LA SIGNIFICATIVITÀ DELLA SIGNIFICATIVITÀ*

Massimiliano Pastore

Department of Developmental and Social Psychology,
University of Padova, via Venezia, 8, 35131 Padova, Italy
Tel.: +39-49-8277481
E-mail: massimiliano.pastore@unipd.it

Immaginate di essere con vostro figlio sul marciapiede di una strada molto trafficata. Avete appena acquistato il vostro panino preferito ed attraversato la strada, ma vi accorgete che avete scordato la maionese (di cui siete molto golosi). Sapendo che attraversando la strada di nuovo per tornare al negozio dei panini avete una probabilità di .95 di ritornare sani e salvi cosa fate? Ora immaginate lo stesso scenario con la differenza che vi accorgete di aver dimenticato dall'altra parte della strada vostro figlio. Sapendo che attraversando la strada di nuovo per andare a riprendere vostro figlio avete la stessa probabilità di .95 di ritornare sani e salvi cosa fate? Abbiamo due scenari con la stessa probabilità di successo e quindi identici in termini di significatività statistica; entrambe le variabili *maionese* e *figlio* sono significative con $p = .05$. Quello che differenzia gli scenari è il peso, evidentemente diverso, che diamo alle due situazioni.

Con questo esempio, preso da Ziliak and McCloskey (2008), si mette in chiara evidenza l'assurdità di utilizzare un criterio statistico uguale di fronte a situazioni dal valore (leggi dimensione dell'effetto) differente. Lo stesso concetto era già stato espresso da Fisher (1959, p. 42): *"... no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he reject hypotheses ... "*.

Nel contributo di Perugini, tra gli aspetti cruciali per una crescita di credibilità della psicologia viene indicato anche l'incremento della cultura metodologica. In pieno accordo con questa posizione, questo contributo vuole approfondire la discussione mettendo particolare enfasi sul fatto che la cultura metodologica si appoggia anche sulla cultura statistica. Molte delle cosiddette pratiche di ricerca discutibili (*Questionable Research Practices*, QRP; John, Loewenstein, & Prelec, 2012), commesse in buona fede, si fondano su un'errata conoscenza e/o comprensione dei metodi statistici (si veda ad es. Ioannidis, 2005; Gelman & Stern, 2006; Wagenmakers, 2007; Simmons, Nelson, & Simonsohn, 2011) e quindi per ridurle è assolutamente necessario che i ricercatori siano in grado di conoscere il significato delle statistiche che usano. In questo lavoro, utilizzando come esempio il coefficiente di correlazione lineare di Bravais-Pearson, si vuole illustrare come la mancanza di riflessione critica possa condurre a conclusioni prive di significato.

IL COEFFICIENTE DI CORRELAZIONE

Tutti conoscono il coefficiente di correlazione di Bravais-Pearson (r), sembra però che molti meno ricordino che tale coefficiente si usa per misurare la relazione lineare tra due variabili. Tale sospetto può essere supportato semplicemente osservando nelle pubblicazioni scientifiche di area psicologica il rapporto tra il numero di volte in cui viene presentata una tabella di correlazione ed il numero di volte in cui sono presenti dei grafici a dispersione. In figura 1 sono riportati quattro esempi

* Commento all'articolo *"La crisi internazionale di credibilità della psicologia come un'opportunità di crescita: Problemi e possibili soluzioni"* di Marco Perugini, Giornale Italiano di Psicologia.

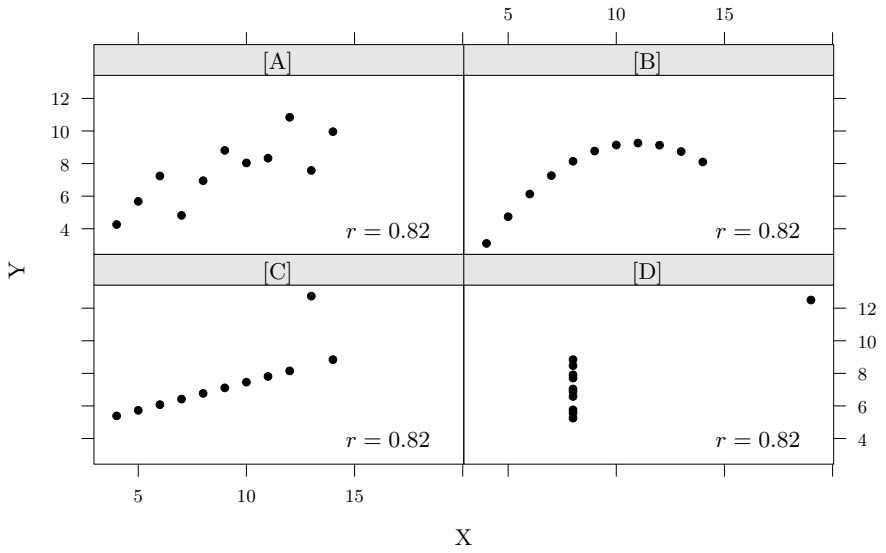

Fig. 1. Casi esemplari in cui configurazioni di punti molto diverse producono la stessa correlazione $r = 0.82$ (Anscombe, 1973).

descritti da Anscombe (1973) in cui osserviamo delle relazioni completamente diverse tra loro ma che producono la stessa identica correlazione $r = 0.82$. Già questo semplice esempio dovrebbe metterci in guardia dall'utilizzare r senza una preventiva osservazione dei dati; si noti in particolare, che nel pannello [D] le due variabili sono praticamente indipendenti ed il valore alto di correlazione dipende solo dalla presenza di un caso anomalo.

Soffermiamoci ora su un altro aspetto peculiare legato a r, ovvero la sua significatività statistica, ed in particolare ci chiediamo: cosa vuol dire significatività statistica di r? Secondo il tradizionale approccio *Null Hypothesis Significance Testing* (NHST; Cohen, 1994) il test su r presuppone la seguente ipotesi nulla: $H_0 : \rho = 0$ in cui ρ rappresenta il valore vero ed incognito della correlazione nella popolazione da cui campioniamo. Quindi, se otteniamo un risultato statisticamente significativo rigettiamo H_0 e concludiamo che $\rho \neq 0$ ovvero che la correlazione è significativamente diversa da zero. Si badi bene però che tale conclusione non implica che la relazione sia forte o che l'effetto sia rilevante. È stato ampiamente dibattuto il fatto che l'aumento della numerosità campionaria produca una riduzione della p associata al test e quindi un "*incremento*" della significatività statistica (si veda, ad. es. Wagenmakers, 2007; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Pastore, 2009; Altoé & Pastore, 2013), ma nel caso della correlazione questo produce effetti particolarmente paradossali.

Si considerino i due casi rappresentati in figura 2, immaginiamo che siano i dati raccolti sulle stesse due variabili X e Y da due ricercatori indipendenti ed ognuno all'insaputa dell'altro. Il primo ha rilevato un campione di $n = 10$ osservazioni mentre il secondo $n = 150$. I due, scrupolosamente, osservano con un grafico a dispersione la relazione tra le due variabili (in figura 2 rispettivamente a sinistra e a destra) e ne calcolano la correlazione. In entrambe le configurazioni la correlazione tra le distribuzioni di punti è $r = 0.2$. A questo punto i due vogliono sapere se la correlazione osservata sia statisticamente significativa ed eseguono un apposito test con approccio NHST. Il primo ricercatore ottiene una statistica test $t = 0.58$, con 8 gradi di libertà e $p = 0.58$ e quindi conclude che la correlazione osservata non è statisticamente significativa. Il secondo ricercatore ottiene una statistica test $t = 2.48$, con 148 gradi di libertà e $p = 0.014$ e quindi conclude che la correlazione osservata è statisticamente significativa. Abbiamo ottenuto il paradosso che la stessa statistica ($r = 0.2$) è diversa da zero, ma anche no. Se i due ricercatori potessero disporre dei dati completi l'uno dell'altro vedrebbero che il risultato ottenuto sul campione più piccolo è perfettamente compatibile

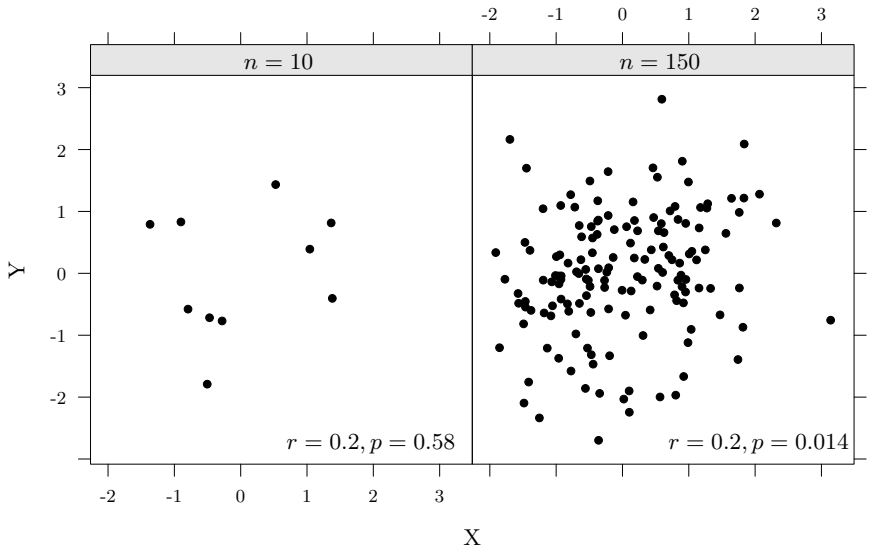

Fig. 2. Grafici a dispersione per due casi con numerosità diverse ($n = 10$ a sinistra e $n = 150$ a destra), stesso valore di correlazione $r = 0.2$ ma con diversi valori di p.

con quello ottenuto sul campione più grande, inoltre una accurata osservazione della figura 2 ci consente di ragionare sul significato della correlazione ottenuta. Supponiamo infatti di voler usare la relazione osservata per stimare i valori attesi su Y in funzione dei valori osservati su X, quanto potrà essere precisa questa previsione? Sembra abbastanza evidente che in un caso del genere la prevedibilità di Y sulla base di quanto osservato in X sia piuttosto bassa e quindi quale può essere l'utilità di avere stabilito, almeno in un caso, che la correlazione tra le due variabili è statisticamente significativa? Ci azzardiamo a dire che l'utilità sia quanto meno limitata: la significatività statistica, per come l'abbiamo utilizzata, nulla ha a che fare con la prevedibilità del comportamento di Y in funzione di X. Ad ulteriore conferma proviamo a calcolare la quota di variabilità spiegata dalla relazione lineare tra le due variabili con il coefficiente di determinazione: $r^2 = 0.2^2 = 0.04$, ovvero la percentuale di variabilità spiegata dalla relazione è del 4%, il che, evidentemente, implica il 96% di variabilità non spiegata dalla relazione e quindi residuo. Insomma, l'informazione data da r che, lo ricordiamo, è una misura di associazione lineare, indica la forza del legame esistente, indipendentemente dalla significatività statistica. In altri termini, r è già di per sé un indice di effect size (Cohen, 1988) e quindi ha ben poco senso testarne la significatività come differenza da zero. Pertanto la valutazione di r deve essere fatta in tali termini, ovvero di grandezza dell'effetto. Si badi bene però, che anche la grandezza di un effetto va considerata ed interpretata in modo relativo (*"The terms 'small,' 'medium,' and 'large' are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation..."*; Cohen, 1988, p. 25). Chi si occupa di variabili raccolte in ambito sociale sa che una correlazione di .4 può essere considerata forte, ma saremmo disposti a farci operare agli occhi con uno strumento automatizzato in cui la posizione definita dal chirurgo correla .7 con la posizione del laser?

CONCLUSIONI

Tra i responsabili della crisi di credibilità della ricerca scientifica Perugini cita le riviste, in quanto orientano di fatto ciò che viene pubblicato. Non sono del tutto d'accordo su questo: le riviste non sono entità astratte, sono pur sempre formate da persone e, sopratutto, i referee delle riviste sono

gli stessi appartenenti della comunità scientifica, e quindi anche noi. Questo ci pone direttamente in una posizione di responsabilità.

Sono fermamente convinto che una appropriata conoscenza dei metodi statistici comporti anche una maggiore sensibilità ad evitare, ovviamente laddove non interviene la frode, le QRP. A tale scopo, è importante acquisire una nuova forma di ragionamento statistico, fondato sull'osservazione dei dati e non sulla ripetizione acritica di schemi comportamentali/procedurali, nati negli anni '30 del secolo scorso con una logica precisa (si veda a tale proposito Berger, 2003) ma che con il tempo è andata persa, trasformandosi in un vero e proprio culto (Ziliak & McCloskey, 2008); senza trascurare il fatto, sul quale non ci siamo soffermati nel presente contributo, che anche le tecniche statistiche si sono evolute rendendo possibile la trattazione di problemi prima non gestibili. Anche l'ossessione verso la significatività statistica deve essere seriamente ridimensionata. Un risultato non può essere considerato "*buono*" solo se statisticamente significativo, ma piuttosto se apporta un contributo significativo alla conoscenza; da questo punto di vista anche l'evidenza di effetti non significativi può essere rilevante (ad es. Gallistel, 2009; Sörqvist, Marsh, & Nöstl, 2013) ed è molto interessante la proposta di Francis (2013) di sostituire il termine *"significantly different"* con *"apparently different"*.

Mi piace immaginare la statistica come un'automobile, ovvero uno strumento utile per muoversi sulle strade della ricerca. Per guidare l'automobile non è strettamente necessario conoscere il funzionamento del motore, saperlo riparare o progettarne di nuovi (è il compito degli statistici), ma è assolutamente obbligatorio saper manovrare il mezzo opportunamente e conoscere il codice della strada. E chi di noi sarebbe disponibile, oggi, a viaggiare con un'automobile costruita nel 1930?

References

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

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*, 17–21.

Berger, J. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, *18*, 1–12.

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

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

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

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

Gallistel, C. R. (2009). The Importance of Proving the Null. *Psychological Review*, *116*, 439–453.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *American Statistician*, *60*, 328–331.

Ioannidis, J. (2005). Why most published research findings are false. *PLOS Medicine*, *2*, 696–701.

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

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

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.

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

Sörqvist, P., Marsh, J. E., & Nöstl, A. (2013). High working memory capacity does not always attenuate distraction: Bayesian evidence in support of the null hypothesis. *Psychonomic Bulletin & Review*, *20*, 897–904.

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

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