# Il test severo

*Una tesi che non si lascia smentire non è nemmeno sbagliata, e una tesi che ha superato soltanto prove incapaci di smentirla non ha guadagnato niente. Dieci schede su come si misura quanto vale una conferma, dentro e fuori dalla statistica.*

CLAUDIO CAMMARANO 30 SETTEMBRE 2026

| | |
|---|---|
| **FONTI** | 10 schede, ago 2005 – set 2026 |
| **ARGOMENTI** | Epistemologia · Scienza · AI |
| **CONCETTI** | crisi della replicazione · not even wrong · preregistrazione · statistica dell'errore · test severo |
| **ORIGINE** | https://claudiocammarano.com/writings/2026-09-30-il-test-severo/ |

**TESI**

Una conferma vale quanto la probabilità che la prova aveva di smentire; la domanda, nata per giudicare gli esperimenti, si applica ai numeri del giornalismo, alle scoperte dei modelli e alle affermazioni di questo archivio.

La catena *Non è nemmeno sbagliato* si chiudeva con una domanda da rivolgere a ogni previsione sull'intelligenza artificiale: che cosa dovrebbe accadere perché risulti sbagliata. È la condizione minima, quella di Popper, e non basta, perché un'affermazione può essere smentibile in linea di principio e aver superato soltanto prove che non l'avrebbero mai smentita. Questa è la seconda catena della serie *Lo standard di prova*. La prima ha ricostruito il p-value e le scuole che se lo contendono; questa prova a portare fuori dal laboratorio lo standard che l'archivio ha adottato, il test severo, e a usarlo su affermazioni che nessun p-value accompagna.

## La prova che avrebbe potuto fallire

Nella definizione di Deborah Mayo, che l'archivio ha preso dal suo post su Peirce, un'ipotesi supera un test severo quando il dato concorda con l'ipotesi e la procedura, se l'ipotesi fosse sbagliata, lo avrebbe segnalato con probabilità molto alta.[1] Mayo lo spiega con un metal detector: la domanda utile è se quel modello, passato su un chiodo, suonerebbe. Da Peirce ricava anche tre ordini di induzione, che uso come scala per qualunque affermazione. Il primo è

l'induzione rozza, che procede per assenza di confutazione ed è in sostanza un argomento dall'ignoranza. Nel secondo la prova è genuina, e la sua forza dipende da quanto la previsione vada contro ciò che ci si aspetterebbe senza l'ipotesi. Il terzo comincia quando quella distanza si può misurare con probabilità d'errore dichiarate.

Popper aveva già l'idea, e la voce di Stephen Thornton la riporta con le sue parole: la corroborazione conta solo se è il risultato di una previsione genuinamente rischiosa.[2] La stessa voce registra l'obiezione di Lakatos su Nettuno, secondo cui nessuna osservazione isolata falsifica una teoria di alto livello, perché il fallimento si può sempre addebitare a un'ipotesi ausiliaria. La mossa di Mayo sposta la valutazione dalla teoria alla procedura. Invece di chiedere se un'osservazione abbia confutato la fisica newtoniana, chiede se la procedura usata avrebbe rilevato quell'errore specifico, e in questo modo può giudicare anche le ipotesi ausiliarie, una alla volta. Mayo stessa scrive che Popper l'idea della severità non l'ha mai incassata adeguatamente.

La tesi di questa catena è che **una conferma vale quanto la probabilità che la prova aveva di smentire**, e che la domanda, nata per giudicare gli esperimenti, si applica ai numeri del giornalismo, alle scoperte dei modelli e alle affermazioni di questo archivio.

---

**1** · Deborah G. Mayo / Error Statistics Philosophy · 11 set 2026 · **RIFERIMENTO TECNICO**

**Happy Birthday C.S. Peirce: Peircean Induction and the Error-Correcting Thesis**

Lo standard con cui l'archivio giudica le affermazioni, da giugno 2026, l'avevo preso di seconda mano: questa è la fonte, e serve a correggere l'equivoco più diffuso — un test è severo per quello che la procedura avrebbe rilevato, non per quanti dati ha raccolto.

https://errorstatistics.com/2026/09/11/happy-birthday-c-s-peirce-peircean-induction-and-the-error-correcting-thesis/

---

## Il primo ordine

L'argomento dall'ignoranza ha una versione istituzionale, e Caty Enders, sul *Guardian*, la trova nella frase con cui Mark Zuckerberg difende le piattaforme, cioè che la scienza sul video in formato breve non è conclusiva.[3] La meta-analisi più recente, su circa settanta studi, associa il formato breve a cambiamenti cognitivi più che al peggioramento della salute mentale, ed Enders dichiara per prima i limiti di quell'evidenza: letteratura scarsa, un solo studio sulla memoria, risultati che chiedono replicazione, ricerca concentrata in Cina su un prodotto che non è lo stesso. La scheda aggiunge il punto che interessa qui. Dire che nessuno ha dimostrato il danno è un'induzione del primo ordine, che dall'assenza di prova ricava un permesso, e il costo dell'attesa ricade su chi non decide mentre chi decide guadagna dal ritardo.

Luciano Floridi, Claudio Novelli e Jessica Morley aggiungono un caso che viene prima della severità stessa.[4] Una previsione che trattiene metodo, dati e incertezza non si può mettere alla prova, e la sua severità è nulla per costruzione, qualunque cifra esibisca. Il loro esempio più istruttivo è il paper di Frey e Osborne del 2013, che annunciava di voler evitare previsioni sulla legislazione e sull'accettazione pubblica della tecnologia. Sembrava prudenza, e metteva al riparo il 47 per cento, perché qualunque cosa avesse fatto poi il mercato del lavoro si poteva addebitare alle variabili escluse. Gli autori lo chiamano il predicamento di Duhem-Quine in miniatura, che è l'obiezione di Lakatos vista dal lato di chi la sfrutta, e ne ricavano una regola che uso da allora: un'affermazione formulata in una metrica è confermata o smentita soltanto

da un esito formulato nella stessa metrica.

## Quando nessuno bara

Il giardino dei sentieri che si biforcano, di Andrew Gelman ed Eric Loken, è la scheda che ha cambiato di più il mio modo di leggere i numeri del giornalismo.[5] Un articolo di dati che sceglie il periodo, il confronto e la categoria dopo aver guardato le serie compie le stesse scelte che loro descrivono, quasi sempre in buona fede, e il risultato che pubblica ha superato una prova che difficilmente avrebbe potuto fallire. Gli autori indicano come rimedio la preregistrazione dell'intero protocollo o la replica prima della pubblicazione, e ammettono con onestà che nei loro progetti applicati hanno imparato tantissimo proprio guardando i dati.

**5** · Andrew Gelman ed Eric Loken / American Scientist · 1 nov 2014 · **STRUMENTO CONCETTUALE**

**The Statistical Crisis in Science**

Il giardino dei sentieri che si biforcano: il p-value si invalida anche quando l'analisi condotta è una sola, perché conta quante ne sarebbero state possibili con dati diversi.

https://doi.org/10.1511/2014.111.460

Il sesto corollario di John Ioannidis aiuta a capire dove aspettarsi i sentieri più numerosi: più un campo è caldo, con più squadre in competizione, meno i suoi risultati sono veri.[6] Nessun campo è oggi più caldo dell'intelligenza artificiale, e il corollario suggerisce di leggere con il sospetto maggiore proprio i risultati che arrivano con più clamore, a cominciare dai confronti sui benchmark. Ioannidis si spinge oltre, e scrive che in molti campi le dimensioni d'effetto dichiarate possono essere semplicemente misure accurate del bias prevalente.

La replica è il modo più diretto di rendere severa una prova, perché la ripete senza il giardino alle spalle, e l'Open Science Collaboration mostra che cosa ne resta: il 36 per cento dei risultati significativi, con effetti dimezzati.[7] Le cautele degli autori valgono anche fuori dalla psicologia. Una replica riuscita misura l'affidabilità di un risultato e lascia aperta la correttezza della sua spiegazione, e un singolo studio, scrivono, non risolve quasi mai una questione in un senso o nell'altro.

## Il terzo ordine

Il terzo ordine si riconosce da un dettaglio: chi fa l'affermazione dichiara la probabilità d'errore della procedura. Il watermark proposto nel 2023 da John Kirchenbauer e dai colleghi dell'Università del Maryland ne è un esempio pulito.[8] Il modello, prima di generare ogni token, favorisce leggermente una lista di token scelta con un hash del token precedente, e il segnale si rileva con un semplice z-test: alla soglia z = 4 i falsi positivi sono tre su centomila, e bastano sedici token. È una prova di cui si conosce la severità prima di usarla. I rilevatori di testo artificiale oggi in commercio non possono contare su un watermark, perché nessun modello diffuso lo implementa, e restano scatole nere statistiche che non dichiarano nulla del genere; i loro verdetti andrebbero pesati di conseguenza, cioè come induzioni di secondo ordine presentate come se fossero di terzo.

All'estremo opposto sta la dimostrazione formale. Quando Anthropic ha annunciato che una versione di Claude aveva migliorato il limite inferiore della frazione di zeri della funzione zeta che soddisfano l'ipotesi di Riemann, dal 41,6 al 67,2 per cento, il risultato è arrivato validato da

quattro matematici, due interni e due esterni, e formalizzato in Lean, un sistema che certifica meccanicamente ogni passaggio logico.[9] Un verificatore formale avrebbe rifiutato la dimostrazione se contenesse un errore, e questo porta la severità vicino al massimo possibile, con un limite che va detto: Lean certifica che la dimostrazione prova l'enunciato formalizzato, e resta ai matematici giudicare se quell'enunciato è quello che intendevano. La catena *Possiamo fidarci di una macchina?* sosteneva che la fiducia nei modelli può poggiare solo su una verifica esterna dei risultati; la severità dà a quella verifica una misura.

---

**9** · Anthropic · 10 ago 2026 · **DATO CHE CORREGGE**

**Learning more about Claude's mathematical capabilities**

Non è un benchmark costruito per far bella figura: è un risultato matematico originale su un problema aperto dal 1859, prodotto come sottoprodotto non intenzionale di un tentativo fallito. Il salto dal 41,6% al 67,2% conta meno del contesto in cui è avvenuto — il modello era scettico sulla propria capacità di contribuire, aveva bisogno di essere incoraggiato, e ha poi formalizzato il risultato in Lean in modo autonomo. La stessa sessione ha prodotto, in nota a piè di pagina, una confutazione della congettura Jacobiana.

https://www.anthropic.com/research/riemann-zeta

---

**Controtesi 10 · Jan-Willem Romeijn / Stanford Encyclopedia of Philosophy**

La severità ha un avversario di principio, e la voce di Jan-Willem Romeijn lo nomina.[10] Per un bayesiano l'evidenza portata da un dato dipende soltanto dal dato osservato, secondo il principio di verosimiglianza, e la statistica classica lo viola: nella sua impostazione, scrive Romeijn, l'impatto dei dati osservati può cambiare a seconda della probabilità di campioni diversi da quello osservato. La severità fa esattamente questo, perché giudica una conferma in base a ciò che la procedura avrebbe fatto in casi che non si sono verificati. Due ricercatori con gli stessi dati, uno che aveva deciso in anticipo quando fermarsi e uno che ha continuato a raccogliere finché il risultato è arrivato, ricevono dalla severità giudizi diversi, e per un bayesiano far dipendere l'evidenza dalle intenzioni di chi la raccoglie è un paradosso.

Prendo l'obiezione sul serio, e questa catena non la risolve. Osservo soltanto che fuori dal laboratorio la dipendenza dalle intenzioni è proprio ciò che serve. Il giornalista che ha scelto il confronto dopo aver visto i dati e quello che lo aveva dichiarato prima mi mettono davanti la stessa tabella, e da lettore voglio sapere quale dei due mi sta parlando. Il bayesiano mi chiederebbe una probabilità a priori che nessuno dei due mi dà, mentre la domanda di Mayo, cioè che cosa avrebbero trovato se la loro tesi fosse sbagliata, posso farla a entrambi. Per l'archivio la regola diventa doppia: a ogni affermazione che entra chiedo se sia smentibile, come voleva Pauli, e poi quante possibilità avesse di non superare la prova che cita. Resta da capire chi decida quale prova basti, ed è l'oggetto della terza catena della serie.

## Nella catena

1. Deborah G. Mayo / Error Statistics Philosophy, *Happy Birthday C.S. Peirce: Peircean Induction and the Error-Correcting Thesis*, 11 set 2026. Riferimento tecnico.
   https://errorstatistics.com/2026/09/11/happy-birthday-c-s-peirce-peircean-induction-and-the-error-correcting-thesis/
   Scheda: https://claudiocammarano.com/curated/2026-09-11-mayo-peirce-tesi-autocorrettiva-errorstatistics/
2. Stephen Thornton / Stanford Encyclopedia of Philosophy, *Karl Popper*, 31 lug 2026. Riferimento tecnico.

https://plato.stanford.edu/entries/popper/
Scheda: https://claudiocammarano.com/curated/2026-07-31-thornton-karl-popper-sep/

3. Caty Enders / The Guardian, *Zuckerberg says the science isn't settled. But the harms of short-form video on the brain are starting to show*, 18 set 2026. Dato che corregge.
https://www.theguardian.com/commentisfree/2026/sep/14/zuckerberg-short-form-video-cognitive-harm
Scheda: https://claudiocammarano.com/curated/2026-09-18-enders-short-form-video-cognizione-guardian/

4. Luciano Floridi, Claudio Novelli e Jessica Morley / SSRN, *Not Even Wrong 1: AI and the Labour Market, From Frey–Osborne to ChatGPT, 2012–2026*, 10 giu 2026. Strumento concettuale.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6830981
Scheda: https://claudiocammarano.com/curated/2026-06-10-floridi-not-even-wrong-1-lavoro-ssrn/

5. Andrew Gelman ed Eric Loken / American Scientist, *The Statistical Crisis in Science*, 1 nov 2014. Strumento concettuale.
https://doi.org/10.1511/2014.111.460
Scheda: https://claudiocammarano.com/curated/2014-11-01-gelman-loken-giardino-sentieri-biforcano-americanscientist/

6. John P. A. Ioannidis / PLoS Medicine, *Why Most Published Research Findings Are False*, 30 ago 2005. Strumento concettuale.
https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124
Scheda: https://claudiocammarano.com/curated/2005-08-30-ioannidis-most-published-findings-false-plosmed/

7. Open Science Collaboration / Science, *Estimating the reproducibility of psychological science*, 28 ago 2015. Dato che corregge.
https://doi.org/10.1126/science.aac4716
Scheda: https://claudiocammarano.com/curated/2015-08-28-open-science-collaboration-riproducibilita-science/

8. John Kirchenbauer et al. / arXiv, *A Watermark for Large Language Models*, 25 gen 2023. Riferimento tecnico.
https://arxiv.org/abs/2301.10226
Scheda: https://claudiocammarano.com/curated/2023-01-25-kirchenbauer-watermark-llm-arxiv/

9. Anthropic, *Learning more about Claude's mathematical capabilities*, 10 ago 2026. Dato che corregge.
https://www.anthropic.com/research/riemann-zeta
Scheda: https://claudiocammarano.com/curated/2026-08-10-anthropic-riemann-zeta-claude/

10. Jan-Willem Romeijn / Stanford Encyclopedia of Philosophy, *Philosophy of Statistics*, 1 ott 2025. Riferimento tecnico.
https://plato.stanford.edu/entries/statistics/
Scheda: https://claudiocammarano.com/curated/2025-10-01-romeijn-filosofia-della-statistica-sep/

**Intervento AI: WR** (Scrittura). Notazione descritta in https://claudiocammarano.com/notazione/.

© Claudio Cammarano. Licenza CC BY-NC-ND 4.0: si può condividere citando l'autore e la fonte, non a fini commerciali e senza modifiche.

Versione originale e aggiornata: https://claudiocammarano.com/writings/2026-09-30-il-test-severo/. Le parti interattive della catena sono disponibili solo online.

Questa catena fa parte della serie *Lo standard di prova*: https://claudiocammarano.com/serie/lo-standard-di-prova/.

Scrivo anche *The Abstract*, in inglese, una volta al mese, sull'economia dei contenuti: https://claudiocammarano.com/newsletter/. C'è anche il feed: https://claudiocammarano.com/feed.xml.