IN QUESTO NUMERO

Annibale Biggeri[1,2]

[1] Dipartimento di statistica, informatica, applicazioni "G. Parenti", Università di Firenze
[2] Società per l'epidemiologia e la prevenzione "GA Maccacaro", Impresa sociale srl, Milano

Corrispondenza: annibale.biggeri@unifi.it

P-value: «Il re è morto, viva il re!»

P-value: «Le roi est mort, vive le roi!»

«Le roi est mort, vive le roi!» era la formula usata per annunciare al popolo la morte del re e la nomina del suo successore, al fine di ribadire la continuità ininterrotta della monarchia.

Questo prima della rivoluzione francese. Così la continuità della pratica di dividere i risultati di uno studio epidemiologico in statisticamente significativi e non si è tramandata nel tempo. A nulla è valso dichiarare scorretto l'uso del p-value rispetto a un livello di significatività prefissato, l'invito a riportare il valore esatto di p, il suggerimento di usare l'intervallo di confidenza, la richiesta di fornire l'intervallo a un livello di confidenza del 90% e così via.

Forse ora siamo arrivati alla rivoluzione francese: qualche testa coronata comincia a cadere nel cestino. Strumentale o meno, l'articolo del 2005 di John Ioannidis "*Why Most Published Research Findings Are False*" ha rivelato un problema di credibilità della ricerca impensabile in passato.[1]

Oggi, a causa del grande numero di lavori sulla riproducibilità della ricerca scientifica, si è generata in una parte non piccola della comunità scientifica la convinzione che si debba cambiare strada. Tagliare le teste, leggi «abbandonare la "significatività statistica"».[2] Già alcune riviste scientifiche, su decisione dei rispettivi direttori scientifici, avevano abbandonato l'uso del livello di significatività per discriminare i risultati di una ricerca, o comunque lo scoraggiavano nelle istruzioni per gli autori. Per quello che più riguarda gli epidemiologi, ricordo i punti di riferimento salienti di questa vicenda: "Confidence Intervals Rather Than P-values: Estimation Rather Than Hypothesis Testing" di Gardner e Altman a metà degli anni Ottanta[3] e il libro sugli intervalli di confidenza tradotto in italiano a cura di Alessandro Liberati per i tipi del Pensiero scientifico nei primi anni Novanta;[4] "Sifting the Evidence. What's Wrong with Significance Tests" dell'allora editor dell'International Journal of Epidemiology, G. Davey-Smith insieme a J. Sterne, nei primi anni Duemila.[5]

La questione è però complessa e l'articolo di **Andrea Saltelli** in questo numero di EpiChange (pp. 121-122) mostra l'orizzonte entro cui va inquadrata, orizzonte nel quale l'integrità morale gioca un ruolo fondamentale. Alcune considerazioni in questa direzione:

1. la prima: il fatto che la proposta di abbandonare il livello di significatività venga presa in considerazione quando diviene grande il numero di ricercatori rivela che si tratta di una proposta interna alla comunità scientifica, volta a recuperare una credibilità sociale compromessa. Se questo è lo scopo, la soluzione deve essere per forza trovata nell'interazione con la società nel suo complesso;

2. la seconda: l'insufficienza di questa soluzione si manifesta anche per la seconda considerazione, cioè che la proposta ha trovato sostenitori e detrattori che si appoggiano su argomentazioni extra-scientifiche relative all'impatto che le diverse soluzioni avrebbero in termini di falsi positivi e falsi negativi e soggettività o arbitrarietà delle decisioni prese dai ricercatori in merito all'importanza dei loro risultati.[6] Si apre la porta

epichange

agli interessi in conflitto nella società e si pretende di coprirli con il velo della soluzione tecnica.

Come *Epidemiologia&Prevenzione* ci eravamo dati delle linee guida, ovviamente da prendere con giudizio:

1. classificare un risultato come "statisticamente significativo" è scorretto;

2. i risultati vanno riportati come stima puntuale e intervallare, scegliendo un livello del 90%;

3. le analisi di sottogruppo vanno giustificate sulla base di un test statistico per l'interazione, riportando ove appropriato il valore esatto di p;

4. i risultati vanno discussi alla luce delle distorsioni che minacciano la validità dello studio;

5. la dimensione dello studio va calcolata alla luce dell'ampiezza desiderata della stima intervallare;

6. regole specifiche valgono per gli studi sperimentali, dove è possibile indicare a priori la differenza clinicamente rilevante e controllare la probabilità di errore di secondo tipo (falsi negativi), la cui giustificazione si trova in alcuni articoli divulgativi pubblicati sulla nostra rivista intorno al 2010-2011.[7-11]

BIBLIOGRAFIA

1. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med 2005;2(8):e124.
2. Amrhein V, Greenland S, McShane BB. Retire statistical significance. Nature 2019;567:305-07.
3. Gardner MJ, Altman DG. Confidence Intervals Rather Than Pvalues: Estimation Rather Than Hypothesis Testing. Br Med J (Clin Res Ed) 1986;292(6522):746-50.
4. Gardner MJ, Altman DG. Gli intervalli di confidenza - Oltre la significatività statistica. Roma, Il Pensiero Scientifico Editore, 1990.
5. Stern JAC, Smith DG. Sifting the Evidence. What's Wrong with Significance Tests? BMJ 2001;322:226-31.
6. Ioannidis. Retiring statistical significance would give bias a free pass. Nature 2019;567:461.
7. Catelan D, Biggeri A, Barbone F. Epidemiol Prev 2011;35(5-6):358-61.
8. Catelan D, Biggeri A, Barbone F. Epidemiol Prev 2011;35(3-4):236-40.
9. Catelan D, Biggeri A, Barbone F. Epidemiol Prev 2011;35(2):150-54.
10. Biggeri A, Catelan D, Barbone F. Epidemiol Prev 2011;35(1):51-52.
11. Barbone F, Biggeri A, Catelan D. Reporting uncertainty. Epidemiol Prev 2010;34(5-6):91-95.

Smettiamola di fingere: quantificare non è un'operazione neutrale

Drop the act: quantification is never neutral

Andrea Saltelli[1,2]

[1] Centre for the Study of the Sciences and the Humanities, University of Bergen (Norway)
[2] Open Evidence Research, Universitat Oberta de Catalunya, Barcelona (Spain)

Corrispondenza: andrea.saltelli@uib.no

Ogni quantificazione che prescinda o che non specifichi in che contesto e a che fine viene prodotta, fatalmente oscura, piuttosto che illuminare. Nell'esercizio di qualsiasi attività di quantificazione, la metodologia non è mai neutrale, dato che non è mai interamente possibile separare l'esercizio di quantificazione dai desideri e dalle aspettative di chi quantifica. Paradossalmente però la forza argomentativa, o retorica, di una quantificazione risiede esattamente nella sua presupposta obiettività e neutralità: «i numeri parlano chiaro», e «il modello non sbaglia» sono espressioni di uso corrente.

Benché il dubbio sulla neutralità della quantificazione sia oggi molto praticato in relazione all'uso di algoritmi per prendere decisioni quali promuovere o bocciare, imprigionare o lasciare in libertà, concedere o non concedere credito,[1] il dubbio è generale. La convergenza dei fenomeni big data e intelligenza artificiale rendono le frontiere fra diversi tipi di quantificazione molto permeabili. Il nascente campo della sociologia della quantificazione si chiede:[2] «Quali qualità sono specifiche di un ranking, di un indicatore, di un modello o di un algoritmo?» In realtà, questa nuova consapevolezza del fatto che ogni numero presupponga una narrazione, una visione del mondo e un possibile obiettivo da conseguire, non riguarda solo l'esempio degli algoritmi, ma si è manifestata in modo molto evidente nel campo della statistica applicata. Qui, l'esistenza di una vera e propria disciplina – la statistica appunto – ha reso la crisi nell'uso e l'abuso dei metodi al tempo stesso meglio compresa e più visibile. Le diagnosi legate all'impiego scorretto dei test di significatività, nei suoi aspetti metodologici e normativi, nutrono le attuali letture degli addetti ai lavori.[3,4]

«Nell'esercizio di qualsiasi attività di quantificazione, la metodologia non è mai neutrale, dato che non è mai interamente possibile separare l'esercizio di quantificazione dai desideri e dalle aspettative di chi quantifica».

Un esempio lampante dell'intreccio esistente fra tecnica e valori è proprio offerto dalla discussione sulla significatività:[5] non sarebbe meglio abolire del tutto il concetto, sostengono alcuni autori,[4] che spingono la loro iniziativa fino a sollecitare il pubblico supporto per questa abolizione, suscitando in molti la domanda legittima: «È appropriato raccogliere centinaia di firme in supporto di un editoriale scientifico?». Per alcuni commentatori, le questioni scientifiche vanno risolte su publicazioni scientifiche, non con petizioni. Altri notano che numerosi articoli scientifici dedicati a illustrare il cattivo uso del concetto di significatività lungo l'arco di più di tre decenni non hanno portato a nessun cambiamento, da cui la necessità di un'azione collettiva di propaganda per affrontare quello che sembra essere un problema di sociologia della scienza.

Epidemiol Prev 2019; 43 (2-3):121-123. doi: 10.19191/EP19.2-3.P121.046