RIVISTA INTERNAZIONALE DI FILOSOFIA E PSICOLOGIA
DOI: 10.4453/rifp.2022.0005
ISSN 2039-4667; E-ISSN 2239-2629
Vol. 13 (2022), n. 1, pp. 57-65

RICERCHE

Fragilità, credibilità e controfattuale

Enrico Ripamonti[(α)], Piero Quatto[(α)] & Donata Marasini[(α)]

Ricevuto: 7 luglio 2021; accettato: 19 novembre 2021

Riassunto Nell'ultimo decennio il *p*-value è stato sottoposto a notevoli critiche soprattutto per l'uso che se ne fa per raggiungere una conclusione dicotomica circa la significatività del risultato sperimentale (significativo o non significativo). Pertanto, da una parte il *p*-value è stato sostituito con approcci differenti, dall'altra è stato affiancato da alcune procedure diagnostiche, tra le quali figurano la fragilità e la credibilità, che hanno il compito di rafforzare o meno la conclusione. La fragilità e l'indice che la misura presentano aspetti di debolezza metodologica. D'altro canto, l'indice di credibilità sembra idoneo per dare o meno supporto alla conclusione e per rafforzare o sostituire l'indice di fragilità, dato che misura la credibilità del risultato osservato quantificando l'informazione *a priori* necessaria per ribaltare il risultato stesso. Il particolare meccanismo delle due procedure, che si fonda su quanto dovrebbe accadere per cambiare la conclusione, suggerisce di inserire le medesime nella prospettiva controfattuale considerandole come nuovi strumenti per la sua misura quantitativa. In questo contributo si presenta questa prospettiva, con particolare riferimento al campo applicativo delle scienze psicologiche.
PAROLE CHIAVE: *p*-value; Indice di fragilità; Distribuzioni a priori; Indice di credibilità; Prospettiva controfattuale

Abstract *Fragility, credibility and counterfactuality* – In the last decade, scientific reliance on *p*-values, especially when used to determine in a dichotomic manner whether a scientific result is significant or not, has been strongly criticized. As a consequence, *p*-values are sometimes replaced by other statistical tools, or supplemented by complementary procedures such as tests for fragility and credibility, which lend further support or challenge the conclusion. The fragility index presents some methodological weaknesses of its own. The credibility index proposed in the literature seems to provide a particularly useful supplement for *p*-values as well as for the fragility index, considering that it assesses the reliability of the result obtained by quantifying the *a priori* information needed to overturn the result. Both procedures rely on what would need to happen in order to modify the conclusion. This suggests that they can be considered as valuable new tools for quantitative measurement within a counterfactual framework. In our contribution we present this perspective, with reference to the psychological sciences.
KEYWORDS: *p*-value; Fragility Index; Priors/Posteriors; Credibility Index; Counterfactual Perspective

[(α)]Dipartimento di Economia, Metodi Quantitativi e Strategie di Impresa, Università degli Studi di Milano-Bicocca, via Bicocca degli Arcimboldi, 8 – 20126 Milano (Italia)

E-mail: enrico.ripamonti@unimib.it (✉); piero.quatto@unimib.it; donata.marasini@unimib.it

1 Introduzione

COME È NOTO NEL CAMPO dell'inferenza statistica, negli ultimi anni il *p*-value, elemento fondamentale dell'approccio NHST (*Null Hypothesis Significance Testing*), è stato sottoposto ad aspre critiche anche per la regola che porta alla conclusione dicotomica: rifiutare o accettare una ipotesi prestabilita secondo che il risultato sperimentale sia significativo o non significativo. Nell'ultimo quinquennio si percepisce che è arrivato il momento di allontanarsi da NHST o, quantomeno, di riadattarlo.[1] Infatti, nel 2018 è pubblicato un articolo a firma di 72 Autori che propone di abbassare la soglia del *p*-value da 0.05 a 0.005[2] e, come contrapposto, esce subito dopo un articolo a firma di almeno 800 autori[3] che invitano esplicitamente al ritiro della significatività statistica.

Una sintesi molto importante della diatriba si trova nel numero speciale della rivista *The American Statistician* del 2019 che, sotto l'egida della *American Statistical Association* (ASA), raccoglie un insieme di 43 contributi dal titolo *Statistical Inference in the 21st Century: A World Beyond p < 0.05*. I lavori riprendono le critiche alla procedura NHST e propongono soluzioni alternative, tutte con l'esplicito accordo di rifiutare la regola come strumento unico per fare inferenza.

Tra gli studiosi, c'è una parte molto consistente che affianca il *p*-value con strumenti di supporto e un'altra che cerca strade alternative al *p*-value stesso. Al primo gruppo appartengono, tra gli altri, Greenland,[4] che riprende l'*s*-value, proposto nel 2017, che è il logaritmo in base 2 del *p*-value e lo impiega come misura descrittiva; Betensky[5] che confronta il *p*-value con una soglia che tiene conto della numerosità campionaria e della misura dell'effetto, per esempio, l'effetto di un trattamento; Fraser[6] che lo inserisce come caso particolare della *p*-value function, già proposta nel 2018, che l'Autore ritiene uno strumento molto generale per trarre inferenza dai dati. Tra gli studiosi appartenenti al secondo gruppo orientati verso l'approccio bayesiano figurano Blume e colleghi[7] che ripropongono il *p*-value di seconda generazione che si basa sugli intervalli di confidenza, ma sconfina in ambito bayesiano con il calcolo delle probabilità delle ipotesi da verificare sia *a priori* sia a esperimento realizzato; Gannon e colleghi[8] che propongono un particolare *p*-value basato sul fattore di Bayes. Mentre nel numero speciale del 2019 si cercano soluzioni alternative, l'anno successivo viene pubblicato un articolo[9] che drasticamente sostituisce il *p*-value con il fattore di Bayes, tra l'altro, nella versione rivisitata di Jeffreys degli anni '30.

In termini di strumenti di supporto al *p*-value, vi sono due proposte molto interessanti antecedenti al 2019, ovvero quella basata sulla fragilità della conclusione e quella basata sulla credibilità della medesima. La prima, già nota nel 1990,[10] ripresa e modificata da Walsh e colleghi[11] e aggiornata di recente,[12] ha la funzione di affiancare il *p*-value sperimentale utilizzando un secondo *p*-value ipotetico. La seconda, introdotta da Matthews nel 2001 (ripresa successivamente dallo stesso Matthews nel 2018, per poi riproporla nel 2019),[13] approfondita da Held,[14] da Quatto e colleghi[15] e da Held e colleghi,[16] ha la funzione di supportare il *p*-value sperimentale impiegando strumenti bayesiani come le distribuzioni *a priori* e *a posteriori*.

Fragilità e credibilità hanno un obiettivo comune che è quello di verificare se la conclusione conseguente a un particolare risultato sperimentale e riassunta nel *p*-value possa ritenersi robusta, nel senso che può essere confermata, ovviamente sempre in termini probabilistici. Anche se l'analisi della fragilità ha ricevuto numerose critiche, l'ultima delle quali è del 2020,[17] e la seconda presenta una sorta di anomalia, si ritengono entrambe validi strumenti nel campo dell'inferenza. Pertanto, nel presente lavoro, si ripropongono entrambe, ritenendo la prima uno strumento che contribuisce a fare luce sul risultato sperimentale e la seconda uno strumento che rafforza, o meno, il *p*-value. Tra l'altro, nelle righe che seguono, l'analisi della credibilità viene accompagnata da un nuovo indice che ha la prerogativa di sintetizzare in un'unica formula la significatività e la non significatività.

Poiché l'analisi della fragilità e quella della credibilità richiamano la prospettiva controfattuale, studiata e approfondita sotto il termine di *potential outcome model* da Rubin[18] e da Rosenbaum,[19] le medesime analisi vengono riproposte anche come nuovi strumenti all'interno dell'approccio controfattuale.

Nella sezione 2 si introduce l'indice di fragilità, nella sezione 3 si introduce la logica della credibilità e il nuovo indice di credibilità. Nella sezione 4, dopo avere introdotto brevemente la prospettiva controfattuale, si verifica come sia possibile inserire fragilità e credibilità nella medesima. La sezione 5 conclude la rassegna. Viene inserita un'appendice che può essere di supporto per la comprensione di alcune formule.

2 L'indice di fragilità

La fragilità può essere considerata come una misura della robustezza di una procedura inferenziale, nel senso che piccole variazioni nei dati possono sovvertire il valore del *p*-value, portandolo da significativo ($p < 0.05$) a non significativo ($p \geq 0.05$) o viceversa. La misura si identifica con l'indice di fragilità (FI) che è dato dal più piccolo numero di eventi che portano a una variazione del *p*-value da significativo a non significativo o viceversa. Tanto più piccolo è l'indice tanto più fragile è il risultato[20] o anche tanto maggiore è FI tanti più eventi occorrerebbero per il cambiamento di stato denotando così un risultato robusto.

Per illustrare la fragilità, si consideri il seguente esempio. In uno studio clinico randomizzato 100 pazienti sono stati assegnati al gruppo di trattamento che prevede la somministrazione di un nuovo farmaco per prevenire l'infarto e altrettanti pazienti nel gruppo di controllo. Tra i trattati 10 pazienti hanno avuto un infarto e 22 nel gruppo di controllo. Se con θ si indica l'effetto del trattamento, l'ipotesi nulla soggetta a verifica è H_0: $\theta = 0$, ossia il farmaco non ha effetto, in contrapposizione a H_1: $\theta \neq 0$, ossia il farmaco ha effetto. Ricorrendo alla differenza $\hat{\theta}$ di proporzioni di infartuati nel trattamento (0.10) e nel controllo (0.22), supposta distribuita come una variabile casuale Normale, si è trovato $p = 0.025$ che essendo inferiore alla soglia $\alpha = 0.05$ fa concludere per la significatività, ossia il farmaco si ritiene efficace nella prevenzione degli infarti. Se nel gruppo dei trattati ci fossero stati due infartuati in più, ricorrendo ancora alla differenza di proporzioni $\hat{\theta}$, si sarebbe trovato $p = 0.057$ che essendo maggiore della soglia avrebbe fatto concludere per la non significatività del risultato, ossia il nuovo farmaco non ha effetto. FI = 2 denota la fragilità del risultato, dal momento che solo due eventi sarebbero in grado di sovvertire la significatività.

Un limite di FI sta nel fatto che può essere applicato solo a dati dicotomici (guarito, non guarito; laureato non laureato; assicurato, non assicurato), ovvero principalmente all'analisi *ex post* di esperimenti randomizzati. In secondo luogo, una delle critiche più frequenti, alla quale è difficile controbattere, è la mancanza di una soglia al di sotto della quale il risultato può ritenersi fragile. Se il risultato dell'esempio precedente avesse richiesto 5 eventi (infartuati) per passare alla non significatività, il valore (FI = 5) che ha ribaltato il risultato sarebbe sufficientemente piccolo? Un'ulteriore criticità riguarda il fatto che FI cresce al crescere dell'ampiezza campionaria *n* segnalando, ad esempio, un risultato robusto, ossia non fragile, solo perché il campione è ampio. Tale debolezza può essere però arginata ricorrendo all'indice relativo FQ = FI/n (quoziente di fragilità), rimuovendo cioè l'influenza della numerosità del gruppo.[21] Tuttavia, l'indice relativo è una soluzione a questo problema solo nel confronto tra più situazioni con diverse numerosità.

3 L'analisi della credibilità

La credibilità ha l'obiettivo di rafforzare un risultato ottenuto tramite il *p*-value (o mostrarne la debolezza) impiegando una procedura bayesiana.

In particolare, l'analisi della credibilità si basa sul fatto che un risultato significativo o non significativo può potenzialmente essere trasformato nel risultato opposto (rispettivamente, non significativo o significativo), scegliendo un'opportuna distribuzione *a priori* nell'ambito dell'inferenza bayesiana.

Seguendo le premesse di Matthews,[22] si supponga che ogni distribuzione di probabilità con cui si lavora sia (almeno approssimativamente) Normale, come la distribuzione della statistica $\hat{\theta}$ utilizzata come stima per il parametro θ nell'approccio NHST, la distribuzione *a priori* che sintetizza tutte le possibili informazioni *a priori* su θ, punto di partenza dell'inferenza bayesiana e la distribuzione *a posteriori*, ottenuta aggiornando l'*a priori* sulla base del risultato sperimentale, anch'essa strumento di inferenza bayesiana.

In tale contesto, ovvero sotto lo specifico assunto di Normalità, l'analisi della credibilità permette di scegliere un'opportuna distribuzione *a priori* in grado di ribaltare il risultato osservato (*reverse Bayes approach*) che ha portato alla significatività, ossia il trattamento ha avuto effetto, o alla non significatività, ossia il trattamento non ha avuto effetto.

Si supponga il caso della significatività, che produce $p < \alpha$, dove: la prefissata soglia è $\alpha = 2P(Z > z_{1-\alpha/2})$, Z è la variabile casuale Normale standardizzata, ossia Z $\sim$ N(0,1), $z_{1-\alpha/2}$ è il percentile di ordine $1 - \alpha/2$, $p = 2P(Z > z)$ e $z = \hat{\theta}/\sigma_{\hat{\theta}}$ è il risultato sperimentale (stima di θ) opportunamente standardizzato tramite la radice della varianza dello stimatore stesso $\sigma^2_{\hat{\theta}}$, ossia l'errore standard. Dal fatto che $p < \alpha$ segue $|z| > z_{1-\alpha/2}$.

In questo caso, la distribuzione *a priori* viene supposta con media $\theta_a = 0$ che avalla l'ipotesi nulla e tanto più è concentrata intorno alla sua media, ossia tanto più la sua varianza σ^2_a è piccola, tanto più è in contrasto con la significatività che vuole $\theta \neq 0$. Si supponga di costruire sulla distribuzione *a posteriori* un intervallo a livello $1 - \alpha$ con un estremo in 0, così da rafforzare ulteriormente l'ipotesi $\theta = 0$, antagonista del risultato ottenuto. Tenendo presente le relazioni esistenti tra medie e varianze delle tre distribuzioni (cfr. *Appendice A1*), è possibile verificare (cfr. *Appendice A2*) che

$$\sigma^2_a = \frac{z^2_{1-\alpha/2}\sigma^2_{\hat{\theta}}}{z^2 - z^2_{1-\alpha/2}}. \qquad (1)$$

Se sperimentalmente è accaduto che $p \geq \alpha$, cioè il risultato è non significativo, allora per caratterizzare la distribuzione *a priori*, secondo Matthews, ferma restando la richiesta che un estremo dell'intervallo costruito sulla distribuzione *a posteriori* sia nullo, si chiede che sia nullo anche l'estremo dell'intervallo costruito sulla distribuzione *a priori* a livello 1-α. In altri termini, la distribuzione *a priori* e quella *a posteriori*, pur favorevoli all'alternativa, non falsificano completamente l'ipotesi nulla. Nel caso in esame si verifica (cfr. *Appendice A3*) che

$$\theta_a = \frac{2z\,\sigma_{\hat{\theta}}\,z^2_{1-\alpha/2}}{z^2_{1-\alpha/2} - z^2} \qquad (2)$$

dove il denominatore è positivo dal momento che la

non significatività garantisce $|z| \leq z_{1-\alpha/2}$ e

$$\sigma_a^2 = \frac{4(\sigma_{\hat{\theta}}^2)^2 z^2 z_{1-\alpha/2}^2}{(z^2 - z_{1-\frac{\alpha}{2}}^2)^2}. \qquad (3)$$

Secondo Matthews, se il risultato sperimentale è esterno all'intervallo costruito sulla distribuzione *a priori* si conferma la conclusione di non significatività. Così operando, la procedura di Matthews presenta un inconveniente nel senso che, se detto intervallo è del tipo (0, A), costruito supponendo come alternativa $\theta > 0$, la stima $\hat{\theta} > 0$ è certamente compresa e altrettanto accade se l'intervallo è (-A,0) e $\hat{\theta} < 0$, legando in tal modo la credibilità solo al segno di $\hat{\theta}$.

Al fine di quantificare la credibilità del risultato osservato e mantenendo le premesse sulle distribuzioni *a priori* e *a posteriori*, si propone il seguente indice di credibilità che confronta le due varianze $\sigma_{\hat{\theta}}^2$ e σ_a^2 tramite il rapporto

$$CI = \frac{\sigma_{\hat{\theta}}^2}{\sigma_a^2}. \qquad (4)$$

L'indice (4), che varia nell'intervallo (0,∞), misura l'informatività della distribuzione *a priori* rispetto a quella della statistica impiegata, di modo che tanto più la distribuzione *a priori* è concentrata intorno alla propria media tanto più è informativa nei confronti dell'ipotesi che sta difendendo e quindi tanto più è in antitesi con il risultato ottenuto che è il suo opposto. Una soglia ragionevole per decretare la credibilità del risultato sembra essere l'unità, corrispondente all'uguaglianza delle varianze $\sigma_{\hat{\theta}}^2$ e σ_a^2. Pertanto se accade che CI ≥ 1, il risultato ottenuto sperimentalmente viene ritenuto stabile (credibile).

In particolare, ricordando la (1), nel caso della significatività l'indice assume la forma

$$\mathrm{CI} = \frac{z^2 - z_{1-\alpha/2}^2}{z_{1-\alpha/2}^2} \qquad (5)$$

e ricordando la (3)

$$\mathrm{CI} = \frac{(z_{1-\alpha/2}^2 - z^2)^2}{4z^2 z_{1-\alpha/2}^2} \qquad (6)$$

nel caso della non significatività. Conviene osservare che se CI > 1, dalla (5) si ottiene la cosiddetta credibilità intrinseca $|z| > \sqrt{2}\, z_{1-\alpha/2}$,[23] che è una condizione più restrittiva della precedente $|z| > z_{1-\alpha/2}$, e dalla (6) si ottiene $|z| \leq (\sqrt{2} - 1) z_{1-\alpha/2}$, anch'essa condizione più restrittiva di $|z| \leq z_{1-\alpha/2}$. In altri termini, la credibilità del risultato conseguito richiede condizioni più stringenti rispetto a quanto accade con il *p*-value che, tramite il risultato sperimentale, ha portato a una conclusione e questa richiesta sembra dare solidità alla procedura.

Si consideri il seguente esempio che riguarda un esperimento fittizio sul tema del processamento subconscio delle informazioni (*abstract unconscious fear processing*).[24] Secondo questa teoria ci sono stimoli che a livello subconscio producono come risposta un elevato livello di paura che può essere rilevato esaminando la attività del cervello. I dati impiegati in questo contesto sono altrettanto fittizi. Un team di psicologi valuta un gruppo di 25 soggetti esposti agli stimoli e un altro gruppo di 25 soggetti che costituisce il controllo. L'ipotesi da verificare è l'assenza di effetti degli stimoli ricevuti, ossia $\theta = 0$ a livello $\alpha = 0.05$, supponendo come ipotesi alternativa $\theta \neq 0$. Sia $\hat{\theta} = 1.98$ la differenza riscontrata con opportuna misura tra i soggetti esposti agli stimoli e quelli di controllo e sia $\sigma_{\hat{\theta}}^2 = 0.5$ la corrispondente varianza. Risulta $z = 1.98/\sqrt{0.5} = 2.80$ di modo che si ha p = $2P(Z > 2.80) = 0.005 < 0.05$, potendosi concludere sulla significatività. Per verificare la stabilità di questa conclusione, si può calcolare la (5) che nel caso in esame porta a CI = 1.04 che conferma la significatività.

4 La prospettiva controfattuale

Con il termine controfattuale si intende l'analisi di che cosa sarebbe accaduto se si fosse verificata una circostanza diversa da quella realmente accaduta. Quest'ultima è una circostanza fattuale mentre la prima, che è solo potenziale, è detta controfattuale. Il controfattuale può sintetizzarsi in "*what-if*" ("che cosa sarebbe accaduto se").[25]

Senza entrare nell'aspetto filosofico del controfattuale che chiama in causa l'analisi causale,[26] nel presente contesto il controfattuale viene declinato in due modi diversi, come emerge dalle due proposizioni: (a) se Gianni si fosse laureato, avrebbe certamente raggiunto una posizione apicale e (b) se Gianni si fosse laureato, avrebbe potuto raggiungere una posizione apicale.

Gianni non si è laureato come emerge da (a) e da (b) e questa circostanza è diversa da quella prevista nel congiuntivo (se si fosse) ed è cioè l'aspetto fattuale della proposizione. L'aspetto controfattuale nel condizionale (avrebbe), sintetizza in (a) il raggiungimento di una posizione apicale e in (b) un possibile raggiungimento. In altri termini, nel primo caso il controfattuale viene ritenuto *certo*, mentre nel secondo come soltanto *possibile*. I due tipi di controfattuale prendono i nomi di "*would-counterfactual*", che rappresenta il controfattuale standard, e, rispettivamente, "*might-counterfactual*" che nel condizionale impiega l'ausiliare potere.[27]

Esempi concreti di analisi controfattuale si trovano in molti campi di ricerca quando l'attenzione è rivolta all'effetto di un trattamento. Nel campo clinico per verificare l'effetto di un particolare

farmaco su determinati pazienti l'aspetto controfattuale è «che cosa sarebbe accaduto al paziente nel caso che non fosse stato trattato»; nel campo politico per verificare le conseguenze di una legge «che cosa sarebbe accaduto se non fosse entrata in vigore»; nel campo sociale per verificare l'effetto della laurea sul reddito da lavoro di un individuo «quale sarebbe il reddito di quel soggetto se non si fosse laureato»; nel campo economico per verificare l'effetto di un intervento bancario «che cosa sarebbe accaduto al mercato immobiliare se la banca centrale avesse mantenuto un tasso d'interesse più alto»; nel campo psicologico «quale sarebbe stato il livello di paura di un soggetto se non fosse stato sottoposto a stimoli».

In psicologia il pensiero controfattuale gioca un ruolo importante "per la mente, nel prendere decisioni, nel plasmare le emozioni".[28] Il pensiero controfattuale può aiutare a porre ipotesi riguardanti il passato: «se avessi avuto più tempo, se avessi saputo», oppure può essere di sostegno nel prendere decisioni per il futuro "se non si fosse soddisfatti, per esempio, di un servizio, si potrebbe cambiare il fornitore".[29] Ci si trova quotidianamente ad avere a che fare con la prospettiva controfattuale: «che cosa sarebbe successo se un individuo avesse frequentato un'altra scuola, se non si fosse sposato, se non avesse fatto la carriera attuale?».

Se la prospettiva fattuale può spiegarsi o misurarsi ricorrendo a fatti concreti, quella controfattuale deve rifarsi a fatti potenziali. Così, con riguardo alla misura di un effetto, la prospettiva fattuale potrebbe misurare l'effetto causale della laurea sul reddito da lavoro come la differenza tra reddito percepito dai laureati e dai non laureati, mentre la prospettiva controfattuale dovrebbe misurare l'effetto causale della laurea come la differenza tra reddito percepito dai laureati e reddito che essi stessi avrebbero percepito se non fossero giunti a laurearsi.[30]

Il problema è per l'appunto la misura del controfattuale. Laddove sono possibili esperimenti che prevedono la suddivisione randomizzata di soggetti tra trattamento e controllo, la procedura è relativamente semplice. Infatti il soggetto sottoposto al controllo è ritenuto simile a quello sottoposto al trattamento e, pertanto, l'insieme dei primi costituisce il controfattuale dei secondi a patto che la numerosità dei due gruppi sia la medesima.

La difficoltà sorge quando non è realizzabile uno studio sperimentale, così come accade, ad esempio, nel verificare l'effetto di una politica pubblica, dal momento che bisogna ricostruire il dato controfattuale. Ci si chiede allora perché in queste circostanze sia opportuno riferirsi alla prospettiva controfattuale. La risposta è che gli effetti osservati, ad esempio sempre con riguardo alla politica, potrebbero non essere dovuti all'attuazione della politica ma ad altri fenomeni che potrebbero portare a sovrastime o sottostime del contributo della politica in questione. Volendo ricostruire il controfattuale mediante metodi quantitativi, esistono diverse tecniche come la regressione che, sotto particolari assunzioni, consente di stimare l'effetto del trattamento a parità di altre condizioni, avendo selezionato opportune variabili osservabili, lo *statistical matching* che consiste nel creare *a posteriori* un gruppo di unità che costituiscono una sorta di controllo il più possibile simili alle unità trattate,[31] l'analisi dei grafi[32] e altre procedure[33] che non è necessario approfondire dal momento che, nel prosieguo, l'aspetto controfattuale viene trattato in modo del tutto originale rispetto a quanto avviene solitamente.

4.1 Fragilità e credibilità nella prospettiva controfattuale

La fragilità e la credibilità tentano di rispondere alla domanda «che cosa dovrebbe accadere (senza che sia accaduto) perché la conclusione accettata a seguito del risultato sperimentale (accaduto) possa essere ribaltata?»

Le risposte sono leggermente diverse tra fragilità e credibilità, così come si è accennato nell'esempio riguardante la laurea di Gianni. Per la fragilità, se si fossero verificati x eventi in più (o in meno) la conclusione sarebbe stata l'opposto di quella ottenuta dal risultato sperimentale, ossia il controfattuale è ritenuto certo.

Nel caso della credibilità, se si fosse adottata una distribuzione *a priori* contrastante con il risultato sperimentale la conclusione avrebbe potuto essere opposta a quella sperimentale e il controfattuale è solo possibile, perché fondato su una distribuzione di probabilità.

Per la fragilità si impiega l'indice FI per quantificare il controfattuale e in base al suo livello si valuta la sua *robustezza*: con x eventi in più o in meno il risultato da significativo si è trasformato in non significativo o viceversa, cioè il controfattuale contraddice il fattuale. Resta però la domanda se siano sufficienti gli x eventi per ritenere il risultato fragile.

Per la credibilità si impiega l'indice C per quantificare il controfattuale e in base al fatto che superi o meno la soglia prefissata si valuta la sua robustezza: se supera la soglia si ritiene che il controfattuale confermi il fattuale e lo contraddica nel caso sia inferiore all'unità.

La fragilità sembra entrare in modo naturale nel controfattuale dal momento che si riferisce a un esperimento randomizzato che, nel caso di uguale numerosità tra trattamento e controllo, mette alla pari il gruppo dei trattati con il gruppo di controllo nel senso che ogni unità sottoposta a trattamento ha il suo *clone* nel controllo, pertanto l'analisi del controfattuale potrebbe arrestarsi nel momento in cui si giunge alla conclusione. Tuttavia, in questo contesto, viene fatto un passo avanti intendendo il controfattuale come una verifica teorica della con-

clusione secondo la logica dell'esperimento mentale o esperimento immaginario.[34]

L'inserimento della credibilità nell'ambito controfattuale è meno naturale se si fa riferimento al fatto che non necessita di esperimenti randomizzati, ma può rientrare ugualmente sotto forma della domanda: «quale distribuzione *a priori* avrebbe potuto ribaltare il risultato fornito dal *p*-value?». Anche in questo caso il controfattuale può farsi coincidere con la verifica teorica della conclusione.

Le due prospettive potrebbero vedersi inserite nell'ambito del trattamento e del controllo se il primo venisse identificato con la procedura inferenziale NHST che ha fornito un risultato sperimentale che ha consentito una conclusione se pure probabilistica e il secondo con la procedura di verifica della conclusione, altrettanto probabilistica, che utilizza i dati osservati con il trattamento impiegando altre ipotesi che consentono il controllo controfattuale.

L'inserimento delle due metodologie nella prospettiva controfattuale consente di affermare che, oltre alle usuali tecniche di analisi di cui si è detto in precedenza, ne esistono altre che possono essere di aiuto nel risolvere situazioni ipotizzate in contrapposizione a quanto accaduto realmente.

5 Conclusione

Come si è già precisato l'indice di fragilità ha subito molte critiche, anche se continua a essere utilizzato come dimostrano i numerosi articoli; infatti solo nell'ultimo biennio (2019-2020) sono stati pubblicati almeno 15 lavori in cui si fa riferimento all'indice di fragilità in molti campi della ricerca medica, come l'oncologia, la cardiologia, l'oftalmologia, l'anestesiologia.

In altri campi di ricerca, come quello sociale o quello psicologico, non si trovano in letteratura riferimenti sull'impiego di FI e altrettanto accade per l'analisi della credibilità, comunque meno nota della fragilità. Il mancato ricorso alla fragilità può essere dovuto alla difficoltà di impostare esperimenti randomizzati (pur non assenti nelle discipline economico-sociali), mentre per la credibilità potrebbe essere il fatto che si basa sulla procedura bayesiana.

Quest'ultima, vista con diffidenza nella ricerca clinica tradizionale, solo in tempi relativamente recenti è stata introdotta nell'ambito delle scienze sociali e nella ricerca psicologica, dove oggi viene sempre più spesso utilizzata, come emerge consultando gli indici di riviste specializzate quali *Psychometrika*, *Journal of Mathematical Psychology*, *Psychological Methods* o *Behavioral Research Methods*. Nel presente contesto si può ricordare l'articolo di Scott e colleghi,[35] che presenta una rassegna dell'analisi bayesiana in sociologia, e, per quanto riguarda la psicologia, i contributi di Kruschke e colleghi,[36] di Wagenmakers e colleghi,[37] e quello di Heck e colleghi.[38] Un volume introduttivo all'analisi bayesiana nelle discipline psicologiche è quello curato da Kruschke.[39]

La credibilità e il conseguente indice, come già precisato, si basano su un assunto bayesiano chiamando in causa distribuzioni *a priori* e *a posteriori*. La sostanziale differenza, per esempio con il fattore di Bayes, fulcro dell'approccio bayesiano per la verifica di ipotesi, sta nel fatto che nell'analisi della credibilità la distribuzione *a priori* non rappresenta una sintesi di informazioni preesistenti e/o soggettive, ma è determinata esclusivamente in funzione dell'obiettivo di favorire l'ipotesi nulla, nel caso di significatività, o l'ipotesi alternativa, nel caso di non significatività.

Per esempio, nel caso di significatività, assunta la distribuzione Normale, la sua media è coerente con l'ipotesi nulla e la sua varianza viene quantificata sfruttando i dati sperimentali.

In particolare, l'indice proposto intende misurare la credibilità del risultato osservato in termini della concentrazione della distribuzione *a priori* necessaria per sovvertire il risultato stesso.

Si ritiene, inoltre, che fragilità e credibilità siano riconducibili a una prospettiva controfattuale, così come possano intendersi strumenti idonei per affrontare le due così dette antimetabole: assenza di evidenza ed evidenza dell'assenza,[40] nel caso della non significatività, e presenza di evidenza ed evidenza della presenza, nel caso della significatività. Con la prima si intende che si può concludere sul non effetto anche se quest'ultimo è presente, mentre si conclude sul non effetto per importanti informazioni contenute nel risultato sperimentale. Con la seconda si intende che si può concludere sull'effetto anche se non è presente, mentre si conclude per la presenza dell'effetto stanti importanti informazioni contenute nel risultato sperimentale.

Appendici

A1

Le tre distribuzioni di interesse possono sintetizzarsi in:

$$\begin{aligned} &\hat{\theta}|\theta \sim N(\theta,\ \sigma_{\hat{\theta}}^2),\\ &\theta \sim N(\theta_a, \sigma_a^2),\\ &\theta|\hat{\theta} \sim N(\theta_p, \sigma_p^2), \end{aligned} \qquad (1a)$$

dove $\hat{\theta}$, θ_a, θ_p, $\sigma_{\hat{\theta}}^2$, σ_a^2 e σ_p^2 sono medie e varianze delle medesime. In particolare la prima delle (1a) si riferisce alla distribuzione della statistica (variabile casuale) impiegata nella procedura NHST, la seconda è la distribuzione *a priori* per θ che, nell'approccio bayesiano, differentemente dall'approccio frequentista, perde la natura di parametro e diventa variabile casuale, e la terza è la distribuzione *a posteriori* per θ. Medie e varianze sono legate dalle seguenti relazioni (cfr. A. GELMAN, G.B. CARLIN, H.S. STERN, D.B. DUNSON, A. VEH-

TARI, D.B. RUBIN, *Bayesian data analysis*, Chapman & Hall, New York 2014):

$$\frac{\theta_p}{\sigma_p^2} = \frac{\hat{\theta}}{\sigma_{\hat{\theta}}^2} + \frac{\theta_a}{\sigma_a^2} \qquad \frac{1}{\sigma_p^2} = \frac{1}{\sigma_{\hat{\theta}}^2} + \frac{1}{\sigma_a^2}. \tag{2a}$$

A2

Indicando con $L_p = \theta_p - \sigma_p z_{1-\alpha/2}$ l'estremo inferiore dell'intervallo costruito sulla distribuzione *a posteriori*, dalle ipotesi $\theta_a = 0$ e $L_p = 0$ dalla prima delle (2a) si ricava

$$\frac{z_{1-\alpha/2}}{\sigma_p} = \frac{\hat{\theta}}{\sigma_{\hat{\theta}}^2}$$

ossia

$$\frac{1}{\sigma_p^2} = \left(\frac{\hat{\theta}}{\sigma_{\hat{\theta}}^2}\right)^2 \frac{1}{z_{1-\alpha/2}^2}$$

che sostituito nella seconda delle (2a) comporta

$$\left(\frac{\hat{\theta}}{\sigma_{\hat{\theta}}^2}\right)^2 \frac{1}{z_{1-\alpha/2}^2} = \frac{1}{\sigma_{\hat{\theta}}^2} + \frac{1}{\sigma_a^2}$$

e con semplici passaggi si ottiene la (1).

A3

Indicando con $L_a = \theta_a - \sigma_a z_{1-\alpha/2}$ l'estremo inferiore dell'intervallo costruito sulla distribuzione *a priori*, dalle ipotesi $L_a = 0$ e $L_p = 0$ si ottiene $\theta_a = \sigma_a z_{1-\alpha/2}$ e $\theta_p = \sigma_p z_{1-\alpha/2}$. Dalla prima delle (2a) si ha

$$\frac{z_{1-\alpha/2}}{\sigma_p} = \frac{\hat{\theta}}{\sigma_{\hat{\theta}}^2} + \frac{z_{1-\alpha/2}}{\sigma_a}$$

e dalla seconda

$$\frac{1}{\sigma_a^2} = \frac{1}{\sigma_p^2} - \frac{1}{\sigma_{\hat{\theta}}^2}$$

e quindi

$$\frac{z_{1-\alpha/2}}{\sigma_p} = \frac{\hat{\theta}}{\sigma_{\hat{\theta}}^2} + z_{1-\alpha/2}\sqrt{\frac{1}{\sigma_p^2} - \frac{1}{\sigma_{\hat{\theta}}^2}}.$$

Dopo alcuni passaggi si ottiene

$$\frac{1}{\sigma_p^2} = \frac{\left(\hat{\theta}^2 + \sigma_{\hat{\theta}}^2 z_{1-\alpha/2}^2\right)^2}{\left(2\hat{\theta}\sigma_{\hat{\theta}}^2 z_{1-\alpha/2}\right)^2}$$

e

$$\frac{1}{\sigma_a^2} = \frac{\left(\hat{\theta}^2 + \sigma_{\hat{\theta}}^2 z_{1-\alpha/2}^2\right)^2}{\left(2\hat{\theta}\sigma_{\hat{\theta}}^2 z_{1-\alpha/2}\right)^2} - \frac{1}{\sigma_{\hat{\theta}}^2}.$$

Sviluppando e facendo le opportune semplificazioni si ricavano la (2) e la (3).

Note

[1] Cfr. R.A.J. MATTHEWS, *The p-value statement, five years on*.
[2] Cfr. D.J. BENJAMIN, J.O. BERGER, M. JOHANNESSON, ET ALII, *Redefine statistical significance*.
[3] Cfr. V. AMRHEIN, S. GREENLAND, B. MCSHANE, *Scientists raise up against statistical significance*.
[4] Cfr. S. GREENLAND, *Valid p-values behave exactly as they should*.
[5] Cfr. R.A. BETENSKY, *The p-value requires context, not a threshold*.
[6] Cfr. D.A.S. FRASER, *The p-value function and statistical inference*.
[7] Cfr. J.D. BLUME, R.A. GREEVY, V.F. WELTY, J.R. SMITH, W.D. DUPONT, *An introduction to second-generation p-values*.
[8] Cfr. M.A. GANNON, C.A. DE BRAGANÇA PEREIRA, A. POLPO, *Blending Bayesian and classical tools to define optimal sample-size-dependent significance levels*.
[9] Cfr. A. LY, A. STEFAN, J. VAN DOORN, F. DABLANDER, ET ALII, *The Bayesian methodology of Sir Harold Jeffreys as a practical alternative to the p value hypothesis test*.
[10] Cfr. A.R. FEINSTEIN, *The unit fragility index*.
[11] Cfr. M. WALSH, S.K. SRINATHAN, D.F. MCAULEY, ET ALII, *The fragility of trial results is frequently fragile: A case for a fragility index*.
[12] Cfr. S.D. WALTER, L. THABANE, M. BRIEL, *The fragility of trial results involves more than statistical significance alone*.
[13] Cfr. R.A.J. MATTHEWS, *Methods for assessing the credibility of clinical trial outcomes*; R.A.J. MATTHEWS, *Beyond "significance"*; R.A.J. MATTHEWS, *Moving towards the post p < 0.05 Era via the analysis of credibility*.
[14] Cfr. L. HELD, *The assessment of intrinsic credibility and a new argument for p <0.005*.
[15] Cfr. P. QUATTO, E. RIPAMONTI, D. MARASINI, *Beyond p < 0.05: A critical review of new Bayesian proposal for assessing the p-value*.
[16] Cfr. L. HELD, R. MATTHEWS, M. OTT, S. PAWEL, *Reverse-Bayes methods*.
[17] Cfr. G.E. POTTER, *Dismantling the fragility index*.
[18] Cfr. D.B. RUBIN, *Estimating causal effects of treatments in randomized and nonrandomized studies*; D.B. RUBIN, *Causal inference using potential outcomes*.
[19] Cfr. P.R. ROSENBAUM, *Design of observational studies*, Springer; P.R. ROSENBAUM, D.B. RUBIN, *The central role of the propensity score in observational studies for causal effects*; P.R. ROSENBAUM, D.B. RUBIN, *Assessing sensitivity to an unobserved covariate in an observational study with binary outcome*.
[20] Cfr M. WALSH, S.K. SRINATHAN, D.F. MCAULEY, ET ALII, *The fragility of trial results is frequently fragile*.
[21] Cfr. W. AHMED, R.A. FOWLER, V.A. MCCREDIE, *Does sample size matter when interpreting the fragility index?*.

[22] Cfr. R.A.J Matthews, *Beyond "significance": Principles and practice of the analysis of credibility*.
[23] Cfr. L. Held, *The assessment of intrinsic credibility and a new argument for p <0.005*.
[24] Cfr. E.J. Wagenmaker, J. Verhagen, D. Matzke, et alii, *The need for Bayesian hypothesis testing in psychological science*.
[25] Cfr. S.L. Morgan, C. Winship, *Counterfactuals and causal inference*.
[26] Cfr. D. Lewis, *Causation*.
[27] V. Morato, *Controfattuali*.
[28] N. Roese, *The psychology of counterfactual thinking*.
[29] Cfr. R.M.J. Byrne, *Counterfactual thinking*.
[30] Cfr. M. Lucchini, *Il contributo del modello controfattuale all'irrobustimento della sociologia*.
[31] Cfr. A. Martini, *Metodo sperimentale, approccio controfattuale e valutazione degli effetti delle politiche pubbliche*.
[32] Cfr. L. Pearl, M. Glymour, N.P. Jewell, *Causal inference in statistics*.
[33] Cfr. S.L. Morgan, C. Winship, *Counterfactuals and causal inference*.
[34] Cfr. L. Held, R. Matthews, M. Ott, S. Pawel, *Reverse-Bayes methods*.
[35] Cfr. M.L. Scott, B. Bartlett, *Bayesian statistics in sociology: Past, present, future*.
[36] Cfr. J.K. Kruschke, T.M. Liddell, *Bayesian data analysis for newcomers*.
[37] Cfr. E.J. Wagenmakers, M. Marsman, T. Jamil, et alii, *Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications*.
[38] Cfr. D.W. Heck, U. Boehm, F. Böing-Messing, et alii, *A review of applications of the Bayes factor in psychological research*.
[39] Cfr. J. Kruschke, *Doing Bayesian data analysis*.
[40] Cfr. C. Keyserse, V. Gazzola, E.J. Wagenmakers, *Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence*.

Riferimenti bibliografici

Ahmed, W., Fowler, R.A., McCredie, V.A. (2016). *Does sample size matter when interpreting the fragility index?*. In: «Critical Care Medicine», vol. XLIV, n. 11, pp. 1142-1143.

Amrhein, V., Greenland, S., McShane, B. (2019). *Scientists raise up against statistical significance*. In: «Nature», vol. DLXVII, n. 7748, pp. 305-307.

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmalers, E.J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C.D., Clyde, M., Cook, T.D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A.P., Forster, M., George, E.I., Gonzales, R., Goodman, S., Green, E., Green, D.P., Greenwald, A.G., Hadfield, J.D., Hedges, L.V., Held, L., Ho, T.H., Hojjtink, H., Hruschka, D.J., Imai, K., Imbens, G., Ioannidis, J.P.A., Jeon, M., Jones, J.H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S.E., McCarthy, M., Moore, D.A., Morgan, S.L., Munafo, M., Nakagawa, S., Nyhan, B., Parker, T.H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F.D., Sellke, T., Sinclair, B., Tingley, D., van Zandt, T., Vazire, S., Watts, D.J., Winship, C., Wolpert, R.L., Xie, Y., Young, C., Zinman, J., Johnson, V.E. (2018). *Redefine statistical significance*. In: «Nature Human Behaviour», vol. II, n. 1, pp. 6-10.

Betensky, R.A. (2019). *The p-value requires context, not a threshold*. In: «The American Statistician», vol. LXXIII, Supplement 1, pp. 115-117.

Blume, J.D., Greevy, R.A., Welty, V.F., Smith, J.R., Dupont, W.D. (2019). *An introduction to second-generation p-values*. In: «The American Statistician», vol. LXXIII, Supplement 1, pp. 157-167.

Byrne, R.M.J. (2016). *Counterfactual thinking*. In: «Annual Review of Psychology», vol. LXVII, pp. 135-157.

Feinstein, A.R. (1990). *The unit fragility index: An additional appraisal of "statistical significance" for a contrast of two proportions*. In: «Journal of Clinical Epidemiology», vol. XLIII, n. 9, pp. 201-209.

Fraser, D.A.S. (2019). *The p-value function and statistical inference*. In: «The American Statistician», vol. LXXIII, Supplement 1, pp. 135-147.

Gannon, M.A., de Bragança Pereira, C.A., Polpo, A. (2019). *Blending Bayesian and classical tools to define optimal sample-size-dependent significance levels*. In: «The American Statistician», vol. LXXIII, Supplement 1, pp. 213-222.

Gelman, A., Carlin, G.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. (2014). *Bayesian data analysis*, Chapman & Hall, New York.

Greenland, S. (2019). *Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values*. In: «The American Statistician», vol. LXXIII, Supplement 1, pp. 106-114.

Heck, D.W., Boehm, U., Böing-Messing, F., Bürkner, P., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H., Klugkist, I., Kuiper, R.M., Lee, M.D., Leenders, R., Leplaa, H.J., Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., Mulder, J., Palfi, B., Schönbrodt, F., Tendeiro, J., van den Bergh, D., van Lissa, C.J., van Ravenzwaaij, D., Vanpaemel, W., Wagenmakers, E., Williams, D.R., Zondervan-Zwijnenburg, M., Hoijtink, H. (2022). *A review of applications of the Bayes factor in psychological research*. In: «Psychological Methods» – doi:10.1037/ met0000454.

Held, L. (2019). *The assessment of intrinsic credibility and a new argument for p <0.005*. In: «Royal Society Open Science», vol. VI, n. 3, Art. Nr. 181534 – doi: 10.1098/rsos.181534.

Held, L., Matthews, R., Ott, M., Pawel, S. (2021). *Reverse-Bayes methods: A review of recent technical advances*, arXiv preprint arXiv:2102.13443.

Keyserse, C., Gazzola, V., Wagenmakers, E.J. (2020). *Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence*. In: «Nature Neuroscience», vol. XXIII, n. 7, pp. 788-799.

Kruschke, J. (2018). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Elsevier, Amsterdam, 2nd edition.

Kruschke, J.K., Liddell, T.M. (2018). *Bayesian data*

analysis for newcomers. In: «Psychonomic Bulletin & Review», vol. XXV, n. 1, pp. 155-177.

LEWIS, D. (1973). *Causation*. In: «The Journal of Philosophy», vol. LXX, n. 17, pp. 556-567.

LUCCHINI, M. (2013). *Il contributo del modello controfattuale all'irrobustimento della sociologia*. In: «Quaderni di Sociologia», vol. LXII, pp. 55-76.

LY, A., STEFAN, A., VAN DOORN, J., DABLANDER, F., VAN DEN BERGH, D., SARAFOGLOU, A., KUCHARSKY, S., DERSK, K., GRONAU, Q.F., RAJ, A., BOEHM, U., VAN KESTEREN, E.-J., HINNE, M., MATZKE, D., MARSMAN, M., WAGENMAKERS, E.J. (2020). *The Bayesian methodology of Sir Harold Jeffreys as a practical alternative to the p value hypothesis test*. In: «Computational Brain & Behavior», vol. III, n. 2, pp. 153-161.

MARTINI, A. (2006). *Metodo sperimentale, approccio controfattuale e valutazione degli effetti delle politiche pubbliche*. In: «Rassegna Italiana di Valutazione», vol. XXXIV, pp. 61-74.

MATTHEWS, R.A.J. (2001). *Methods for assessing the credibility of clinical trial outcomes*. In: «Drug Information Journal», vol. XXXV, n. 4, pp. 1469-1478.

MATTHEWS, R.A.J. (2018). *Beyond "significance": Principles and practice of the analysis of credibility*. In: «Royal Society Open Science», vol. V, n. 1, Art. Nr. 171047 – doi: 10.1098/rsos.171047.

MATTHEWS, R.A.J. (2019). *Moving towards the post p < 0.05 era via the analysis of credibility*. In: «The American Statistician», vol. LXXIII, pp. 202-212.

MATTHEWS, R.A.J. (2021). *The p-value statement, five years on*. In: «Significance», vol. XVIII, n. 2, pp. 16-19.

MORATO, V. (2019). *Controfattuali*. In: «APhEx», vol. XX, pp. 1-58.

MORGAN, S.L., WINSHIP, C. (2014). *Counterfactuals and causal inference*, Cambridge University Press, Cambridge.

PEARL, L., GLYMOUR, M., JEWELL, N.P. (2016). *Causal inference in statistics*, Wiley, New York.

POTTER, G.E. (2020). *Dismantling the fragility index: A demonstration of statistical reasoning*. In: «Statistics in Medicine», vol. XXXIX, n. 26, pp. 3720-3731.

QUATTO, P., RIPAMONTI, E., MARASINI, D. (2022). *Beyond p < 0.05: A critical review of new Bayesian proposal for assessing the p-value*. In: «Journal of Biopharmaceutical Statistics», online: 4 March 2022 - doi: 10.1080/ 10543406.2021.2009497.

ROESE, N. (2009). *The psychology of counterfactual thinking*. In: «Historical Social Research», vol. XXXIV, n. 2, pp. 16-26.

ROSENBAUM, P.R. (2010). *Design of observational studies*, Springer, Berlin/New York.

ROSENBAUM, P.R., RUBIN, D.B. (1983). *Assessing sensitivity to an unobserved covariate in an observational study with binary outcome*. In: «Journal of the Royal Statistical Society», vol. XLV, n. 2, pp. 212-218.

ROSENBAUM, P.R., RUBIN, D.B. (1983). *The central role of the propensity score in observational studies for causal effects*. In: «Biometrika», vol. LXX, n. 1, pp. 41-55.

RUBIN, D.B. (1974). *Estimating causal effects of treatments in randomized and nonrandomized studies*. In: «Journal of Educational Psychology», vol. LXVI, n. 5, pp. 688-701.

RUBIN, D.B. (2005). *Causal inference using potential outcomes: Design, modeling, decisions*. In: «Journal of the American Statistical Association», vol. C, n. 469, pp. 322-331.

SCOTT, M.L., BARTLETT, B. (2019). *Bayesian statistics in sociology: Past, present, future*. In: «Annual Review of Sociology», vol. XLV, pp. 47-68.

WAGENMAKER, E.J., VERHAGEN, J., MATZKE, D., STEINGROEVER, H., ROUDE, J.N., MOREY, R. (2017). *The need for Bayesian hypothesis testing in psychological science*. In: S.O. LILLIENFELD, I.D. WALDMAN (eds.). *Psychological science under scrutiny*, Wiley, New York, pp. 123-138.

WAGENMAKERS, E.J., MARSMAN, M., JAMIL, T., LY, A., VERHAGEN, J., LOVE, J., SELKER, R., GRONAU, Q.F., SMIRA, M., EPSKAMP, S., MATZKE, D., ROUDER, J.N., MOERY, R.D. (2018). *Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications*. In: «Psychonomic Bullettin & Review», vol. XXV, n. 1, pp. 35-57.

WALSH, M., SRINATHAN, S.K., MCAULEY, D.F., MRKOBRADA, M., LEVINE, O., RIBIC, C., MOLNAR, A.O., DATTANI, N.D., BURKE, A., GUTATT, G., THABANE, L., WALTER, S.D., POGUE, J., DEVERAUX, P.J. (2014). *The fragility of trial results is frequently fragile: A case for a fragility index*. In: «Journal of Clinical Epidemiology», vol. LXVII, n. 6, pp. 622-628.

WALTER, S.D., THABANE, L., BRIEL, M. (2020). *The fragility of trial results involves more than statistical significance alone*. In: «Journal of Clinical Epidemiology», vol. CXXIV, pp. 34-41.