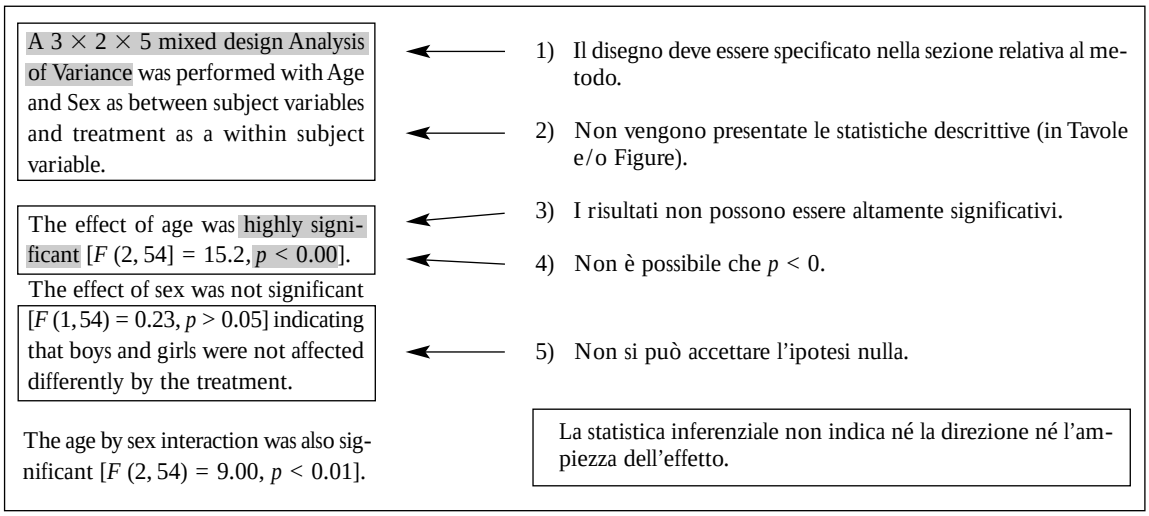
Nella scheda 1 sono rappresentati alcuni errori di una presentazione tipica dei risultati. Gli errori indicati sono di carattere stilistico, metodologico/statistico e logico.

Supponiamo che un ricercatore di psicologia dello sviluppo inizi la sezione dei risultati di uno studio sperimentale nel modo seguente: «È stata eseguita un'Analisi della Varianza sui dati di un disegno misto: $3 \times 2 \times 5$ con i fattori Età (3 livelli) e Sesso (2 livelli) come variabili tra soggetti e trattamento (5 livelli) come variabile ripetuta. L'effetto dell'età è risultato altamente significativo [F (2, 54) = 15.2, $p < 0.00$]. L'effetto Sesso non è risultato significativo [F (1, 54) = 0.23, $p > 0.05$] indicando che maschi e femmine non sono influenzati in modo diverso dal trattamento. Anche il fattore Età $\times$ Sesso risulta significativo [F (2, 54) = 9.00, $p < 0.01$]».

Non presentare subito la descrizione dei dati in Tavole e/o Figure, ma bensì i risultati delle analisi statistiche senza ricordare al lettore che cosa si sta studiando, non facilita la comprensione di quali questioni teoriche sono affrontate nell'analisi dei dati. Nella sezione dei risultati è quindi opportuno riprendere il quesito teorico e specificare che cosa è stato misurato per rispondere alla domanda teorica.

Il ricercatore lavora come un detective; se non presenta i dati

Scheda 1 - Un modo frequente di presentare i risultati di una ricerca con alcuni errori di interpretazione dei dati.



descrittivi, il lettore non riesce a ricostruire la trama della storia. Le statistiche inferenziali (*t*, *F*, ecc.) non si adattano per nulla al linguaggio narrativo della storia ed è per questa ragione che è meglio iniziare la sezione dei risultati con la descrizione dei dati.

Abelson, in un testo intitolato *Statistics as principled argument* (1995), suggerisce di sviluppare la storia secondo 5 criteri: 1) la grandezza dell'effetto, 2) il grado in cui si riesce ad articolare i dati, 3) la generalizzazione dell'effetto, 4) il grado di interesse della storia e 5) il livello di credibilità. Il compito del ricercatore consiste nel risolvere un caso interessante: il pattern unico dei risultati della nostra ricerca; al contempo, bisogna escludere le alternative e mettere in difficoltà gli scettici che mettono in discussione i dati raccolti.

Nella maggior parte delle ricerche pubblicate la sezione riguardante i risultati viene, invece, presentata in modo simile a quello che abbiamo raffigurato nella scheda 1.

La Verifica di Ipotesi Nulla: limiti e controversie

I motivi che hanno condotto all'impoverimento e all'automatizzazione nella presentazione dei risultati hanno un'origine storica.

Il metodo comunemente accettato di analisi statistica nelle ricerche sperimentali è chiamato la Verifica di Ipotesi Nulla (in inglese si usa l'acronimo NHST che sta ad indicare *Null Hypothesis Significance Testing*).

Quella che Murray e Dosser (1987) chiamarono «the inference revolution» in psicologia avvenne tra il 1940 e il 1960, e «the Intro Stats Method» (Dixon e O'Reilly, 1999) venne largamente adottato nei libri di testo, all'interno degli insegnamenti universitari (negli Stati Uniti e poi in Europa), e le riviste lo considerarono *il* metodo per testare le ipotesi. Per almeno 60 anni quindi il *Null Hypothesis Significance Testing* (NHST) è stato parte integrante e fondamentale della ricerca nelle scienze del comportamento e dell'educazione ma, da almeno 60 anni, è oggetto di controversie.

L'affermazione del NHST ha sicuramente portato notevoli vantaggi, come la velocissima crescita delle scienze del comportamento a partire dal 1945 e l'utilizzo di un linguaggio comune con un consequenziale avvicinamento a quelli che sono i canoni delle scienze naturali. Come ogni metodo che però viene elevato a dogma, l'utilizzo del NHST ha subito dei costi elevatissimi, e, come hanno ben

Tab. 1 - Interpretazioni solitamente date da 70 ricercatori a $p < .01$.

Affermazione	Frequenze	%
1) H_0 è assolutamente confutata.	1	1.4
2) È stata trovata la probabilità di H_0.	32	45.7
3) H_1 è assolutamente dimostrata.	2	2.9
4) Si può dedurre la probabilità di H_1.	30	42.9
5) Si conosce la probabilità che la decisione presa sia errata.	48	68.6
6) Una replica ha una probabilità di .99 di essere significativa.	24	34.3
7) Si conosce la probabilità che i dati provengano da popolazioni in cui è vera l'ipotesi nulla.	8	11.3

dimostrato Anderson, Burnham e Thompson (2000), soprattutto a partire dagli anni '70, il numero di articoli che mettono in discussione l'utilità dei test è salito in maniera esponenziale.

L'applicazione dogmatica del NHST ha avuto non poche conseguenze in psicologia tra cui la persistente confusione nell'interpretazione dei risultati da parte di studenti, autori di libri di testo e ricercatori (Finch, Cumming e Thomason, 2001).

Tra i motivi principali di confusione vi sono l'errata interpretazione del valore *p* e l'identificazione del livello *alpha* di .05 come criterio dicotomico di demarcazione per la valutazione della significatività o meno dei risultati (Abelson, 1997; Kline, 2004).

Valori bassi di *p* sono erroneamente interpretati come determinanti l'improbabilità dell'ipotesi nulla (Falk e Greenbaum, 1995) e, inoltre, si considera spesso il valore *p* come una sorta di indicatore inverso della dimensione dell'effetto, dell'importanza di un risultato (Thompson e Snyder, 1998) e della replicabilità (Gigerenzer, 1993; Oakes, 1986), ma un basso valore di *p* non implica necessariamente che siamo di fronte a un risultato importante.

Ci si può chiedere se i ricercatori sono immuni da errori nell'interpretazione del valore *p*.

In tabella 1 sono riportati i risultati di una ricerca nella quale è stato chiesto a 70 ricercatori universitari di psicologia quale interpretazione essi adottano generalmente per dare significato a $p < .01$. Gli interrogati potevano dare più di una risposta. Delle sette affermazioni riportate nella tabella, solo l'ultima è corretta, ed è stata indicata appena da 8 intervistati su 70 (11%).

Circa il 50% di essi sostiene le affermazioni 2 e 4 riportate nella tabella 1: i valori *p* indicherebbero la probabilità condizionata dell'ipotesi nulla oppure dell'ipotesi alternativa. La maggioranza degli intervistati ha sostenuto (erroneamente) che i valori *p* sono probabilità a posteriori dell'Errore di I tipo (affermazione numero 5); inoltre, circa un terzo degli intervistati sostiene che il complemento del valore *p* (cioè $1 - p$) indica la probabilità di ottenere gli stessi risultati nella replica dell'esperimento (affermazione numero 6).

Non è difficile trovare errori simili a quelli riportati nella tabella 1 anche nelle pubblicazioni. Kirk (1996) afferma che «[…] il cambiamento potrà avvenire negli anni futuri se verranno rettificate le istruzioni date agli autori degli articoli scientifici nell'ambito della psicologia e dell'educazione. Questa svolta causerà una reazione a catena: i docenti di statistica muteranno i programmi dei loro corsi, gli autori dei libri di testo pubblicheranno delle revisioni e gli autori di articoli pubblicati nelle riviste scientifiche modificheranno le loro strategie inferenziali».

La dimensione dell'effetto e gli intervalli di fiducia: una svolta possibile

Nel 1996 la sempre maggiore messa in discussione del NHST e le consolidate e diffuse difficoltà relative alla presentazione

dei risultati portano il *Board of Scientific Affairs of the American Psychological Association* (APA) a convocare la Task Force sull'Inferenza Statistica (TFSI) per fare chiarezza su alcune questioni controverse inerenti all'applicazione della statistica, questioni di cui fanno parte anche la significatività statistica e le sue possibili alternative.

In quell'occasione viene enfatizzato che il *valore p* non è un indice adeguato della dimensione dell'effetto: per questa ragione è emersa l'importanza di riportare sempre una qualche misura della dimensione dell'effetto (*effect size*, ES) e di indicare gli Intervalli di Fiducia attorno alle medie. Inoltre viene discusso il ruolo dei test statistici tradizionali per cercare di dare una svolta al dibattito relativo all'efficacia del loro utilizzo.

Le indicazioni emerse possono essere riassunte nel seguente modo (vedi l'articolo di Wilkinson e della Task Force on Statistical Inference, 1999):

1) Usare il numero minimo di analisi necessarie;

2) Non riportare i risultati delle analisi effettuate mediante software statistici senza conoscere il loro significato;

3) Documentare le ipotesi relative alla popolazione o alla stima della potenza statistica dello studio. Usare gli Intervalli di Fiducia per definire i risultati ottenuti;

4) Riportare sempre le statistiche descrittive per permettere repliche dello studio e facilitarne la meta-analisi;

5) Specificare gli Intervalli di Fiducia per gli indici della dimensione dell'effetto;

6) Dimostrare che i dati sostengono effettivamente le ipotesi statistiche.

Con l'ultima edizione del Manuale APA (2001), quelli che nel 1996 erano soltanto degli incoraggiamenti diventano linee guida ufficiali alla luce di questioni che non possono più essere ignorate, quali, appunto, l'importanza pratica e "reale" dei risultati, la replicabilità e la meta-analisi. Nell'ultima edizione del Manuale APA viene ufficialmente sancito che «... affinché il lettore possa comprendere completamente l'importanza dei risultati ottenuti in una ricerca, risulta necessario riportare sempre una qualche misura indicante la dimensione dell'effetto o la consistenza della relazione tra le variabili» e che riportare sempre le misure degli indici di dimensione dell'effetto accanto al valore di *p* «aiuta, grazie ad una rapida osservazione, a collocare l'effetto sperimentale in un contesto teorico e pratico... Dobbiamo insistere ancora sul fatto che riportare e interpretare le misure degli indici di dimensione dell'effetto, anche nel contesto di studi presentati in passato, è essenziale per una buona ricerca» (Manuale APA, 2001; per la controversia suscitata da queste raccomandazioni si veda Fidler, 2002).

L'errore che spesso si vede nei resoconti dei risultati è proprio quello di assumere che la grandezza di *p* sia un indice numerico dell'ampiezza di un certo effetto. Una prima conseguenza è usare frasi come "il risultato è altamente significativo" oppure "è estremamente significativo" (si veda scheda 1); una seconda conseguenza riguarda la non comprensione dell'arbitrarietà della scelta di $p < .05$ (con frasi del tipo "i risultati si avvicinano alla significatività"); una terza conseguenza è confrontare risultati di ricerche diverse sulla base dei livelli di probabilità.

I risultati dei test statistici e i rispettivi valori *p* dipendono entrambi dalla dimensione del campione e dall'*effect size*, così che un effetto non molto ampio può risultare statisticamente significativo in un campione sufficientemente grande. Se il campione è sufficientemente ampio, bassi valori di *p* semplicemente confermano che il campione era ampio; questa però è una tautologia (Thompson, 1992).

Quest'ultimo caso può essere illustrato dall'analisi dei dati presentata in tabella 2 (Keppel, Saufley e Tokunaga, 2001).

Nella parte superiore sono riportati i risultati di un esperimento in cui 5 bambini vengono casualmente assegnati a ciascuna delle tre condizioni (lode, critica, oppure condizione di controllo). È stata effettuata un'Analisi della Varianza ad un fattore con F ottenuto = 4.22 (con 2 e 12 gradi di libertà, $p < .05$). Nella parte inferiore della tabella 2 il numero dei partecipanti assegnato a ciascuna delle tre diverse condizioni è stato

Tab. 2 - Effetto del raddoppio di N.

Condizioni	Lode	Critica	Controllo
Dati grezzi	7	9	2
	8	4	7
	6	6	5
	10	9	3
	7	8	5
Media	7.6	7.2	4.4

ANOVA

	SS	*df*	*MS*	*F*
A	30.4	2	15.2	4.22*
S/A	43.2	12	3.6	
Totale	73.6	14		

Condizioni	Lode	Critica	Controllo
Dati grezzi	7 7	9 9	2 2
	8 8	4 4	7 7
	6 6	6 6	5 5
	10 10	9 9	3 3
	7 7	8 8	5 5
Media	7.6	7.2	4.4

ANOVA

	SS	*df*	*MS*	*F*
A	60.8	2	30.4	9.5*
S/A	86.4	27	3.2	
Totale	147.2	29		

raddoppiato. Le medie sono equivalenti, ma questa volta *F* ottenuto (con 2 e 27 gradi di libertà) è uguale a 9.5 ($p < .01$). Da questo semplice esempio si può osservare come la grandezza dell'indice *F* è tanto maggiore quanto maggiore è la dimensione del campione. Confrontare i risultati sulla base dei livelli di probabilità rappresenta un uso non appropriato del test delle ipotesi.

La stima della dimensione dell'effetto sperimentale (*ES*) è, invece, un indice non influenzato direttamente dal numero di soggetti appartenenti alle condizioni sperimentali (vedi l'esempio in tabella 2), e quindi non può coincidere né col valore *F*, né con la probabilità associata al valore ottenuto (il valore *p*).

Introduciamo brevemente il significato degli indici di dimensione dell'effetto. Supponiamo ad esempio di voler studiare l'efficacia di una certa strategia di memorizzazione; a tal fine creiamo due gruppi, A e B, formati ciascuno da 50 individui.

Il gruppo A viene addestrato ad imparare a memoria del materiale con il metodo che stiamo studiando, mentre al gruppo B non viene insegnato nessun metodo particolare. Si procede con la raccolta dei dati e con la verifica statistica delle ipotesi, dalla quale si ottengono indicazioni sulla probabilità che i nostri dati siano conformi all'ipotesi nulla.

Tab. 3 - Differenze di medie standardizzate per due diversi gruppi di contrasti (Kline, 2004).

Studio	$M_1 - M_2$	$\sigma_{stimato}$	*d*
	Differenza tra le medie diversa, uguale dimensione dell'effetto		
1	75.00	100.00	.75
2	11.25	15	.75
	Differenza tra le medie uguale, diversa dimensione dell'effetto		
3	75.00	500.00	.15
4	75.00	50.00	1.50

Dopo aver effettuato l'analisi dei dati la domanda che è necessario porsi è: quanto è stato efficace il trattamento al quale sono stati sottoposti i soggetti? Per rispondere a ciò si calcola la misura dell'indice di dimensione dell'effetto più appropriata alla situazione sperimentale. L'*effect size* quindi ci fornisce proprio la misura di quanto è ampio il cambiamento che si è prodotto nel gruppo di soggetti sottoposti all'esperimento, cambiamento che è in gran parte dovuto al trattamento sperimentale.

È importante che i ricercatori riportino sempre almeno un indice della dimensione dell'effetto ottenuto nei propri esperimenti perché solo in questo modo è possibile confrontare poi i risultati di due o più distinte ricerche in maniera diretta.

Ad ogni modo non è questa la sede appropriata per discutere le molteplici tipologie di indici di dimensione dell'effetto che esistono nella letteratura, ma, a scopo prettamente esemplificativo, riportiamo l'indice *d* di Cohen (1988) il quale rappresenta la forma più semplice e più utilizzata di *effect size* (per una maggiore trattazione si veda Fern e Monroe, 1996).

L'indice *d* di Cohen è una differenza media standardizzata e la sua applicazione è relativa alle situazioni in cui il fine di una ricerca è comparare le performance di due gruppi (ad esempio il gruppo di trattamento *vs* il gruppo di controllo, uomini *vs* donne) sulla base di variabili di tipo continuo.

Viene definito come la differenza tra le medie di due gruppi divisa dalla deviazione standard stimata della popolazione di ciascun gruppo (assunta essere la medesima):

$$d = \frac{M_1 - M_2}{\sigma_{stimato}}$$

L'interpretazione di *d* è semplice: se, per esempio, $d = .50$, allora M_1 è mezza deviazione standard più grande di M_2. Il segno di *d* è arbitrario ma bisogna sempre spiegare come è stato derivato, in quanto non esistono simbologie universali che lo definiscano. In linea di principio può accadere che *d* assuma anche valori molto alti, e dire che $d = 4.00$ significa che tra le medie c'è una differenza di quattro deviazioni standard! Ad ogni modo è relativamente difficile trovare differenze così ampie all'interno delle scienze del comportamento (Sedlmeier e Gigerenzer, 1989; Kline, 2004).

La tabella 3 rappresenta i risultati di quattro studi ipotetici che mettono in evidenza come il soffermarsi semplicemente sulla differenza tra le medie di due gruppi sia molto spesso fuorviante e non indicativo di un effetto sperimentale realmente solido e cospicuo. Nella parte superiore della tabella viene mostrato come una misura della dimensione dell'effetto ampia possa sussistere sia con una differenza tra le medie elevata che con una differenza tra le medie molto più ridotta; nella parte inferiore, invece, ad un identico valore della differenza tra le medie, che potrebbe far supporre un'equivalenza tra i due studi, corrispondono misure della dimensione dell'effetto molto diverse.

È per questo motivo che la sola differenza tra le medie di due gruppi non è sintomatica della presenza di un effetto sperimentale elevato neppure se ampia e tanto più se confrontata fra studi diversi.

Una posizione ancora più estrema (Loftus, 1993; 2002; Masson e Loftus, 2003) suggerisce: *a*) di presentare i dati in una figura in cui vengono illustrate le medie del campione con gli appropriati intervalli di fiducia attorno alle medie e *b*) di non includere un'analisi statistica relativa alla verifica di ipotesi che sia ridondante rispetto alle informazioni già presenti nelle figure.

Per illustrare l'importanza dell'uso degli intervalli di fiducia Loftus e Masson (1994) riportano i dati di uno studio ipotetico sulle prestazioni di memoria di figure o parole. In questo studio gli intervalli di ritenzione variano da 0 a 14 giorni dal momento della presentazione degli stimoli. Normalmente la modalità di presentazione dei risultati per questo tipo di esperimento è la seguente: viene illustrata una tabella con le medie che corrispondono alle 8 condizioni sperimentali (2 tipi di stimolo e 4 intervalli di ritenzione). Poi vengono riassunte all'interno del testo le statistiche inferenziali e con molta enfasi si riportano i valori *p* (dando quindi importanza all'errore di primo tipo).

Nel caso in cui si preferisca presentare le medie con una figura non viene solitamente rappresentata nessuna misura di variabilità (si veda la figura 1, riprodotta come Figura 1a da Loftus e Masson, 1994).

Che cosa c'è di errato in questa presentazione dei risultati? Questo modo di presentare i dati non fornisce nessuna informazione sulla precisione delle medie. Si supponga che ci sia poca variabilità attorno alle medie (si veda fig. 2, riprodotta come Figura 1b in Loftus e Masson 1994): in tutti e quattro i livelli di ritenzione gli intervalli di fiducia attorno alle medie non si sovrappongono. Gli effetti sono chiaramente significativi (non c'è bisogno di presentare altre statistiche *t*, *F*, ecc.).

Si ipotizzi, invece, che ci sia ampia variabilità attorno alle medie come illustrato dalla figura 3 (riprodotta come Figura 1c in Loftus e Masson, 1994). In questo caso si può osservare che gli intervalli di fiducia attorno alle medie si sovrappongono: gli effetti sono chiaramente non significativi dato che la variabilità attorno alle medie è molto elevata.

Si possono individuare tre ragioni per seguire le precedenti indicazioni: 1) la presentazione grafica delle medie dei campioni fornisce delle intuizioni relative agli andamenti dei dati; 2) con gli intervalli di fiducia si hanno maggiori informazioni riguardo alla posizione delle medie della popolazione; 3) gli intervalli di fiducia rispondono a considerazioni sulla potenza

Fig. 1 - Rappresentazione dell'andamento delle medie senza Intervalli di Fiducia attorno alle medie.

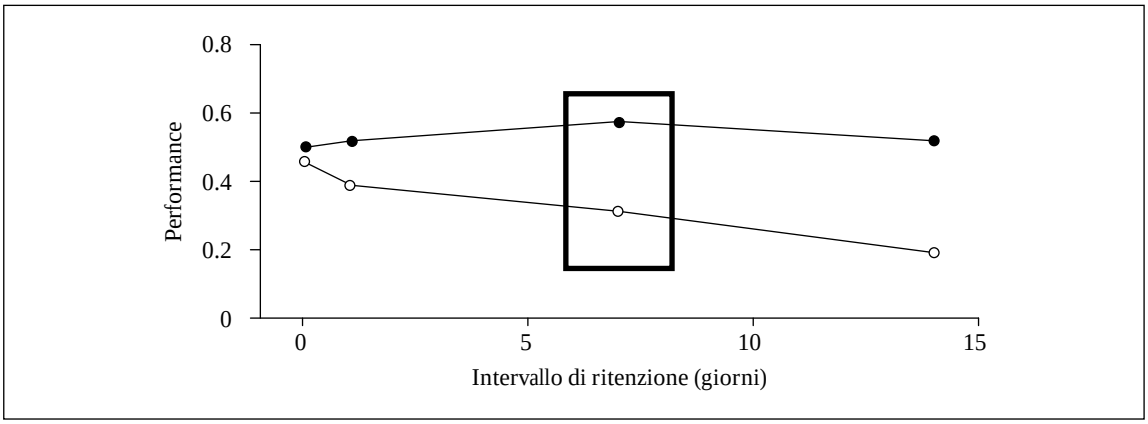


Fig. 2 - Rappresentazione delle medie con i relativi Intervalli di Fiducia che indicano una bassa variabilità attorno alla medie.

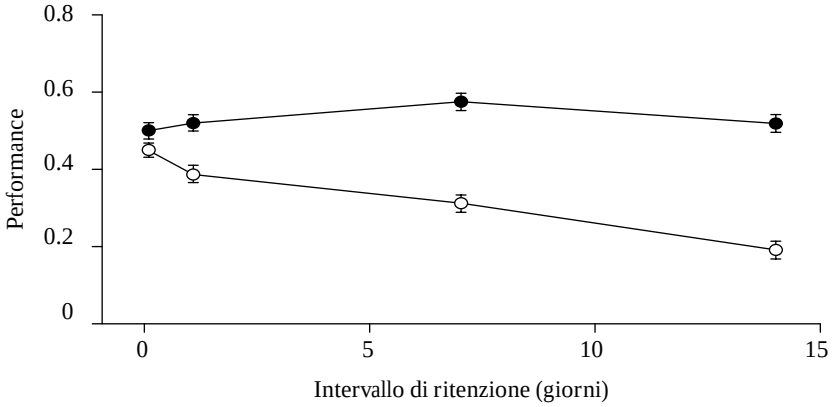
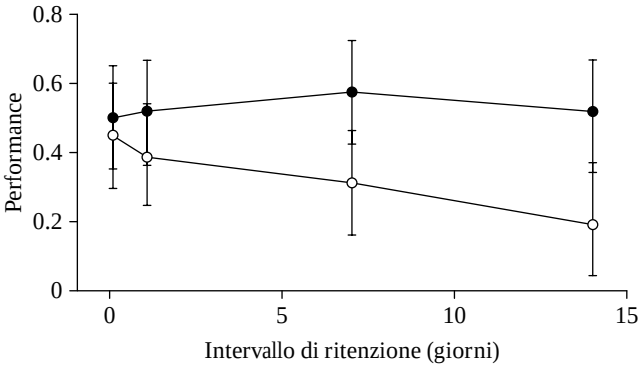


Fig. 3 - Rappresentazione delle medie con i relativi Intervalli di Fiducia che indicano una altissima variabilità attorno alle medie.



del test statistico e sull'adeguatezza dell'ampiezza del campione usato nell'esperimento.

Nella tabella 4 viene indicato un esempio su come riportare gli intervalli di fiducia nel caso in cui la presentazione dei dati sia fatta tramite tabella piuttosto che figura.

In questa tabella vengono riportati i dati di uno studio (Olds, Henderson e Tatelbaum, 1994) che confronta il quoziente intellettivo a 48 mesi ed il peso alla nascita dei rispettivi figli di madri fumatrici e di madri non fumatrici. Al di là delle differenze tra le medie, gli intervalli di fiducia attorno alle differenze permettono di avere una misura della stima della differenza nella popolazione: possiamo concludere che la differenza di peso alla nascita potrebbe essere di soli 167.1 grammi oppure di 594.9 grammi. In ogni caso si può affermare che madri fumatrici avranno

Tab. 4 - Modo corretto di presentare gli Intervalli di Fiducia avvalendosi dell'uso di tabelle.

	Medie			
	0 Sigarette	10+ Sigarette	Differenza	(95% IF)
Livello di istruzione della madre (in anni di scolarità)	11.57	10.89	.67	(0.15, 1.19)
Peso alla nascita	3416	3035	381.0	(167.1, 594.9)
Stanford – Binet (QI) 48 mesi	113.28	103.12	10.16	(5.04, 15.30)

bambini con un peso inferiore alla nascita rispetto a madri non fumatrici.

Significatività pratica e importanza reale di un effetto

Nell'ultima parte di questo articolo affronteremo il problema dell'importanza reale dei risultati statistici. La significatività pratica di un risultato è sempre funzione di un'accurata analisi statistica ma anche di variabili che attengono alla sfera sociale e di giudizio soggettivo dalle quali nessun risultato scientifico può sottrarsi (Kirk, 2001).

Ma la grandezza di un indice di dimensione dell'effetto è sempre proporzionale alla sua importanza? La risposta a questa domanda potrebbe sembrare intuitivamente ovvia, ma non è così, in quanto una misura di dimensione dell'effetto ampia non garantisce necessariamente il fatto che siamo di fronte ad un risultato importante a livello pratico; del resto è anche vero che un effetto di modesta grandezza può non essere di scarsa importanza (Kline, 2004).

Come abbiamo avuto modo di vedere, gli indici di dimensione dell'effetto possiedono molte proprietà positive: hanno delle convenzioni che consentono di definire la loro ampiezza, possono essere usati per comparare quantitativamente i risultati di due o più ricerche, e sono utili nell'analisi della potenza per indirizzare la decisione su quanti soggetti sono necessari in una determinata ricerca. In sintesi, gli indici di dimensione dell'effetto sono una misura quantitativa semplice e di facile comprensione, misura che fornsce indicazioni utili sulla reale importanza di un effetto sperimentale (Prentice e Miller, 1992); gli indici di dimensione dell'effetto dovrebbero essere sempre riportati sia per risultati significativi che no (Rosnow e Rosenthal, 1989; Thompson, 2000).

Tuttavia, se ci fermassimo a questo punto, cioè applicando gli indici di dimensione dell'effetto in maniera acritica, commetteremmo lo stesso tipo di errore che molti autori imputano alla pratica del *Null Hypothesis Significance Testing* (NHST), cioè di farne un uso ritualistico. Non dobbiamo dimenticare che le linee guida interpretative fornite da Cohen (1988) per descrivere qualitativamente la grandezza e la portata della dimensione di un effetto non hanno una base empirica e non devono quindi essere applicate rigidamente né, tanto meno, condurre ad erronei automatismi di ragionamento che stabiliscano una sorta di relazione bidirezionale tra grandezza di un effetto ed importanza dello stesso. Inoltre non possiamo perdere di vista il contesto all'interno del quale una ricerca viene condotta in quanto la dimensione di un effetto e la sua portata a livello teorico e pratico si riconducono alla specifica disciplina di appartenenza (Lenth, 2001).

Supponiamo che sia stata compiuta una ricerca dalla quale è emerso che la differenza media di altezza fra uomini e donne è circa pari a due deviazioni standard ($d = 2$); una differenza così ampia ha anche un'importanza pratica altrettanto elevata? La risposta, chiaramente, dipende dal contesto di ricerca. In termini di indagine psicologica una tale differenza di genere in altezza è

probabilmente irrilevante, ma, nel contesto della sicurezza automobilistica, invece, potrebbe essere cruciale.

Kline (2004) riporta una questione che era sorta alla fine degli anni '90 sulla strutturazione dell'air bag anteriore all'interno delle autovetture: il problema riguardava il fatto che, in caso di incidente, la forza con cui l'air bag si gonfiava avrebbe potuto ferire, o addirittura soffocare, una persona di bassa statura, cioè comportava un grosso rischio specialmente per le donne. Per le automobili, invece, dotate di un sistema "intelligente" di scoppio dell'air bag che si regolava automaticamente a seconda dell'altezza del guidatore, un'ampia differenza di altezza tra uomini e donne sarebbe risultata, ovviamente, molto meno importante.

La stessa logica sottende il fatto che effetti ritenuti di piccole dimensioni siano invece, sul piano pratico, importanti, e questo è proprio quello che hanno sostenuto Prentice e Miller (1992) argomentando che un effetto piccolo può essere molto rilevante sul piano teorico e può diventare di ampie dimensioni nel corso del tempo; inoltre, in certi contesti come le scelte di politica sanitaria, decisioni importanti vengono prese sulla base di effetti che sono quantitativamente insignificanti (Gage, 1978), fatto non poco comune in ambito medico. Pensiamo ad esempio al ruolo dell'aspirina nel prevenire l'infarto; il suo effetto spiega soltanto circa il 2% della varianza totale ma questo risultato diventa molto importante nel momento in cui significa, sul piano reale, salvare un numero cospicuo di vite.

In psicologia il concetto di importanza di un effetto si snoda lungo due strategie di definizione diverse spesso in disaccordo: l'una statistica e l'altra metodologica. L'approccio statistico è molto più adeguato in aree della psicologia in cui l'operazionalizzazione della variabile indipendente e la scelta della variabile dipendente possono essere chiaramente definite dal problema stesso, e quindi gli indici di dimensione dell'effetto sono una misura completamente appropriata nella definizione di importanza, consentendo, inoltre, un'accurata meta-analisi.

L'approccio metodologico coinvolge invece aree di ricerca in cui i disegni sperimentali sono molto più complessi, e le variabili in questione sono difficili da controllare. Questo tipo di ricerche, anche se gli indici di dimensione dell'effetto sono di modeste dimensioni e condurre una meta-analisi è difficoltoso, riescono tuttavia a dimostrare effetti importanti, importanti nel senso di influenti nel vivere quotidiano.

Queste considerazioni ci conducono verso il problema di come non ci sia accordo nel definire cosa rende un effetto realmente importante e quali sono i criteri, gli standard per arrivare alla definizione di ciò. Vero è che i ricercatori non possono esimersi dal rispondere a tre interrogativi basilari (Kirk, 2001): *a*) un certo effetto è presente o dovrebbe essere attribuito al caso? *b*) se l'effetto è presente, quanto ampio è?; *c*) l'effetto è ampio abbastanza da essere considerato importante?

Per rispondere a queste domande gli indici di dimensione dell'effetto sono sicuramente la componente fondamentale per determinare quanto un effetto è significativo, importante, sul piano pratico. Per fare ciò non possiamo trascurare il fatto che l'importanza pratica di un effetto viene stabilita dallo specifico contesto di ricerca in cui è applicato e che il concetto di significatività pratica dipende comunque dal tessuto sociale, da considerazioni personali e dal calcolo dei costi e dei benefici (Kirk, 1996).

Sfortunatamente non esistono analisi statistiche in grado di misurare direttamente la significatività pratica di un effetto; tuttavia gli indici di dimensione dell'effetto possono aiutare in ampia misura il ricercatore a decidere se un risultato è significativo e importante a livello pratico (Kirk, 2001).

Le persone che lavorano nel campo della ricerca o della relativa applicazione pratica devono sempre tener conto dei fattori contestuali relativi sia al momento dell'esecuzione dell'esperimento (o quasi-esperimento, o analisi di un caso, …) che alla fase applicativa.

«Non esistono semplici convenzioni per determinare l'importanza pratica. Come per i

bambini, vale anche per gli indici di dimensione dell'effetto la regola che per capirli bene è meglio studiarli sempre in riferimento al loro contesto» (McCartney e Rosenthal, 2000).

Gigerenzer (1998) afferma che è importante che chi lavora nel campo psicologico conosca i diversi strumenti statistici a sua disposizione, per poter così scegliere quelli più adatti ad ogni situazione e, inoltre, secondo l'autore è necessario sviluppare il pensiero statistico (*statistical thinking*), che è un pensiero creativo, artistico, «[...] è un'arte, non una procedura meccanica».

È chiaro che stiamo uscendo dal campo propriamente matematico, ma ciò si rende necessario. Quando un ricercatore prende una decisione sull'importanza pratica dei suoi risultati non può utilizzare solamente dei criteri oggettivi, dei *rituali* (Gigerenzer, 1998). In questo caso interviene il giudizio personale che, come afferma Kirk (1996):

«[...] inevitabilmente implica una varietà di considerazioni, incluso il sistema di valori del ricercatore, considerazioni sulla società, sul bilancio tra costi e benefici, eccetera. I ricercatori hanno il compito di prendere una complessa serie di decisioni nella progettazione e nell'esecuzione di un esperimento ma è curioso che, in nome dell'oggettività, non spetti a loro decidere se i dati che hanno raccolto ed analizzato hanno validità pratica».

Questa ambiguità molto diffusa (Dixon, 1998) è in realtà estremamente fuorviante perché un test di significatività dell'ipotesi nulla ci dice soltanto la probabilità che abbiamo di ottenere un certo effetto se l'ipotesi nulla è vera ma non ci dice quanto un effetto è ampio, importante o utile (Kirk, 2001); «the statistical significance test does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does» (Cohen, 1994).

È alla luce di queste considerazioni che, di pari passo con l'esponenziale crescita delle pubblicazioni che criticano l'uso indiscriminato dei test statistici, gli indici di dimensione dell'effetto hanno assunto un ruolo sempre più preponderante come misura in grado di definire sia la grandezza che l'importanza reale di un effetto (Cohen, 1988; 1990; Kirk, 1996; 2001; Thompson, 1996; 1999; 2002).

Bibliografia

ABELSON R.P. (1995), *Statistics as principled argument*, Erlbaum, Hillsdale.

ABELSON R.P. (1997), «On the surprising longevity of flogged horses: Why there is a case for the significance test», *Psychological Science*, *8*, 12-15.

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

ANDERSON D.R., BURNHAM K.P. THOMPSON W.L. (2000), «Null hypothesis testing: Problems, prevalence, and an alternative», *Journal of Wildlife Management*, *64*, 913-926.

COHEN J. (1988), *Statistical power analysis for the behavioural sciences* (2nd ed.), Erlbaum, Hillsdale.

COHEN J. (1990), «Things I have learned (so far)», *American Psychologist*, *45*, 1304-1312.

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

DIXON P. (1998), «Why scientists value *p* values», *Psychonomic Bulletin & Review*, *5*, 390-396.

DIXON P., O'REILLY T. (1999), «Scientific versus statistical inference», *Canadian Journal of Experimental Psychology*, *53*, 133-149.

FALK R., GREENBAUM C.W. (1995), «Significance tests die hard: The amazing persistence of a probabilistic misconception», *Theory & Psychology*, *5*, 75-98.

FERN E.F., MONROE K.B. (1996), «Effect size estimates: Issues and problems in interpretation», *Journal of Consumer Research*, *23*, 89-105.

FIDLER F. (2002), «The fifth edition of the APA *Publication Manual*: Why its statistics recommendations are so controversial», *Edu-*

cational and Psychological Measurement, *62*, 749-770.

FINCH S., CUMMING G., THOMASON N. (2001), «Reporting of statistical inference in the "Journal of Applied Psychology": Little evidence of reform», *Educational and Psychological Measurement*, *61*, 181-210.

GAGE N.L. (1978), *The scientific basis of the art of teaching*, Teachers College Press, New York.

GIGERENZER G. (1993), «The superego, the ego and the id in statistical reasoning». In G. Keren, C. Lewis (Eds.), *A handbook for data analysis in the behavioural sciences: Methodological issues*, Erlbaum, Hillsdale, 311-339.

GIGERENZER G. (1998), «We need statistical thinking, not statistical rituals», *Behavioural and Brain Sciences*, *21*, 199-200.

KEPPEL G., SAUFLEY W.H., TOKUNAGA H. (2001), *Disegno sperimentale e analisi dei dati in psicologia*, EdiSES S.r.l., Napoli.

KIRK R.E. (1996), «Practical significance: A concept whose time has come», *Educational and Psychological Measurement*, *56*, 746-759.

KIRK R.E. (2001), «Promotion of good statistical practices: Some suggestions», *Educational and Psychological Measurement*, *61*, 213-218.

KLINE R.B. (2004), *Beyond significance testing. Reforming data analysis methods in behavioural research*, American Psychological Association, Washington.

LENTH R.V. (2001), «Some practical guidelines for effective sample size determination», *American Statistician*, *55*, 187-193.

LOFTUS G.R. (1993), «A picture is worth a thousand p-values: On the irrelevance of hypothesis testing in the computer age», *Behaviour Research Methods, Instrumentation and Computers*, *25*, 250-256.

LOFTUS G.R. (2002), «Analysis, interpretation, and visual presentation of data». In *Stevens' handbook of experimental psychology, third edition*, John Wiley and Sons, New York, vol. 4, 339-390.

LOFTUS G.R., MASSON M.E.J. (1994), «Using confidence intervals in within-subject designs», *Psychonomic Bulletin & Review*, *1*, 476-490.

MASSON M.E.J., LOFTUS G.R. (2003), «Using confidence for graphically based data interpretation», *Canadian Journal of Experimental Psychology*, *57*, 203-220.

MCCARTNEY K., ROSENTHAL R. (2000), «Effect size, practical importance, and social policy for children», *Child Development*, *71*, 173-180.

MURRAY L.V., DOSSER D.A. (1998), «How significant is a significant difference? Problems with the measurement of magnitude of effect», *Journal of Counseling Psychology*, *34*, 68-72.

OAKES M. (1986), *Statistical inference*, Wiley, New York.

OLDS D.L., HENDERSON C.R., TATELBAUM R. (1994), «Intellectual impairment in children of women who smoke cigarettes during pregnancy», *Pediatrics*, *93*, 221-227.

PRENTICE D.A., MILLER D.T. (1992), «When small effects are impressive», *Psychological Bulletin*, *112*, 160-164.

ROBINSON D., LEVIN J. (1997), «Reflection on statistical and substantive significance, with a slice of replication», *Educational Researcher*, *26*, 21-26.

ROSNOW R.L., ROSENTHAL R. (1989), «Statistical procedures and the justification of knowledge in psychological science», *American Psychologist*, *44*, 1276-1284.

SEDLMEIER P., GIGERENZER G. (1989), «Do studies of statistical power have an effect on the power of studies?», *Psychological Bulletin*, *105*, 309-316.

THOMPSON B. (1992), «Two and one-half decades of leadership in measurement and evaluation», *Journal of Counseling and Development*, *70*, 438-438.

THOMPSON B. (1996), «AERA editorial policies regarding statistical significance testing: Three suggested reforms», *Educational Researcher*, *25 (2)*, 26-30.

THOMPSON B. (1999), «Journal editorial policies regarding Statistical Significance Tests: Heat is to fire as *p* is to importance», *Educational Psychology Review*, *11*, 157-169.

THOMPSON B. (2000), «Reporting practices and APA editorial policies regarding statistical significance and effect size», *Theory & Psychology*, 10 (3), 413-425.

THOMPSON B. (2002), «"Statistical", "practical", and "clinical": How many kinds of significance do counsellors need to consider?», *Journal of Counseling & Development*, *80*, 64-71.

THOMPSON B., SNYDER P.A. (1998), «Statistical significance and reliability analyses in recent *Journal of Counseling & Development* research articles», *Journal of Counseling & Development*, *76*, 436-441.

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