# Advancing statistics reform: Ways to improve thinking and practice in the face of resistance,

evolved from:

**'Breaking the statistical tyranny over rational cognition'**

**'Cognition and causation before probability and inference'**

**'How to not lie with statistics: use descriptive imagery, not inference'**

**'The need for cognitive science and causality in statistics teaching and practice'**

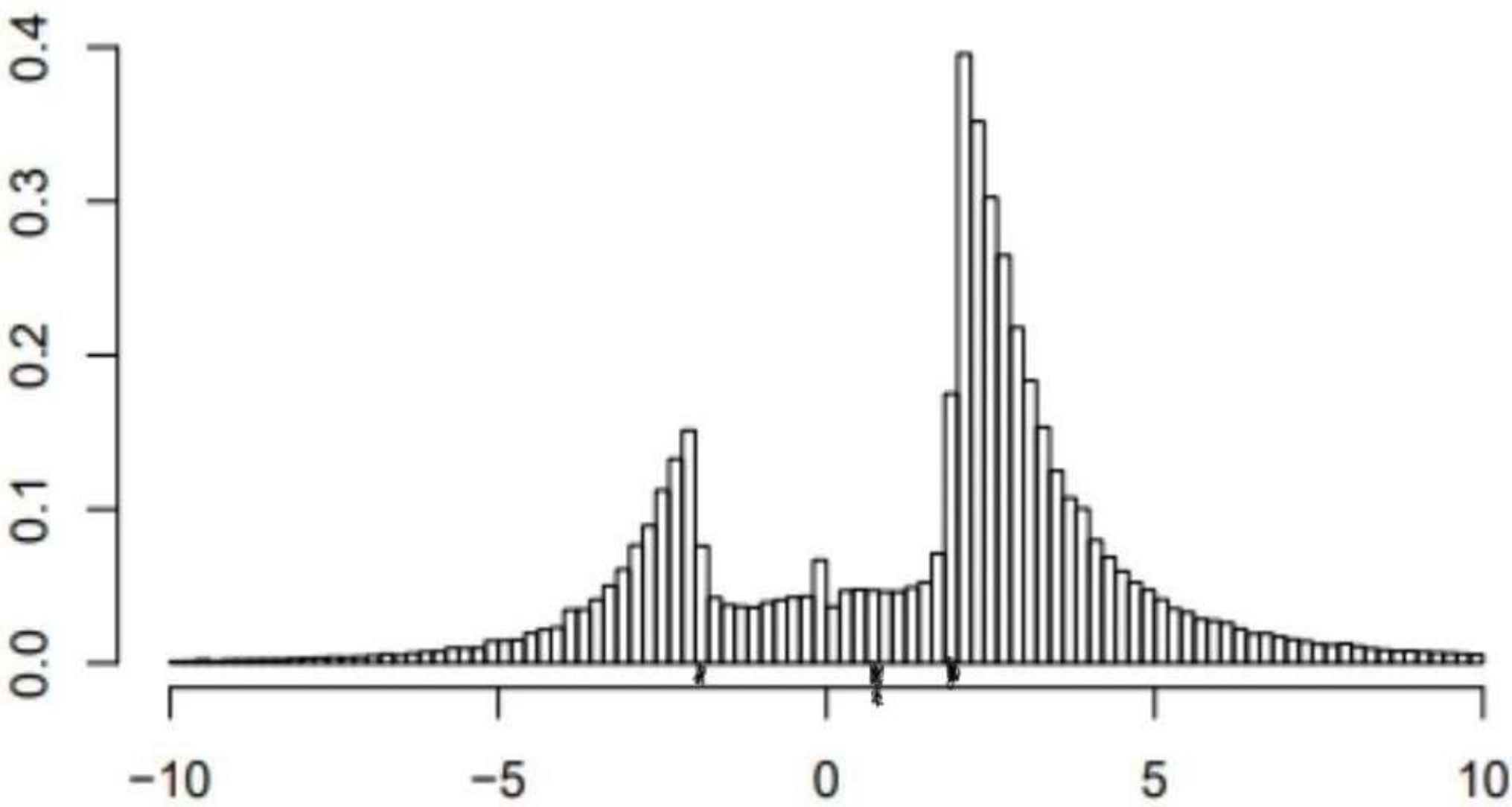
**'Statistics as a condemned building: plans for demolition and reconstruction'**

Sander Greenland, Dept of Epidemiology and Dept of Statistics, UCLA
Please report errors and send comments to Sander Greenland at lesdomes@ucla.edu

# Science progresses funeral by funeral, but in statistics authority is immortal

- **Heroic narrative**: Science progresses by each generation challenging the ideas of its predecessors, discarding those that fail stringent **empirical** tests.
- In contrast, **statistics has decayed by enshrining traditional methodologies**, and then defending them by academic mathematical and philosophical appeals, along with underplaying harms to public information.

Consequence: Fig. 1 van Zwet & Cator 2021
Over a million z-values from Medline 1976-2019.
Imputed curve right-skewed with >75% above 0:



- **Sanktification of cognitive biases (like nullism and dichotomania) as "scientific principles", treatment of mathematical frameworks as if physical realities (reification), and neglect of human biases (such as craving certainty and finality) have rotted the core of statistics.**
- **A solution: Reconstruct statistics as an information science, not as a branch of probability theory, with cognitive science and causality theory as core components.**

In the radical Bayesianism of DeFinetti, all probability is "subjective" – describing only properties of observer's minds. In that view

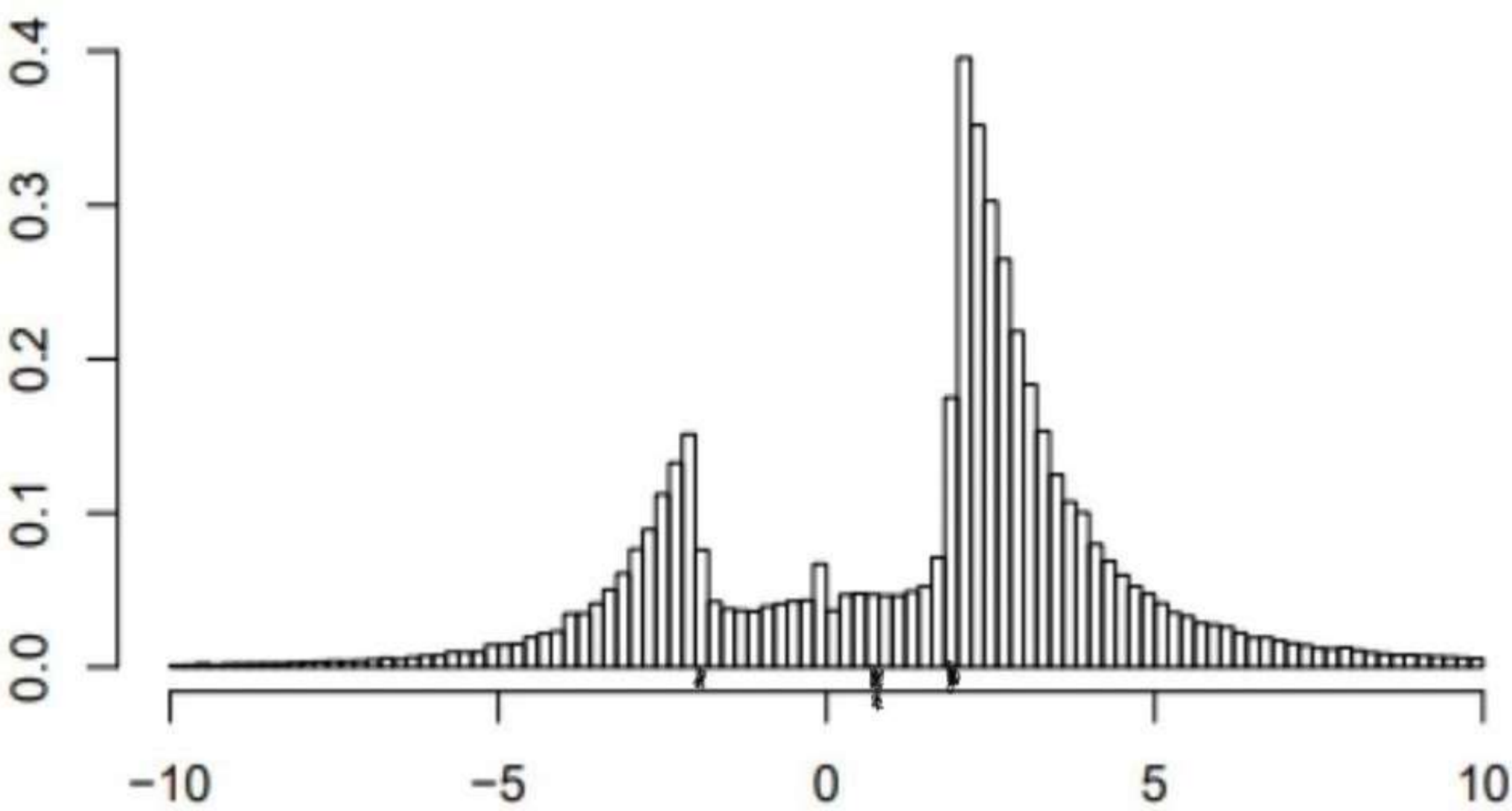
- The idea that patterns are "caused by chance" is absurd as a causal statement about the world;
- Rather, we seek **causal explanations** for a recognized pattern by considering a highly nonrandom (biased) selection of the few causal possibilities that are put forth as plausible;
- We then reify the residual infinitude of unconsidered causal explanations as forming a metaphysical cause called "chance".

- **All analyses should be viewed as part of a vastly incomplete sensitivity analysis.**
- **The frequentist vs Bayes controversy is a religious dispute that disappears under detailed logical analysis**.
- Boxian view: Bayesian tools are for method and model building, frequentist tools are for their evaluation (many other useful combinations).
- Until recently, both "schools" failed to cover the essential causal/contextual dimension from which their methods should be derived.

# What unifies these inference concepts? - Not probability, but causation:

- Past causes: **What caused** ("explains") **our observations?** which is asking about **physical mechanisms**, *not* abstractions of their behavior such as probabilities.
- Future effects: **How will actions affect the future?** which is asking how to change the behavior of mechanisms, such as actual event frequencies, **not** probability distributions.
- Example: What will be the effect of reforms?...

Answer: **Any** reform that still leads to selective reporting based on study outcomes will distort the distribution of available outcomes relative to the total



# What is INFERENCE?

- Dictionary example: "**A conclusion reached on the basis of evidence and reasoning.**"
- **Scientific inference is a complex but narrowly moderated judgement about reality,** based on this assumption:

  There is a logically coherent "objective" (observer-external) reality that **causes** our perceptions according to discoverable laws:

**My perception ← Reality → Your perception**

- **This makes inference part of cognitive science**

# Contrast scientific inference to

- "**Statistical inference**," which in all formalisms, "schools" or toolkits, **has become taking output from a data-processing program** (learning algorithm) **and generating "inferences" via decontextualized rules**.
- It converts oversimplified models of the mechanisms generating the data – the **causes** of the data – into abstract probability distributions.
- **The semantic void it leaves causes inferential errors and facilitates deception (**self or other)

- **Statistics ignored or denigrated semantics and ordinary language, favoring instead deceptive jargon promising "significance" and "confidence" even when studies provide nothing close without huge leaps of faith.**
- This was done to sell technical products and services based on dense formalisms, notation, and **artificial precision** whose assumptions and dangers are poorly understood by most users and consumers in "soft sciences"

**- note the parallel with medical-product sales!**

# The scientific community eagerly contributed to the degeneration of statistical science

Rules that were apparently successful for narrow automated environments induced destructive feedback loops in teaching and research:

- Students want explicit practice rules for memorization to ensure correct answers.
- Instructors want ease of grading.
- Researchers want rules for submitting acceptable reports.
- Reviewers and editors want to ease reviewing and publication decisions.

# The prevailing rules became especially popular and destructive via enforced dichotomies

- **Dichotomies satisfy human drives for definitive conclusions**, since they apply even when the study (the real physical data generator) is incapable of forcing such conclusions if critically scrutinized.*

*apart from "more research is needed", although often even that isn't justified in light of cost/benefit considerations and other studies.

The degeneration of statistical science into a collection of mathematical skeletons left behind explication of and training in essential components of **scientific inference**:

- How **causal networks (not probabilities) produce data, inferences, and decisions**.
- How **cognitive biases as well as procedural problems enter those causal networks**.
- How **valuations** (motivations, goals, real costs and benefits) **affect cognition and are implicit in all methodologies**.

- **Ugly fact: The main problems of P-values will extend to any statistic, because they stem from truth-subverting (perverse) incentives and cognitive biases, not P-values**
- **Perverse incentives create cognitive biases (wishful thinking, mind projection) to see what the incentives dictate. These biases pervade reports in fields like medicine.**
- **Perceptions are currently manipulated to see incentives for positive reporting while ignoring incentives for negative reporting…**

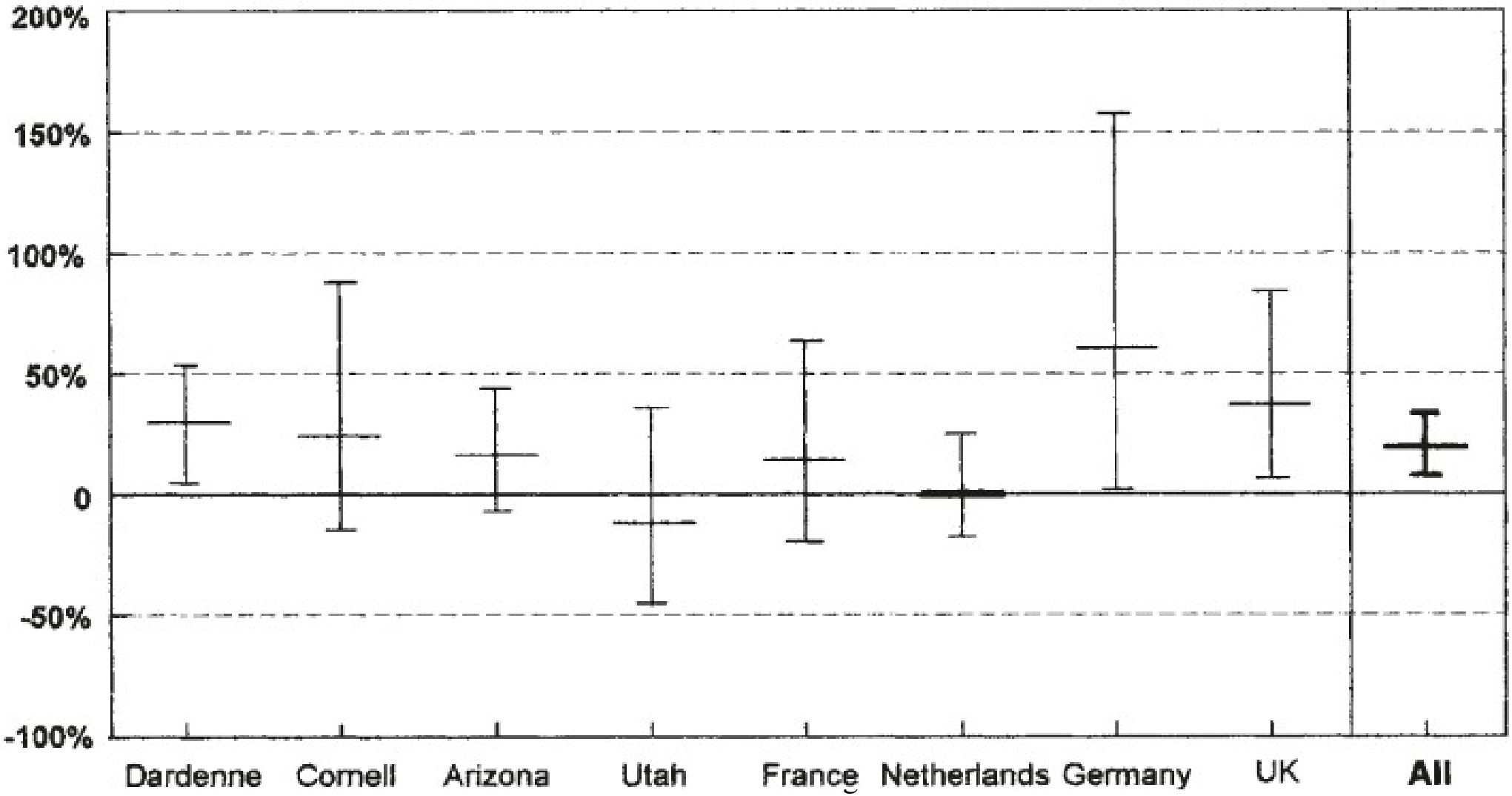
- **Reasoning motivated by commitment to past teaching, past practice, and financial stakes drives resistance to serious reform**

Example – a common label on dairy products:

"*MILK from cows not treated with rBST.
***No significant difference** has been shown between milk derived from cows treated with rBST and those not treated with rBST"

**- Here, a special-interest group forced a statement of fact to be accompanied by a misleading technical claim to benefit sales.**

Millstone et al. *Nature* 1994: 8 trials, 19% average increase in somatic cell count (pus) in milk from cows treated with rBST (meta p=0.004):



- The “replication crisis” is constantly portrayed as one of perverse incentives to make discoveries by searching out “statistical significance”, producing publication bias.
- **Lowering significance thresholds only increases the bias.**
- **Any** selective publication based on outcomes damages goals of building complete, unbiased public data repositories.
- **Yet defense and promotion of significance selection continues unabated…**

More subtly, the standard "replication crisis" story ignores instances of perverse incentives to find and report **negative** results (e.g., by **upward P-hacking or by misreporting ambiguous results as negative**), for example

- When researchers, sponsors, and editors want to dismiss undesirable associations; or
- When "replication failures" or other challenges to an association are more publishable than mere replication.
- Or both…

**A typical example** (Brown et al., "Association between serotonergic antidepressant use during pregnancy and autism spectrum disorder in children", JAMA 2017;317:1544-52), abstract:

- "[Cox-model] adjusted HR, **1.59** [95% CI, **1.17, 2.17**]). After IPTW HDPS, the association was not significant (HR, **1.61** [95% CI: **0.997, 2.59**])." (**p = 0.0505**)
- Their conclusion: "**in utero exposure was not associated with autism spectrum disorder**"
- Their earlier meta-analysis got **HR 1.7 [1.1,2.6]**

Articles decrying this sort of misreporting date at least back to Karl Pearson **1906**:

- "The absence of significance relative to the size of the samples is too often interpreted by the casual reader as a denial of all differentiation, and this may be disastrous."

Innumerable others have repeated this caution for over a century since…

Why does it continue in such naked forms?

Is it mere ignorance? No, I posit **it's forced on authors to protect industry against litigation**.

"…**the distinction between statistical significance and social importance should be apparent to all research workers**…upon us is placed the responsibility of determining whether real differences exist and then of indicating their social importance **and their cost**. **When we fail to find any statistically significant differences, we are not justified in concluding at once that no real differences exist.**" – P. 118 of JW Tyler, Educational Research Bulletin, Mar. 4, **1931**

"One of the most pernicious abuses of automated decision making occurs when clinical treatments are asserted to be equivalent, based on a nonsignificant P-value for the observed difference…we should continue to resist any attempts to automate our decisions, as in formal hypothesis testing."

- Claire Weinberg, "**It's Time to Rehabilitate the P-value**", *Epidemiology* 2001; 12: 288-290.

Brown et al. cited their own report of the same increased risk in their own meta-analysis of **4** earlier cohorts with **HR 1.7 [1.1, 2.6] but…**

- **They did not attempt to combine their new study with those studies and**
- **They did not cite a 2016 meta-analysis by Healy et al. of 16 cohort studies with HR 1.74 [1.19, 2.54] and 5 case-control studies with HR 1.95 [1.63, 2.34]**

**Why no discussion of the consistent association of 60-70% higher risk among the exposed?**

That's because most were certain this highly replicated association was pure confounding:

- *Medscape* 2017: "**Use of antidepressants before and during pregnancy does not cause autism or ADHD new research shows. Three studies demonstrate that antidepressant use in pregnant women is likely not responsible for autistic spectrum disorders (ASDs) in children and that the association found in previous studies was likely due to confounding factors.**"

**The dominant social bias talks as if all incentives are to "discover" rather than to refute effects. This meta-bias is rampant in the "replication crisis" literature, which uncritically ignores differences in incentives across topics and authors.**

- The Brown et al. example has the appearance of **CI-hacking to increase width** by adjusting until the CI finally includes 1 (even though adjustments beyond the initial Cox model have the appearance of overadjustments, inflating variance without removing bias).

The point is **not** to argue that prenatal SSRIs cause ASD (massive topic!), but rather that

- **"Spin" is the driver through The Garden of Forking Paths: "objective" statistics are perceived, selected, and reported based on preferred causal stories and, in high-stakes settings, political and litigation concerns.**
- **Examples abound throughout health and medical sciences** – which should scare you!
- Statistical training that pretends otherwise obscures and fosters this manipulation.

- **The causal stories that "we" (researchers, reviewers, and editors) want believed causally affects analysis choices and output interpretation. The result is that reports often function as lawyering for those stories.**
- **A major source of blindness to the problem is pundits in statistics and "meta-research" neglecting their own cognitive and political biases and training deficiencies, as well as the deficiencies of developers, instructors, users, and consumers of statistics.**

- **Romantic heroic-fantasy science: Committed to fact-finding and dissemination of valid facts regardless of the social consequences…**
- **but almost no one would disseminate all valid facts regardless of the consequences.**
- **Harsh reality: Much of statistics serves commitments of major social networks to warp portrayal of facts into propaganda to direct society according the network's valuations and special interests.**

Example: The endless expert "EBM" promotion of randomized trials as "gold standards" when they are no such thing due to

- **Huge generalization bias** due to exclusion of high-risk patients on ethics and liability grounds, and placebo formulas that have real side effects
- **Numbers too small and follow-up too short to discern adverse effects, resulting non-significance reported as "replication failure"**
- **Hidden protocol violations plus selective publication, reporting, and discussion …**

**A typical example**: RCT by Vallejos et al. 'Ivermectin to prevent hospitalizations in patients with COVID-19' BMC ID 2 July 2021…

- Abstract: OR = **0.65**; 95% CI **0.32**, **1.31**; p = .23 reported as "Ivermectin had no significant effect on preventing hospitalization".
- Gideon M-K "Health Nerd" (Medium 16 July 2021) wrote that the trial "found no benefit for ivermectin on death" – **BUT** p. 5 of Vallejos et al.: OR = **1.34**, 95% CI **0.30**, **6.07** **from 4 ivermectin + 3 placebo deaths.**
- **The trial was too small to show anything!**

Survey from *MedPage Today* May 21, 2021:

"Bait and Switch in IBD Trials? Primary outcomes often go unreported or changed midstream"

- "Analysis of 57 phase III trials with published results indicated that **half [~50%] either never reported at least one of the prespecified primary outcomes (17.5%) or at least one was changed without explanation (33.3%).**"

**Other studies found many *registered* trials are never published despite stated intent to do so.**

Empirical fact: **We are all stupid** (if not corrupt)

Amos Tversky: "**My colleagues they study artificial intelligence; me, I study natural stupidity**."

"**Whenever there is a simple error that most laymen fall for, there is always a slightly more sophisticated version of the same problem that experts fall for**."

Example: P-value = "probability of the null" vs. P-value = "probability chance alone produced the association" – but "chance alone" **is** the null!

# Empirical fact:

## **Incompetence among the exalted is the norm**

Tversky: "It's frightening to think that you might not know something, but more frightening to think that, by and large, the world is run by **people who have faith that they know exactly what is going on**."

– Equally true in research **and** methodology!

- **The Covid-19 pandemic has supplied us with vivid real-world examples – but no agreement about who those examples are.**

- Kahneman: "**People assign much higher probability to the truth of their opinions than is warranted.**"
- **By sanctifying pure opinion, Bayesian methods open statistics to even more abuse via prior spikes and "elicited priors" (summary expressions of biases, literature misreadings, and personal prejudices).**
- - Example: **Claiming Pr(null)=0.5 is "indifference" is massive null bias, <u>not</u> indifference.**

Yet more Kahnemann:

- "We can be blind to the obvious, and we are also blind to our blindness."

And most relevant to statistics in soft sciences:

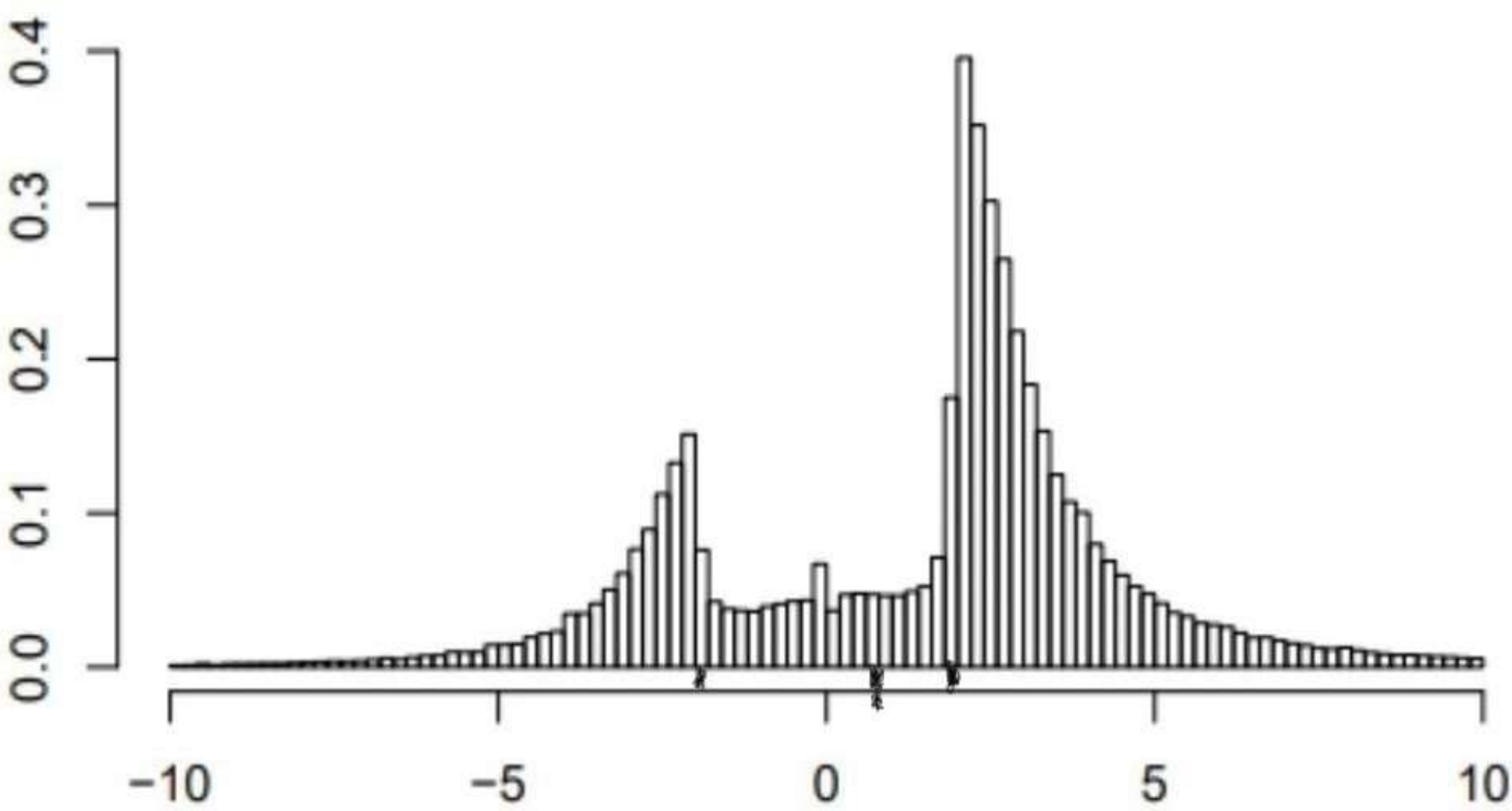
- "…**illusions of validity and skill are supported by a powerful professional culture. We know that people can maintain an unshakeable faith in any proposition, however absurd, when they are sustained by a community of like-minded believers.**"

**- See: Any defense of significance testing…**

**Example**: "**if the p-value for the effect is greater than the journal's threshold p-value, then the editor can immediately reject the paper**, which saves the journal from spending any more time on the (unconvincing) paper…if a result is statistically significant, this means no more than that there is enough weight of evidence for the studied effect to make the paper reporting the effect **worth considering for publication**."

- Fisher 1920s? No, Statistics 2021: Mcnaughton, *The War on Statistical Significance*.

Ignores a fact noted by the 1950s: **Any** selective reporting based on study outcomes will distort the distribution of available outcomes relative to the total



**Any instruction purporting to cover the basics of inference needs to include cognitive science to deal with social delusions and biases such as**

- **Nullism**: Confusion of our need for parsimony (or shrinkage to zero) with reality.
- **Dichotomania**: Confusion of our need for summarization (simplification) and decision with our preference for black-or-white thinking.
- **Reification**: Faith that **formal methods** for reasoning, inference, and decision suffice for **real-world** reasoning, inference, and decision.

# Nullism has a long and glorious history among physics idolaters as pseudo-skepticism

(empirically indefensible certainty about nulls):

- "**Heavier than air flying machines are impossible**" – Lord Kelvin 1895, repeated 1902
- "**Continental drift is out of the question**" because no mechanism is strong enough – **Sir Harold Jeffreys**, geophysicist originator of **spiked priors = formalized overconfidence**.
- See also Fisher arguing against cigarettes causing lung cancer, despite extensive evidence.

- **Against Nullism**: Reality is under no obligation to be simple or decisive.
- **Against Dichotomania**: Many if not most important decisions are not or **should not be** binary: Where do you set your oven? Your thermostat? **Your medication level**?
- **Hidden Reification**: Researchers routinely publish "inferences" that ignore vast model uncertainties – they don't know a rationale for neglecting all the simplifications in their models, and they just don't think about them.

# Many other cognitive biases contribute to design, analysis, reporting, publication biases

https://en.wikipedia.org/wiki/List_of_cognitive_biases

**All of the following and more should form part of basic training for moderating inferences**:

- **Anchoring** to perceived consensus and desired yet erroneous belief **even after correction**.
- **Confirmation bias: ** selective focus on desirable evidence and neglect of undesirable evidence.
- **Courtesy bias: ** Tendency to be obscure about criticisms that will cause offense.

- **Failure to test alternatives** ("congruence bias")
- **Selective criticism** of undesirable evidence.
- **Selective reasoning** to desired conclusions via selective assumptions, explanations, and data.
- **Dunning–Kruger effects:** The less expertise, the more the overestimation of one's competence (as in researchers' overestimation of their statistical expertise, e.g., statistical editors of med journals).
- **Overconfidence, validity illusions**: The tendency to think methods or judgments are as accurate about the world as they are in the math (thought experiments) used to derive them.

- **Familiarity bias** – over-reliance on familiar methods, ignoring alternative approaches ("gets me grants and papers, so no need to change").
- **Territorial (exclusionary) bias** – promoting familiar methods as exclusively correct approaches, thus protecting self-authority and preventing competition from gaining ground ("Strictly Ballroom" effect: You can't be an authority about what you haven't studied and used extensively).
- **Groupthink and herd-behavior biases** such as **repetition bias** (echo-chamber effect, group reinforcement causing overcount of evidence).

- **Mind-projection fallacies**: Imbuing inert quantities with attitudes, opinions, values, inferences, judgments, and decisions.

- Rampant in statistical discussions, thanks to using **value descriptors** like "significance", "confidence" and "severity" for narrow math concepts that cannot capture the word meanings.

- **Top example of nonsense: "P-values overstate evidence." P-values only provide the position of a statistic in a reference distribution (e.g., chi-squared) derived from a model. Any evidence overstatement is by the viewer.**

These are not absolute or sharp categories, but rather are heuristic triggers to avoid getting lulled or suckered by colleagues (however well-meaning), "experts," and most of all **ourselves**. Example:

- A Dunning-Kruger form of **overconfidence bias that is rampant among medical pundits** (and not only when they comment on statistical methods): We may know our specialty superbly, but not realize how that expertise doesn't instantly generalize to other topics. **True even for topics we think are close to our specialty, but actually have a lot more literature than we are aware of.**

# Systemic problems are major reasons why 'most published research findings are false':

- Like everyone, **stat instructors**, users, and consumers suffer from **dichotomania**, **nullism**, and **reification**: **They crave true-or-false conclusions for null hypotheses and so will accept them from oversimplified models**.
- But in "soft-science" applications, **observations (even from RCTs) can never provide such absolute certainties**, **and can't even provide accurate assessments of uncertainties**.

- Statistics caters to our cravings by providing sophisticated **decision theories** which make it **appear** to users that observations can provide definitive risk and uncertainty assessments.
- **"Confidence intervals" perpetuate these illusions by deceptively appearing to capture all the uncertainty sources in applications, when the only uncertainty they capture is that given the model used to compute them.**
- **Worse, standard presentations rarely mention their neglect of model uncertainty!**

- Statistics also freely indulges in the **ludic fallacy** of treating all uncertainty as if from games of chance (random draws from a distribution of **known** form) instead of addressing our deep uncertainties about **the form and causes of variation and bias**.
- **These problems underscore the need for coverage of causal reasoning errors and cognitive biases as an essential component of any specialty claiming to promote sound scientific inference from data.**

- **Mathematizations amplify overconfidence in these fallacies, making statistical theory a fountain of real-world misinformation**:
- **Math derivations only warrant certainty in conclusions (such as "optimality") given their assumptions (e.g., that a small class of model candidates can approximate reality well)**.
- Yet the conclusions are then treated as self-evident truths, a feeling reinforced by commitment to previous training, teaching, and practice. **The resulting shared cognitive biases are then reinforced by social feedback loops**.

- **The result is groupthink, hidden bias and circularity in arguments given by the most technically skilled proponents!**
- Typified by common arguments for Neymanian and Bayesian primacy (worse than that seen in writings by Neyman or Bayes).
- Examples: Demanding calibration or Bayesian coherency as prime directives for real-world conclusions and decisions. Those are only directives **within their representations**, and can lead to disaster from **uncaptured context**.

# STOP treating mathematical justifications as if they are sufficient practical justifications

- No matter how complex they look, math results are only thought experiments to test methods in idealized settings far simpler than real practice.
- Performance in these simple cases can provide guidance for practice, with warnings about problems and suggesting improvements for methods. **But**,
- Problems seen simple settings can get worse in complex settings, and
- Neither math "optimality" results nor failure to find problems in math settings guarantee good performance in real applications.

**Value bias** pervades statistical methodology, most often in the form of **nullism** (values biased toward "accepting" the null)

- Call a methodology **value-biased** when it incorporates cost/benefit assumptions that are not shared by all stakeholders.
- **These biases are usually obscured from public recognition by adherence to statistical traditions and mathematics that hide the values in the obscure cost/benefit assumptions of NHST and its Bayesian analogs.**

- Example: **Consistent use of the null as the test hypothesis, to the point of failing to distinguish between null and test hypotheses** (a mistake traceable to Fisher).
- This is an example of **nullism**, value bias toward the null favoring those with stakes on the null (as found in product surveillance) and those with metaphysical beliefs in nulls (pseudo-skeptics who confuse parsimony heuristics with natural laws).
- Many researchers do not realize that **any effect size can be given a P-value** ("tested").

- Via NHST, nullism has been taught as an integral part of Neyman-Pearson testing – even though it is not! From Neyman, *Synthese* 1977 p. 104, 106 (emphases added):
- "**According to circumstances and according to the subjective attitudes of the research worker, one error may appear more important to avoid than the other; the error which is the more important to avoid will be called 'error of the first kind'**" ["Type-I" error, alpha error, **incorrect rejection of the test hypothesis H**]

- "the [**hypothesis**] the unjust rejection of which constitutes the error of the first kind, **will be called 'the hypothesis tested**'."
- **Note how this description allows that the test hypothesis H may be non-null.**
- "From the point of view of the manufacturer [of a chemical A] the error in asserting the carcinogenicity of A is (or may be) more important to avoid than the error in asserting that A is harmless. Thus, **for the manufacturer, the 'hypothesis tested' may well be: 'A is *not* carcinogenic'.**"

- "**On the other hand, for the prospective user of chemical A the hypothesis tested will be unambiguously: 'A *is* carcinogenic'. In fact, this user is likely to hope that the probability of error in rejecting this hypothesis be reduced to a very small value!**"
  - **This means anyone teaching, promoting, and using statistical tests must justify their choice of test hypothesis H as well as the cutoff used (whether that cutoff is for a P-value or a Bayes factor or a likelihood ratio).**

- Neyman thus provided a clear description of the role of values in choosing test hypotheses and how those can (and often will) vary within a topic across stakeholders.
- **Yet many "opinion leaders" maintain rigid practices of testing only the null,** based on faith in grossly oversimplified biological models, generalizations from selective observation, treating simplicity or parsimony heuristics as if they were metaphysical principles, and of course **hidden valuations including service to sponsors or ideologies**.

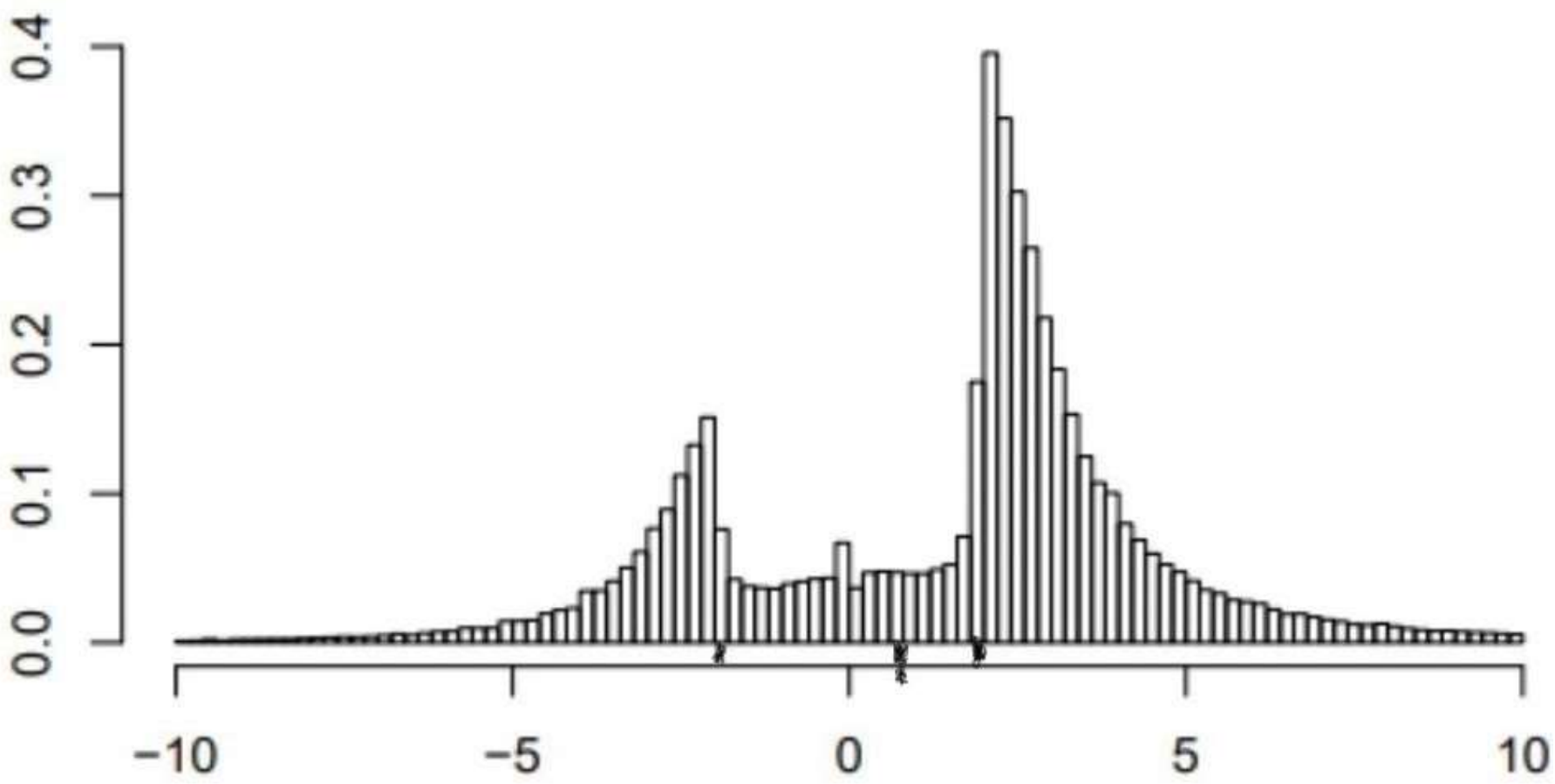
- We might be confident that any effect present is "small enough" so that the cost of ignoring it is acceptable - **but that's a value judgment!**
- Statistical tests can be constructed for superiority, inferiority, non-superiority, non-inferiority, and even equivalence of treatments – but require artificially precise specifications of the effect sizes defining such declarations.
- P-values for such "boundary" effect sizes should be presented – and **a P-value graph can show the P-values for all such choices**.

# Multiplicity adjustments worsen value bias

- They traditionally take the **joint ensemble null** as the hypothesis most important to not reject incorrectly, and apply the Type-I error rate of 0.05 to **the entire ensemble of nulls**.
- **They thus assume false-positive costs are always more than false-negative costs, and that these cost ratios always increase with the number of hypotheses - This valuation applies to drug companies monitoring adverse effects but not to patients!**

- **Null bias also afflicts a large portion of the Bayesian literature**, where **null spikes** are used to misrepresent a belief that a parameter "differs negligibly" from the null.
- In most medical-research settings, concentration of prior probability around the null has no basis in real data. In fact **prior spikes usually contradict genuine prior information. For example, potential medications are studied *precisely because* they are known to affect the targeted physiological system**.

Example: Again, fig. 1 of van Zwet & Cator 2021. Some Bayesians would shrink estimates toward 0 despite an imputed curve right-skewed with >75% above 0. Empirical Bayesians would instead use shrinkage toward estimated topic-specific means.



# Part I summary:

- **Blind acceptance of mathematical frameworks, deification of "great men" and their conceptual errors, and neglect of cognitive problems have rotted the core of statistical training, practice, and the resulting store of public information.**
- **The "replication crisis" hysteria continues the problem via its nullism (neglect of testing alternatives), pointless dichotomania, and pernicious model reification, all of which are enshrined in NHST and null-spiked priors.**

- The persistent practical mistake promoted by statistical methodology: **That we should want to construct real-world inferences using deductions from only one study, one set of background assumptions, one formal reasoning system, or one interpretation**.
- Most writers accept the need for varied designs (not just "replications") and varied assumptions (sensitivity analysis), **yet seem unaware of (some even fight) the need for varying methodology and interpretations.**

- **Statistical rules can worsen bad practices** because their theories assume we will use only perfect interpretations of carefully controlled experiments, with a clear view of error costs.
- But **most "data analysis" in soft-science research has been about applying decision rules to statistical outputs, based on defaults whose value-laden nature is not seen by most users and readers**, e.g., requiring $P<0.005$ to report "association", or misinterpreting $P>0.05$ as "no association."

The bare, psychosocial facts:

- **Most "objective" descriptions of statistical outputs are subjective interpretations**, usually decision rules misrepresented as inference rules – which they are not, especially since decision rules require justified cost functions.
- **Worse, the verbal definitions and descriptions found in most primers, tutorials, and textbooks are flat-out wrong**, e.g.

  Cassidy et al. "Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly", AMPPS 2019

An urgent, neglected, yet easy step toward reform:

- Teach that **data transforms are all that statistics *can* supply**. Examples: P-values, posterior probabilities, interval estimates.
- **Observers supply meaning for statistics via causal models of the physical research reality.**

**Justified "statistical inferences" thus require**

- **showing how statistical assumptions can be derived from physical research reality; and**
- **showing where the data fall relative to what the assumptions taken together predict - which is what P-values do!** Leading to…

# **Reform step**: **Extricate P-values from the dung-heap of null-hypothesis significance testing**

- Both critics and defenders of significance testing misidentify P-values with their traditional use in NHST. This is like calling all knives "weapons": **It's confusion of a tool with its misuse** (as in use of P-values in NHST to murder thought).
- That confusion is destructive because **P-values can be used instead for many tasks divorced from NHST**: **as measurements of model fit, as estimation devices, and to build logical bridges connecting frequentist and Bayesian statistics**.

# Challenges facing P-value rehabilitation once dichotomania and "significance" are banished

- Instructors and users want P-values to be the probability of a test hypothesis H (usually, a null hypothesis of no association or no effect).
- A P-value is usually not near that probability.
- **Yet the teaching and research literature encourages subtle fallacious verbal descriptions that are equivalent to treating a P-values *as if* they were hypothesis probabilities ("P-inversion").**

# Ugly Fact: Valid interpretations of "inferential statistics" seem beyond most sources

- **The literature is filled with botched descriptions of P-values that confuse frequentist and Bayesian interpretations.**
- **Examples: inversions like "*P* is the probability the results are due to chance", nonsense like "*P* is the probability of a chance finding".**
- **Many descriptions of confidence intervals are actually defining posterior intervals, yet…**
- **95% "confidence" intervals typically get treated as nothing more than 5%-level tests.**

**Inversion fallacies** include misinterpreting *P*-values as probabilities that "randomness" or "chance alone" **produced** an association…as in Harris & Taylor *Medical Statistics Made Easy*,* 2nd ed, 2008, p. 24-25 say a *P*-value is

"**the probability of any observed differences having happened by chance**" (alone?)

- **If the tested ("null") model (of no effect or bias or mismodeling) is correct, what is the probability that a nonzero difference happened by chance alone?** Answer: **100%**

*(is "Made Easy" code for "Made Wrong"?)

- Sound analyses need to see results as very fuzzy, often in an asymmetric way. But,
- Concepts of evidence and uncertainty can only be quantified relative to explicit models to which the data supposedly pertain, e.g.:
  - Data contrasted against model predictions (compatibility checking = "tests of fit", as in **frequentist** diagnostics), or
  - Data merged with models to update predictions or bets (as in **Bayesian** posterior computation).

# Reconstruct statistical training:

## STOP perpetuating the mistakes of "great men" of statistics and the cognitive biases they displayed, created, and encouraged

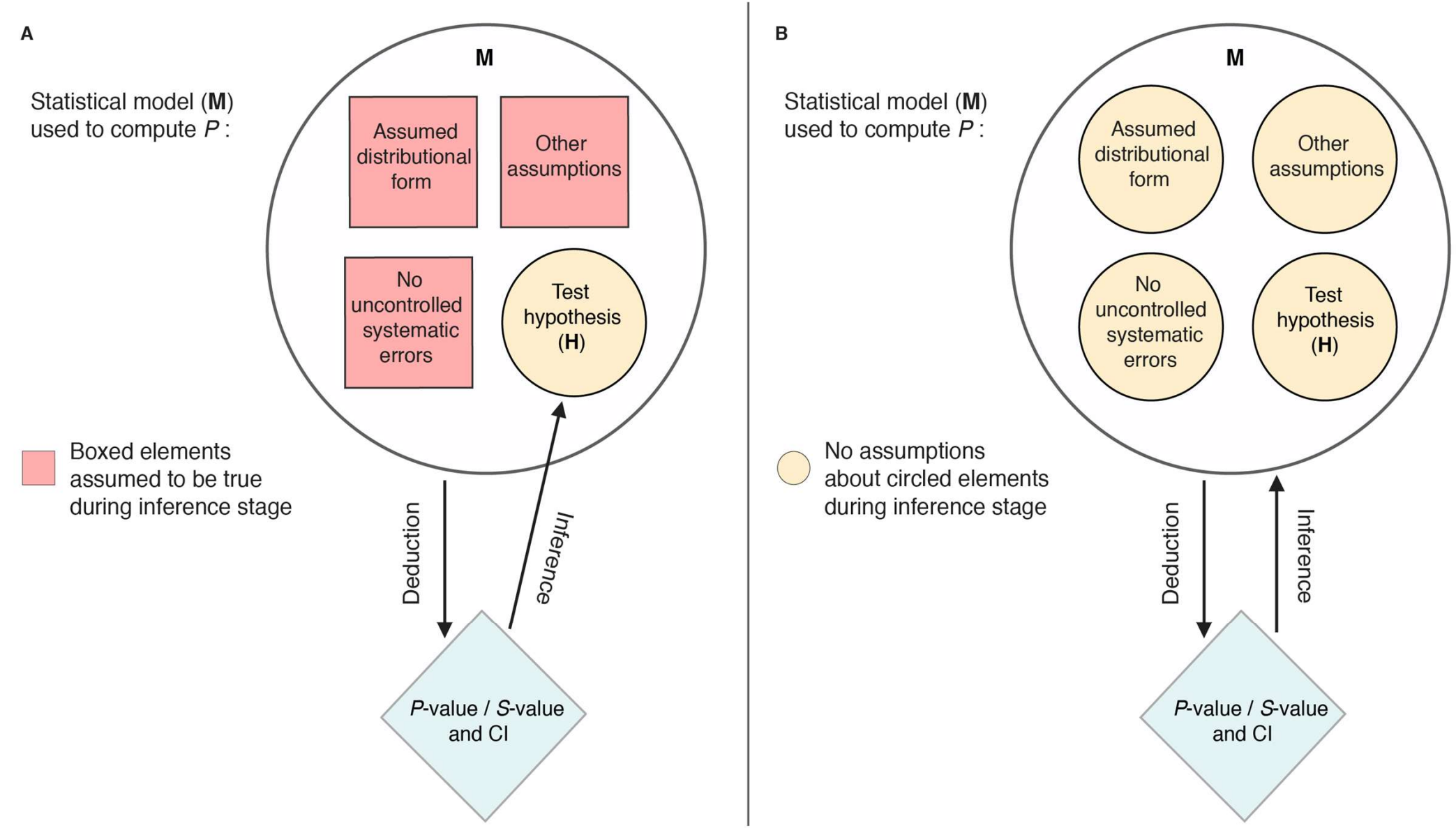
- Statistics texts assume instructors and students understand logic and semantics enough to see through **bad terminology** and to discriminate mathematical from contextual meanings.
- As shown by complaints from before Fisher's prime, **that was never true and only worsened in the mid-20$^{th}$ century research explosion**.

# Shift emphasis away from conditional "hypothesis-testing" interpretations to unconditional descriptive interpretations

- The norm: "**The P-value is the probability of getting a test statistic as or more extreme if the test hypothesis H is correct**", **which leaves the background assumptions (model) implicit**.
- Instead **bring the assumptions forward, as in**

  **A P-value p is the percentile at which the test statistic falls under the test model.**
- **The test model includes the test hypothesis H and all other assumptions used to compute p.**

# from Greenland & Rafi "Semantic and cognitive tools to aid statistical science" http://arxiv.org/abs/1909.08583



- **A test statistic measures discrepancy of data from predictions of a test model that contains both H and background assumptions.**
- **The P-value is thus computed assuming the entire test model, not just H.**
- **Deconditioning emphasizes that violations of *any* assumption used to derive the P-value can be responsible for its size**, not just violation of the test hypothesis **H**.
- **A large P-value does not support or confirm H or the test model (absence of evidence is not evidence of absence** of model violations).

# Overthrow misleading traditional jargon (Statspeak) to realign statistical terminology with ordinary language:

- **Replace** "significance" (Edgeworth 1885) and "confidence" (Neyman 1934) with **compatibility*** measured by the P-value p as it ranges from 0 = no compatibility to 1 = full compatibility of data with the test model used to compute p, in the direction measured by the test statistic. [*"Consistency" is nearly equivalent but is used for too many other concepts.]

- **Why? Because typical modern users depend on words – for them the maths are simply symbolic incantations they must take on faith to get funded and published.**
- **"That's just semantics" irresponsibly fails to grasp the essential analogical information conveyed by the semantics**. That failure is common among the mathematically able, who place syntax and deduction above analogical processes, or dismiss or overlook the role of analogy in mapping between reality and math.

# Stop repeating Fisher's error of using "null hypothesis" for any test hypothesis

**( an error which openly invites nullistic bias)**

**"Null" in English Dictionaries**:

- Oxford: adj. 2. **Having or associated with the value zero**; noun 1. **Zero**.
- Merriam-Webster: adj. 6. **Of, being, or relating to zero**; noun 7. **Zero**.
- Instead, use Neyman's term **tested (or test) hypothesis, and emphasize testing directional, non-null, and interval hypotheses instead of point null hypotheses.**

# Get rid of Neyman's "confidence trick"

- **Assigning high "confidence" is not distinct from assigning high probability.**
- So: Rename and reconceptualize "CI" as **compatibility intervals showing parameter values found most compatible with the data** under some compatibility criterion like P > 0.03 (which as shown below is about 5 coin-flips worth of evidence or less against any parameter value in the interval).
- **This involves no computational or numeric change! It's all about perception**…

## "Compatible" is far more cautious (and logically much weaker) than "confidence":

- There is always an infinitude of possibilities (models) compatible with our data. **Most are unimagined, even unimaginable given current knowledge**.
- We should recount the dogmatic denials by "great men" like Kelvin and Jeffreys of what became accepted facts.
- "Confidence" implies belief and encourages the inversion fallacies that treat the CI as a credible posterior interval. In contrast…

# Compatibility is no basis for confidence:

- **False stories can be compatible with data *and* lead to effective interventions**.
- Example: "Malaria is caused by bad air that collects near the ground around swamps."
- Implied effective solutions: raise dwellings, drain swamps – compatible cause (bad air) and actual cause (mosquitos) are both reduced by those interventions.
- But **confidence in the story will eventually mislead,** e.g., it leads away from use of nets.

# Problem: The stated ("nominal") coverage of a CI is a purely hypothetical frequency property in which we should have no confidence!

- **"Confidence" requires us to know for certain that the actual relative frequency with which the algorithmic interval covers the "true value" for the generator is as stated** (eg 95%).
- **But the actual generator frequencies are unknown, so no such confidence is warranted.**
- The stated coverage thus refers only to repeated draws from a **hypothetical** data-generating algorithm, **not** to a causal story we are sure of.

## In contrast, compatibility is merely an observed relation between data and models

- Compatibility only means the data set is "not far" (in percentile terms along the tested direction) from where it would be expected if it had come from the data-generating algorithm derived from the model under scrutiny.
- A 95% compatibility interval (or region) shows results for every model having $p > 0.05$ in the tested direction. This a region of "high compatibility" when translated into a simple coin-tossing experiment, as described below.

An honest report of Brown et al. JAMA 2017, "Association between serotonergic antidepressant [SSRI] use during pregnancy and autism spectrum disorder [ASD] in children", could be:

- Abstract: The Cox-model adjusted HR was **1.59**, 95% compatibility limits (CL) **1.17, 2.17**. Using IPTW HDPS, the HR estimate was much less precise (HR **1.61,** 95% CL: **1.00, 2.59**).
- Conclusion: **Under our HDPS model, the data appear most compatible with associations ranging from zero to a 2.6-fold elevation of ASD in children with *in utero* SSRI exposure.**

An honest report of Vallejos et al. "Ivermectin to prevent hospitalizations in patients with COVID-19" BMC ID 2 July 2021:

- Abstract: The hospitalization odds ratio was **0.65,** 95% compatibility limits (CL) **0.32**, **1.31**; the mortality odds ratio was **1.34**, 95% CL **0.30**, **6.07**.
- Conclusion: **The results were too imprecise to determine the size or direction of the effect**, being most compatible with hospitalization odds from 68% lower to 31% higher and mortality odds from 70% lower to 500% higher in the ivermectin group compared to the placebo group.

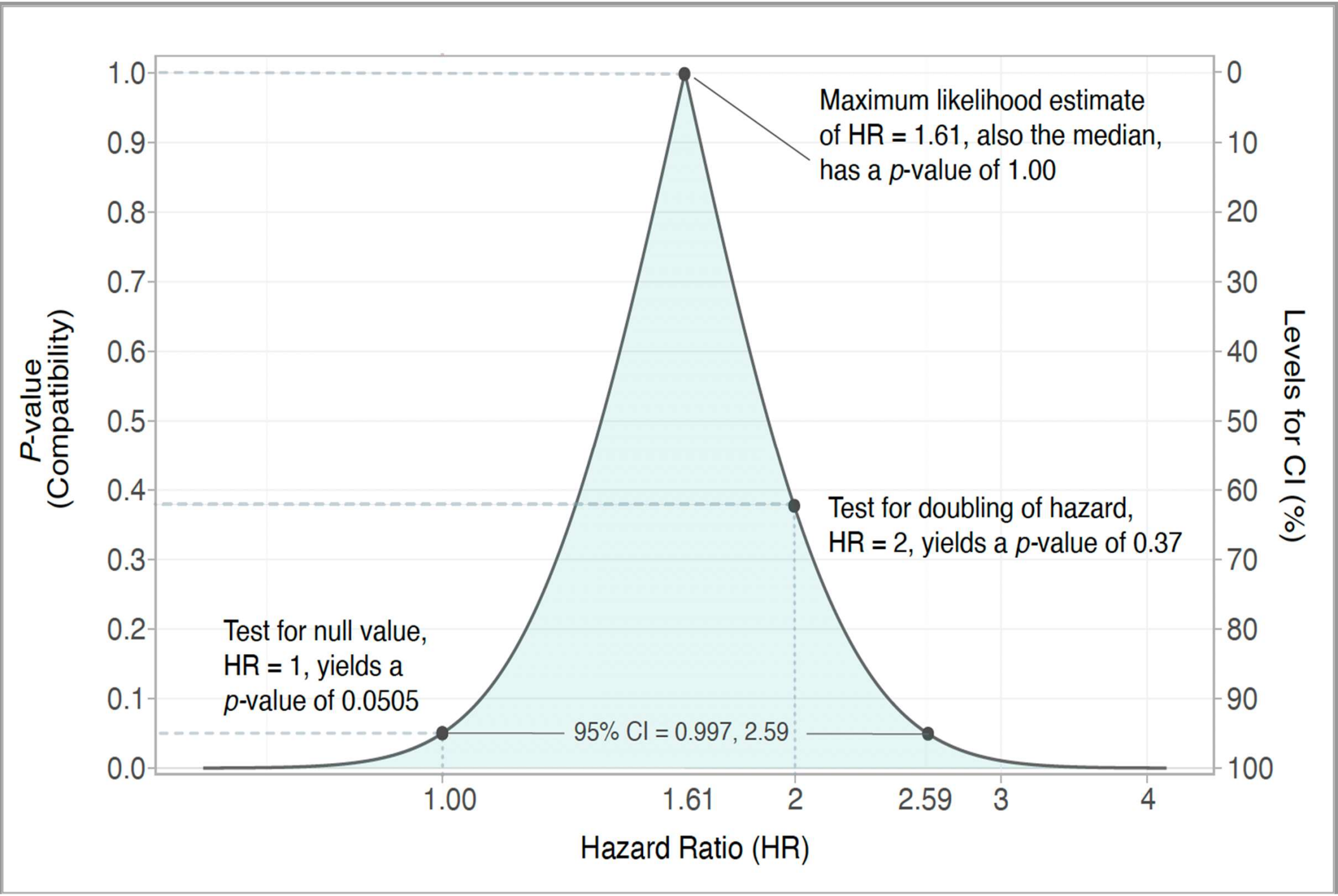
# STOP repeating the massive error of NOT treating P-values as estimation tools

**(another error openly inviting nullistic bias)**

**"The distinction between significance testing** [meaning: P-values] **and estimation is artificial"**
– Edwin Jaynes, Bayesian informationalist

- Indeed, **the distinction has been entirely destructive in focusing tests and decisions on just one hypothesis (the null) or model in an entire spectrum of hypotheses and models.**
- **Visualize P-values and CLs as indicating points on an entire P-graph** (P-value function)

# from Rafi & Greenland BMC Med Res Methodol 2020



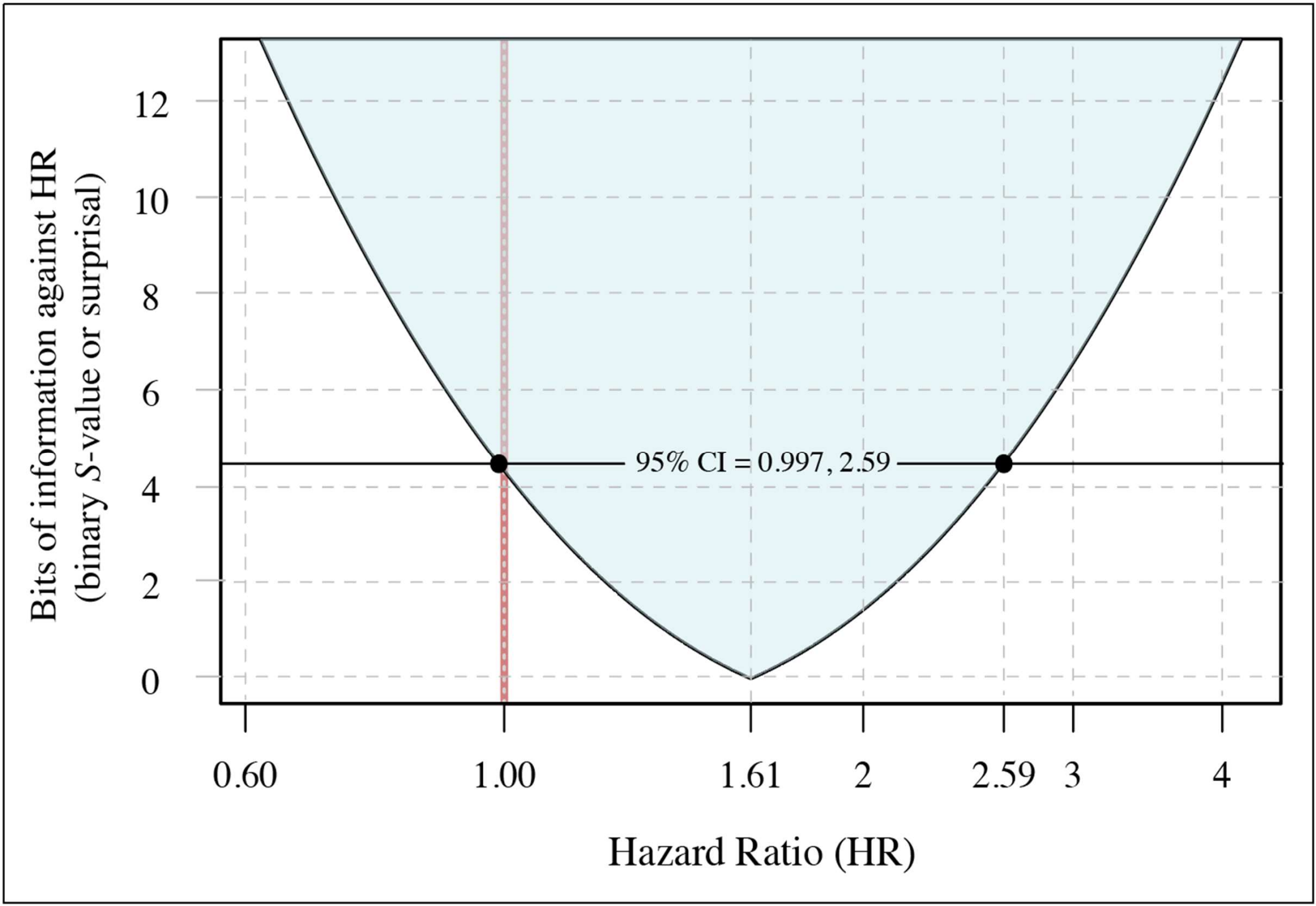
# TRANSLATE P-values to S-values (surprisals) to gauge the evidence supplied by test statistics

- A central aspect of the Fisherian treatment of P-values is their provision of a shared scale of evidence **against** hypotheses or models across different settings and tests.
- To express this scale in everyday currency, any P-value can be compared to the probability $\mathbf{½^s}$ of all heads from a sequence of **s** coin tosses that are independent and "fair" (chance of heads = ½)
- Given a P-value p, find the number of heads **s** in a row that gives back p …

- All heads in s tosses would give $p = ½^s$
- Solving for s gives $\mathbf{s = \log_2(1/p) = -\log_2(p)}$, so
- $p = ½^4 = .0625$ becomes s = 4 heads in 4 tosses
- $p = ½^5 = .0313$ becomes s = 5 heads in 5 tosses
- $p = 0.04 = ½^{4.6}$ becomes $s = -\log_2(.04) = 4.6$. Thus $p = 0.04 = ½^{4.6}$ provides the same evidence against the model used to compute p as about 4 or 5 heads in a row provides against the hypothesis that the tosses are independent with chance of heads no more than ½.
- **$-\log_2(0.05) = 4.3 \approx$ 4 heads in 4 tosses**
- **$-\log_2(.005) = 7.6 \approx$ 7 heads in 7 or 8 in 8 tosses**

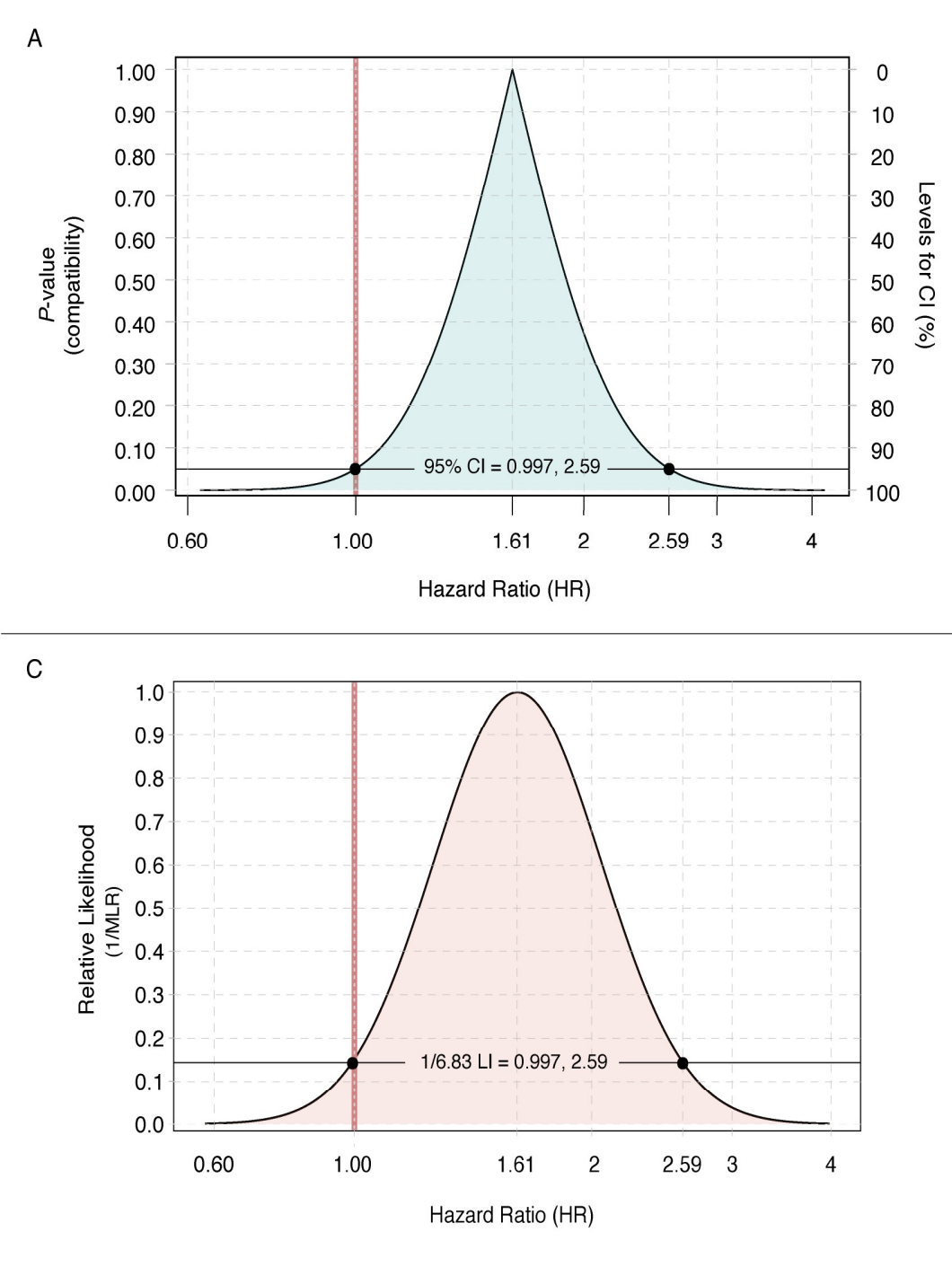
- **The binary S-value (surprisal, logworth) s measures the information the P-value p supplies against the model used to compute p**.
- **The units of s are** called Shannons or **bits**.
- **The P-value scale is highly nonlinear**: In terms of information against a model, **the difference between 0.001 and 0.05 is large, yet the difference between 0.95 and 0.999 is trivial**, despite their being the same distance apart.
- **S-values show their information difference**: $-\log_2(\mathbf{.001}) = 10$, $-\log_2(\mathbf{.05}) = 4.3$, $\Delta = \mathbf{5.7}$ bits; $-\log_2(\mathbf{.95}) = .07$, $-\log_2(\mathbf{.999}) = .01$, $\Delta = \mathbf{.06}$ bits

from Rafi&Greenland http://arxiv.org/abs/1909.08579



- S-values have resurfaced repeatedly since the 1950s when theorists needed to gauge the evidence or information in P-values, and examine test behavior under alternatives.
- **S-values are hard to confuse with Bayesian probabilities because they range far above 1.**
- **S-values do not require a prior distribution. But they *can* incorporate a prior distribution** by computing p from a test of fit of a compound sampling model that treats the prior as a parameter-sampling distribution (a "random-effects" model). Relation to likelihood ratios…

# from Rafi & Greenland BMC Med Res Methodol 2020



**Some background and further readings on my views**

(all should be open access at the links given)


Greenland S. For and against methodology: Some perspectives on recent causal and statistical inference debates. *Eur J Epidemiol*, 2017;32:3-20. https://link.springer.com/article/10.1007%2Fs10654-017-0230-6

Greenland S. The need for cognitive science in methodology. *Am J Epidemiol* 2017;186:639-645. https://academic.oup.com/aje/article/186/6/639/3886035

Greenland S. The causal foundations of applied probability and statistics. In Dechter R, Halpern J, Geffner H, eds. *Probabilistic and Causal Inference: The Works of Judea Pearl*. ACM Books 2022; 36: 605-624, https://arxiv.org/abs/2011.02677 (version with corrections)

Greenland S. Analysis goals, error-cost sensitivity, and analysis hacking: essential considerations in hypothesis testing and multiple comparisons. *Ped Perinatal Epidemiol* 2021;35:8-23. https://doi.org/10.1111/ppe.12711 20-01105-9

Greenland S. Valid P-values behave exactly as they should: some misleading criticisms of P-values and their resolution with S-values. Am Stat 2019; 73: 106-114. http://www.tandfonline.com/doi/pdf/10.1080/00031305.2018.1529625

**Educational readings for students, authors, editors and instructors**

Greenland S, Senn SJ, Rothman KJ, Carlin JC, Poole C, Goodman SN, Altman DG. Statistical tests, confidence intervals, and power: A guide to misinterpretations. *The American Statistician* 2016;70 suppl. 1, https://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/utas_a_1154108_sm5368.pdf

Amrhein V, Greenland S, McShane B. Retire statistical significance. *Nature* 2019;567:305-307. https://www.nature.com/articles/d41586-019-00857-9

Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics. *The American Statistician* 2019;73 suppl 1:262-270. www.tandfonline.com/doi/pdf/10.1080/00031305.2018.1543137

Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: Replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology* 2020;20:244 https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-01105-9

Greenland S, Rafi Z. To aid scientific inference, emphasize unconditional descriptions of statistics. 2021, http://arxiv.org/abs/1909.08583

- **'Pearl's thesis': Around 1900 science and statistics took a serious misstep in dismissing, even attempting to ban causation from theory.**
- A motive: Temporal symmetry in physical laws.
- Yet that overlooked the asymmetry emergent in thermodynamics, and the causal structure of information transmission as seen in $c$ = max speed of energy flow, ***c****ommunication*, ***c****ausation*.
- Causal (path) diagrams and potential-outcome models date from c. 1920, but did not fully develop and begin to spread widely until c. 1990.
- **They should be integrated into basic statistics!**

# Graphical example of a cognitive blindness: **The parsimony fallacy** to defend causal nulls in observational research

- Due to their qualitative nature, graphs say nothing about bias-variance tradeoffs, and so are often dismissed by those limited to pure predictive or potential-outcome models.
- Yet graphs show how statistical criteria fall short for evaluating causal effects, because **causality involves constraints that cannot be captured by probability alone**.

# Every missing arrow in a graph is an assumed causal null hypothesis

In "soft" sciences, we can rarely distinguish 'no effect' from alternatives that are within an interval around it containing important effects.

- Technically: A discontinuous distribution (one with mass concentrations) cannot be effectively distinguished *empirically* from a nearby continuous distribution. And…
- **The approximation error from replacing continuities with point masses can multiply through a causal network into huge errors.**

Ironically for those who deny specific effects are present on the grounds of parsimony, a null hypothesis is rarely the most parsimonious *causal* explanation for nonexperimental observations. In fact

- When *any* association is present, **the null hypothesis of 'no effect' is not parsimonious** because, under the null, the association requires indirect explanations, which are *causally* more complex than direct causation.

Suppose *causal parsimony* is defined as preferring the simplest causal diagram compatible with the observed (nonparametric) data distribution.

Then

- There is no basis for dismissing a reported effect without appeal to a more complex system of mechanisms that produces the association: The necessary causal diagram requires more arrows and larger effects.

Consider: If an X-Y association is observed, what is the simplest single explanation?:

a) Simple confounding: X←C→Y

b) Simple selection bias: X→[S]←Y

c) Differential error: X→X*←Y or X→Y*←Y

(X or Y observed with error as X* or Y*)

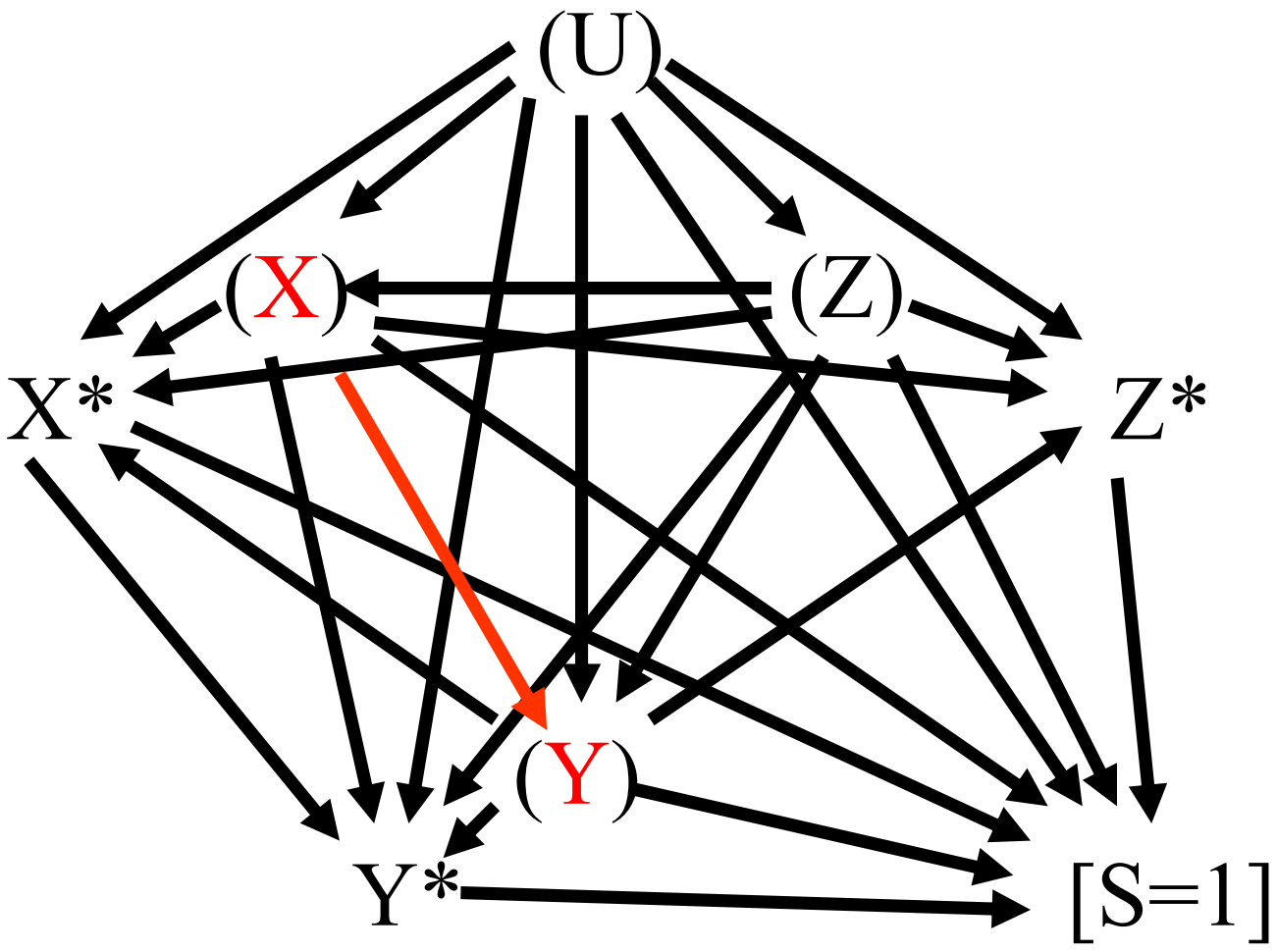
d) Simple random error: X  Y←ε

**e) Simple causation: X→Y**

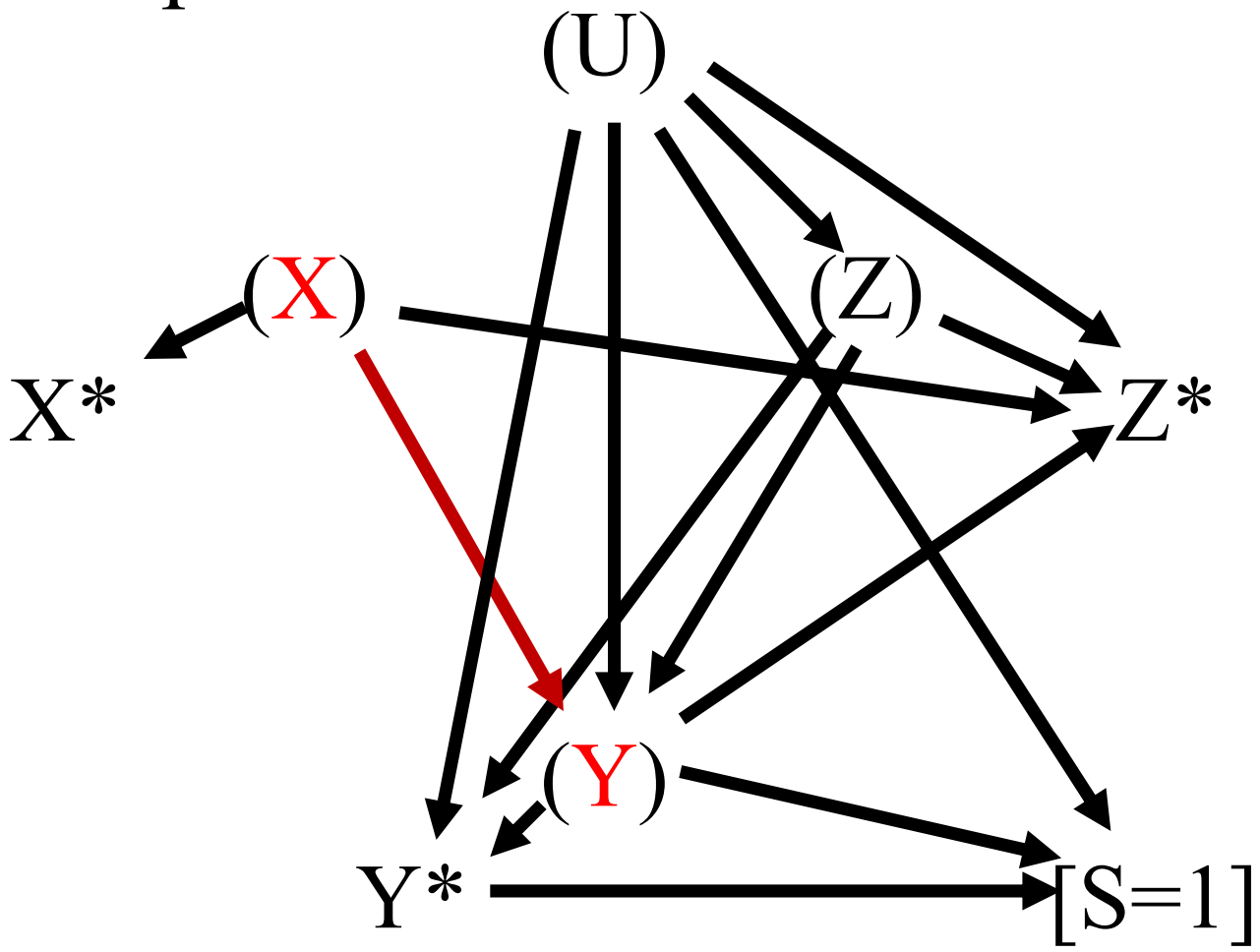
Asserting the null (a-d) requires an extra variable (node) or effect (arrow) relative to causation (e).

- In sum: If *any* association is observed (regardless of whether it falls within bounds for declaring it "nonsignificant"), **maintaining the null requires alternative explanations for the association**.
- Those alternative explanations are always more complex than the direct causal explanation (rejection of the causal null) if we define complexity as the minimum number of variables and arrows needed for the causal diagram (cDAG) of the explanation.

The complex observational reality: Any effect of X on Y is buried within a web of bias sources (confounding, selection bias, measurement error):



The simplest realistic DAGs with X*-Y* associations and hidden variables include those with X* and Y* d-connected *only* through X→Y, for example as in



- The hypothesis that there is no uncontrolled bias (no open noncausal path from X to Y) is the most parsimonious explanation one can provide for an observed relation of X to Y.
- Yet those who offer parsimony in defense of null hypotheses don't apply parsimony to biasing (open nondirected) paths and are supremely confident in more complex alternatives.
- This behavior reveals cognitive illusions influenced by hidden value biases toward the null (specific or general).

# Valid counterpoint: Parsimony is misleading when it fails to reflect essential complexity

Paraphrasing Neil de Grasse Tyson: 'Nature is under no obligation to be simple for you';

Twain: 'It ain't what you model that gets you into trouble, **it's what you don't model that's there**.'

- Every arrow missing between two graphed variables is a null hypothesis.
- Every exogenous variable missing from a graph represents a **set** of null hypotheses, one for every arrow from that variable to a graphed variable.

What **causally** warrants deleting arrows or nodes?

Answer: Forcing deletion by **causal** design – e.g., cohort matching (blocking), randomization.

- If X is randomized we can drop arrows to X.
- Random selection allows dropping arrows to S.

**But**, *by the definition* of observational studies,

- The study treatment X isn't randomized.

Furthermore, in health-science reality

- Selection and participation S is **not** random.

**No randomization = no 'objective' statistics**, only conditional statements of "under this model…"

- In “soft sciences”, prior distributions tightly concentrated near the null rarely have a basis in genuine evidence. They *may* have some support in some settings (e.g., genomics).
- *If* all causal paths were random walks, actual effects might cluster *near* nulls, making most effect sizes “unimportant”… But
- “Importance” is value laden. Declaring an effect to be exactly null buries this problem.

- Under continuity, there are almost no "false positives", because almost all **associations** are nonzero ("true positives").
- The "false-positive problem" is a distortive oversimplification of the problem of when to prune or ignore effects, which are decisions that require loss (penalty) functions.
- Effective pruning algorithms can retain prior continuity, e.g., use absolute-distance (LASSO, Laplacian) instead of squared-distance (Gaussian) penalties.