Proper presentation and interpretation of statistics for causality assessment and policy input: Modern methods and reforms for scientific inference in health and medical sciences are needed to address cognitive problems of statistics and the false claims those produce

Sander Greenland
Department of Epidemiology and Department of Statistics
University of California, Los Angeles

Key Points:

- **We need to learn how to systematically deal with and teach about cognitive biases, as we have done with mechanical biases like confounding.**
- **These biases are larger, more pervasive and socially more important than recognized in current methodologic texts and literature.**
- **Their coverage deserves to displace many finer points of statistical methodology, which itself is a source of cognitive biases.**

Big problems with our research environment:

- **Almost no statistical analysis accounts for all sources of uncertainty about inferential targets (such as effects). This failure often causes overstatement of conclusions.**
- **Most statistics primers and study reports suffer from misinterpretation of already unrealistic statistical results.**
- **The resulting mistakes get amplified in discussions, reviews, and press coverage.**
- **Often, *motivated reasoning* determines the direction of errors and biases.**

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**])." **Not given**: **p = 0.0505**
- Its conclusion: "**in utero exposure was not associated with autism spectrum disorder**"
- Their earlier meta-analysis got **HR 1.7 [1.1,2.6]!**

Example of a reformed presentation:

- **Adjustment using Cox regression produced an HR of 1.59, p = 0.003 for HR =1, and all HR from 1.17 to 2.17 had p > 0.05.**
- **Using instead IPTW HDPS for adjustment, the association was the same, HR 1.61, p = 0.05, but all HR from 1.00 to 2.59 had p > 0.05.**
- **In utero exposure was associated with autism spectrum disorder to about the same extent as in our earlier meta-analysis. The association may however be due to residual confounding or other uncontrolled biases** [they did say that].

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

- **Heroic narrative**: Science progresses by each generation challenging the ideas **and methods** of its predecessors, discarding those that fail stringent **empirical** tests.
- In contrast, **research statistics has decayed by overemphasizing methodologies for highly controlled experiments**, often defended with academic mathematical and philosophical appeals, while underplaying harms to actual research environments and public information.

Articles decrying null misinterpretation of nonsignificance date at least back to Karl Pearson **1906**:

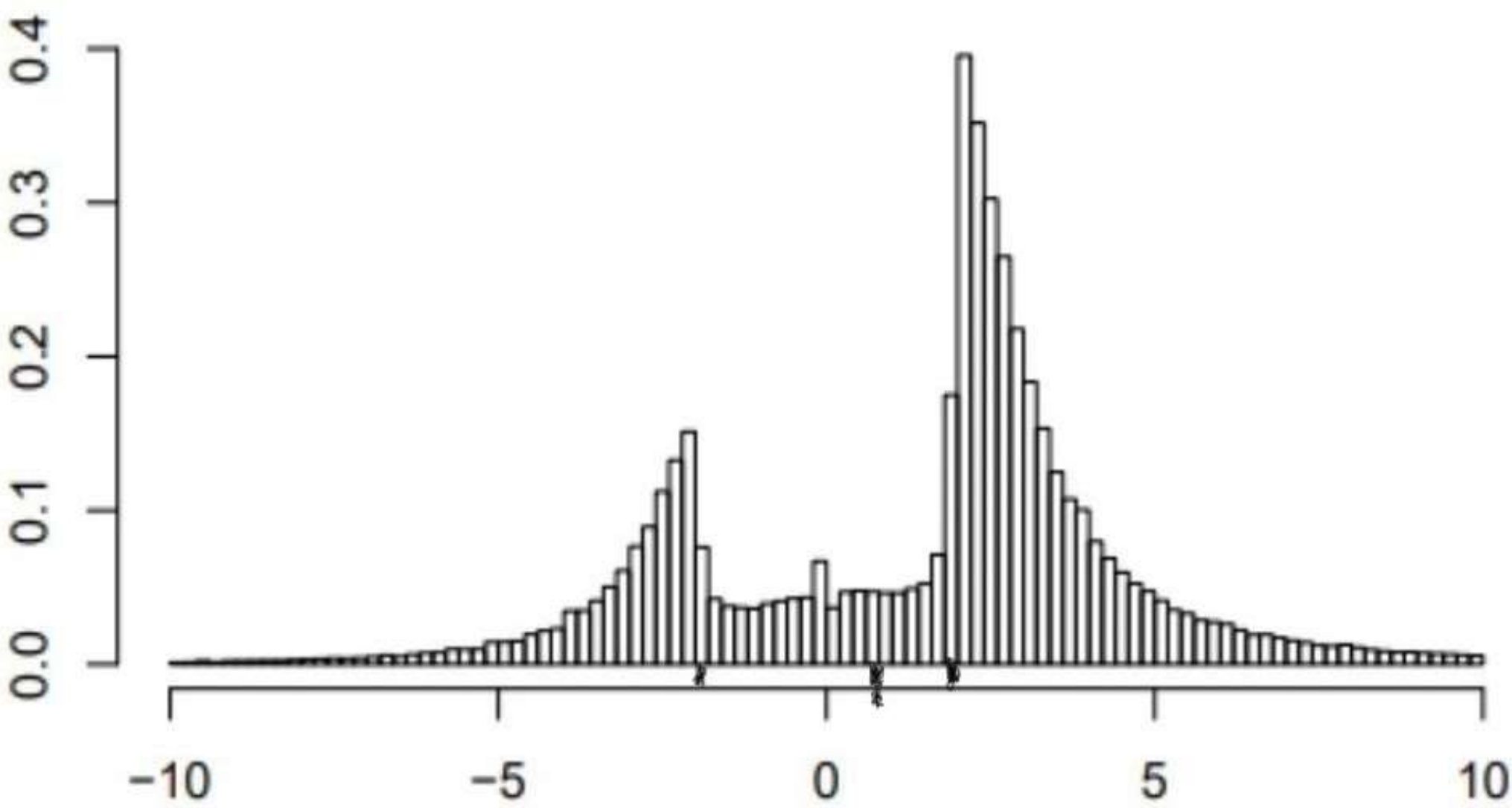
- "**The absence of significance relative to the size of the samples is too often interpreted by the casual reader as a denial of all differentiation, and this may be disastrous**."

Many others have repeated this caution since, including R.A. Fisher.

Why then does misreporting **ambiguous** results as "null" (**nullism**) continue and even remain enforced by some medical journals, despite…

The information damage from nullism

Fig. 1 from van Zwet & Cator 2021:
Over a million z-values from Medline 1976-2019.
Imputed histogram has >75% above 0



"…**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, U.S. NIEHS statistician, in "**It's Time to Rehabilitate the P-value**", *Epidemiology* 2001; 12: 288-290.

Nullism continues as a norm because it enables

- **an illusion of simplicity where none exists** (the religion of parsimony) **forgetting how "nature is under no obligation to be understandable to you"** cf. Tyson (Neil de Grasse, not Mike)
- **an illusion of learning and knowledge or certainty when your own study's information is actually sparse and your results are ambiguous**.
- **imposition of values and preferences of those who believe in or have stakes on the null, without revealing those values or stakes**.

Nullism is a cognitive and value bias, NOT a matter of statistics philosophy!

- **Declarations like "there was no association" when there was an association but p > 0.05 or the CI included the null aren't the fault of P-values and are not fixed by Bayesian "tests"**
- **They are instead the fault of a statistics and science culture that encourages or demands declarations of "findings" from ambiguous results, which most results are.**
- **This vice is synergized by lower publication prospects for honestly reported ambiguity.**

- **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 certain social networks to warp portrayal of facts into propaganda to bias inferences and policies in favor of the network's valuations and special interests.**

- **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.**

- **Statistical training and practice has been undermined by sanktification of cognitive biases (such asnullism and dichotomania) as "scientific principles"; treatment of mathematical frameworks as if physical realities (reification); and catering to human desires for certainty and finality.**
- **A mitigation: 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".

- **Reporting ambiguous statistical results as "negative"or "no association" generates enduring misimpressions that studies conflict, that RCTs refute observational studies, and that there is a "replication crisis" in health+medical sciences.**
- **It also creates spurious claims of conflict or refutation even though the studies agree, e.g., initial studies get $p<0.05$ and later, often smaller studies (as RCTs tend to be due their expense) get $p>0.05$.**

Typical example, Eur J Epid 2016;31:947-51:

- Abstract: "**use of statins was not associated with risk of glioma: OR for ≥90 prescriptions = 0.75; 95% CI (0.48, 1.17). Our findings do not support previous sparse evidence of a possible inverse association**"
- Discussion: "**This matched case–control study revealed a null association between statin use and risk of glioma.**"
- Prev. studies: **0.72 (0.52,1.00)**; **0.76 (0.59,0.98)**
- **3 combined: OR = 0.75 (0.62,0.90) p = 0.0016**

Example of a reformed presentation:

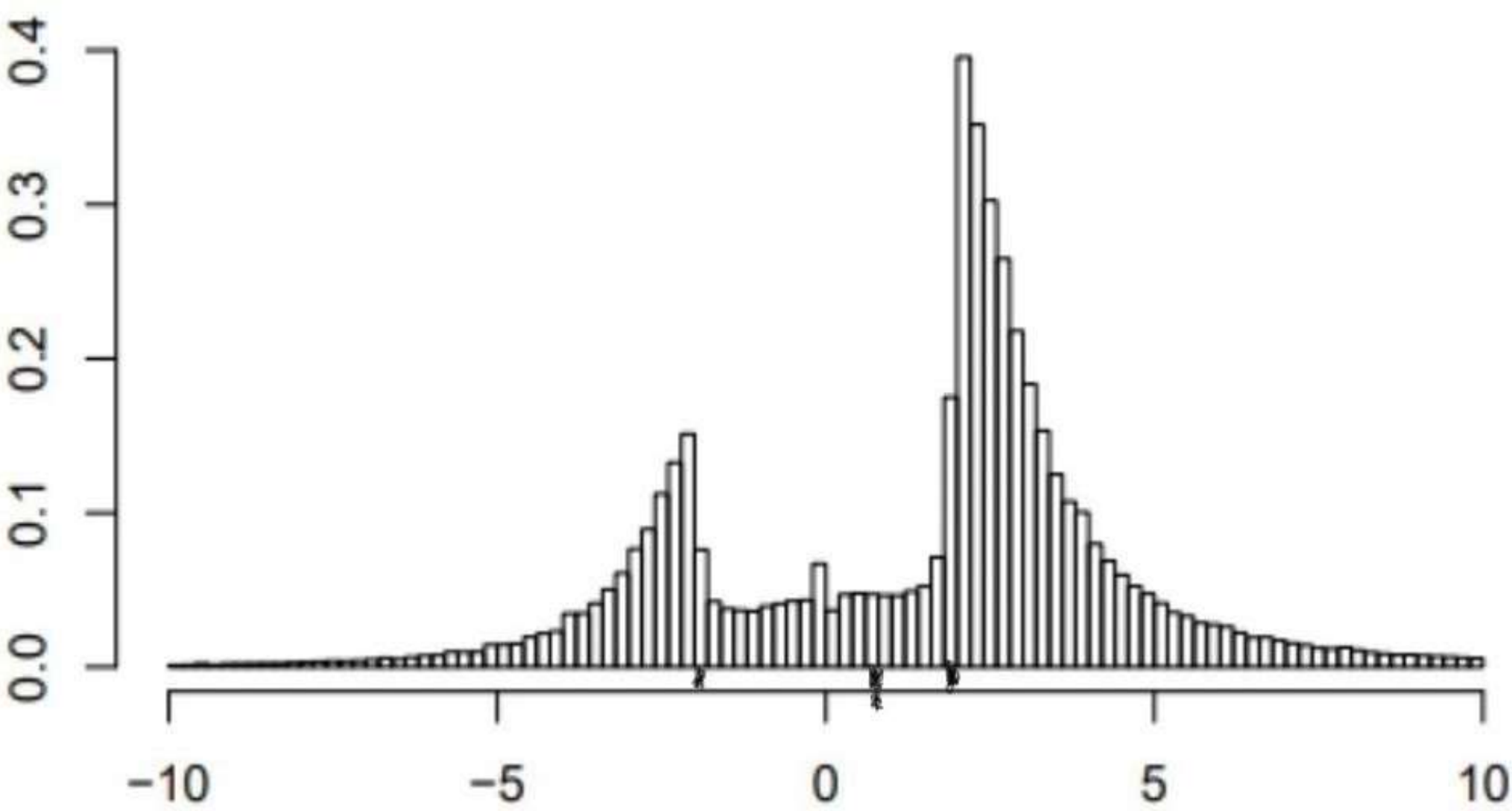
- **Use of statins was associated with risk of glioma: OR for ≥90 prescriptions was 0.75, but all OR from 0.48 to 1.17 had p > 0.05.**
- **The results agree closely with previous studies, which reported inverse associations of 0.72 (0.52, 1.00) and 0.76 (0.59, 0.98).**
- **When all 3 studies were combined the OR was 0.75 (0.62, 0.90), p = 0.0016 for no association. The association may however be largely or wholly due to residual confounding or other uncontrolled biases.**

- **There are many ways to analyze and interpret data that fall within what is acceptable to some reviewers.**
- **There are many analysis choices that must be made that are not dictated by universally accepted methods, guidelines, or rules.**
- **Social experiments in which the same data is given to different teams have resulted in a vast spectrum of results. Consequently,**
- **All analyses should be viewed as part of a vastly incomplete sensitivity analysis.**

What is the foundation of scientific inference?
- Not probability, but causation:

- Past causes: **What caused** ("explains") **our observations?** This is asking about **physical mechanisms**, *not* abstractions of their behavior such as probabilities.
- Future effects: **How will actions affect the future?** This 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 results will distort the distribution of available results relative to all results



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 jargon, 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.

Statistics education has neglected essential features of **scientific inference based on how**

- **causal networks including procedural problems(not probabilities) produce data,**
- **motivations, goals, and valuations (costs and benefits) affect cognition and are implicit in all methodologies,**
- **cognitive biases and social forces produce inferences and decisions,**
- **statistical analyses should be viewed as small points in a vast sensitivity analysis.**

- **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**

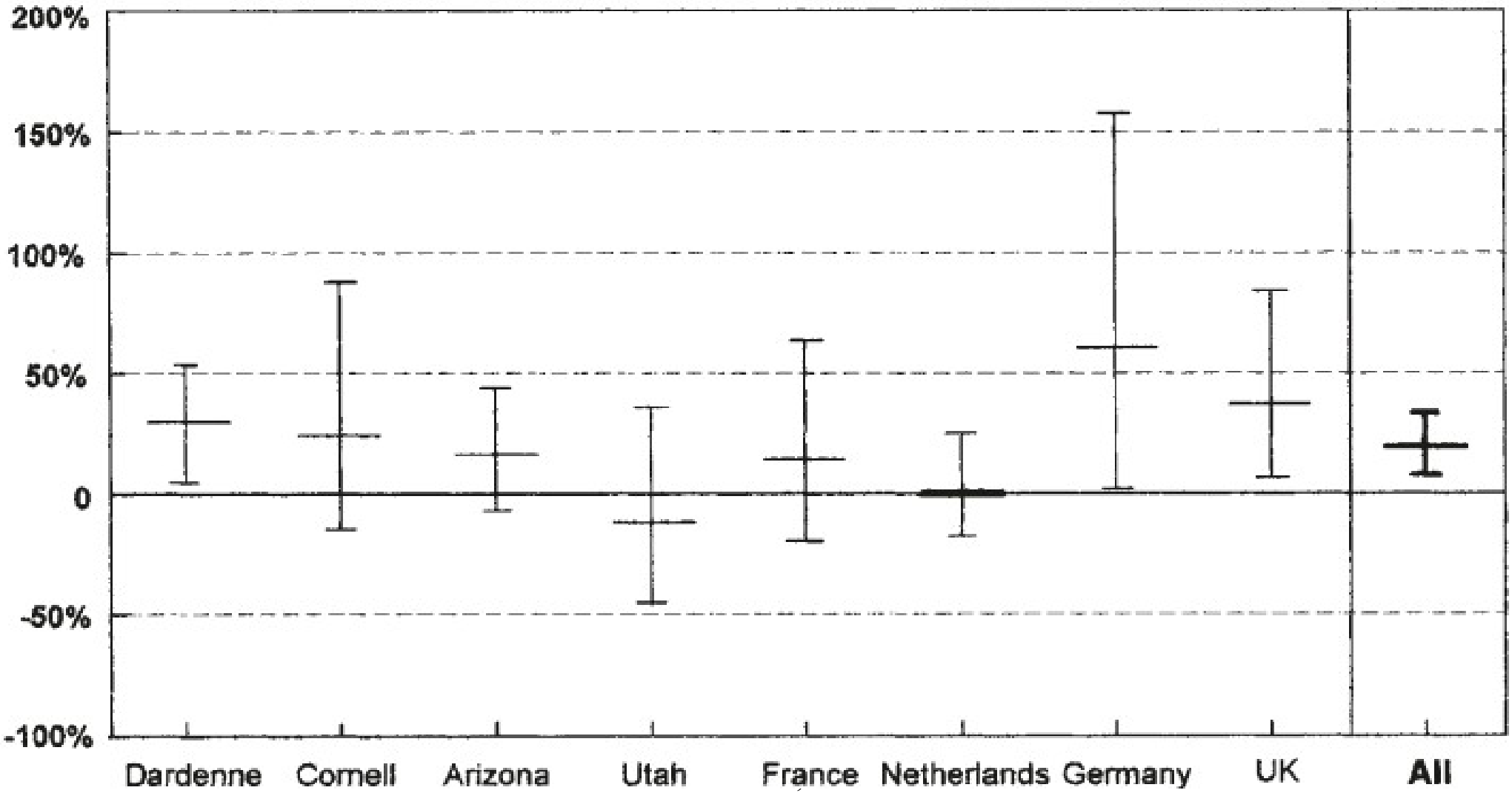
Example – labeling of U.S. dairy products:

"*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"

- A special-interest group forced a statement of fact to be accompanied by the misleading claim in red to defend rBST use.

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 publication bias.**
- **Any selective publication based on results 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 fish out and report **negative** results (**upward P-hacking**: selecting or focusing on results that give $p > 0.05$) **or misreporting of 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.**

Consider again 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]**

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 even 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?

One reason: 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.**"

It's also because they ignored or were unaware that **the hypothesis of no effect has no special plausibility**: embryonic neurogenesis involves serotonin signaling, and SSRI use during pregnancy has been linked to neural-tube defects.

- *Medscape* 2023: "prenatal SSRI exposure was consistently associated with 5%-10% lower brain volume in the frontal, cingulate, and temporal cortex throughout the age range studied." from *JAMA Psych* 2023; doi:10.1001/jamapsychiatry.2023.3161

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 and the null bias of null testing

- 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 CI width and P-values 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.

Reforms and tools to aid statistical inference:

- **To prevent confusion of statistical significance and statistical confidence with posterior probability: Replace them with compatibility interpretations** (1930s) **and surprisals** (1940s).
- **To prevent confusion of statistical (non)significance with practical (in)significance**: **present P-values for interval hypotheses** (1950s) **and for multiple alternative hypotheses** (1960s)

Some tools for scientific inference:

- **To aid identification of bias sources and proper adjustment covariates: causal diagrams** (1920s; as cDAGs: 1990s).
- **To account for uncertainty about uncontrolled observational bias sources: bias-sensitivity analysis** (1950s; 2000s).
- **To account for human cognitive biases and motivated reasoning: Work in progress, but includes statistical reform because…**

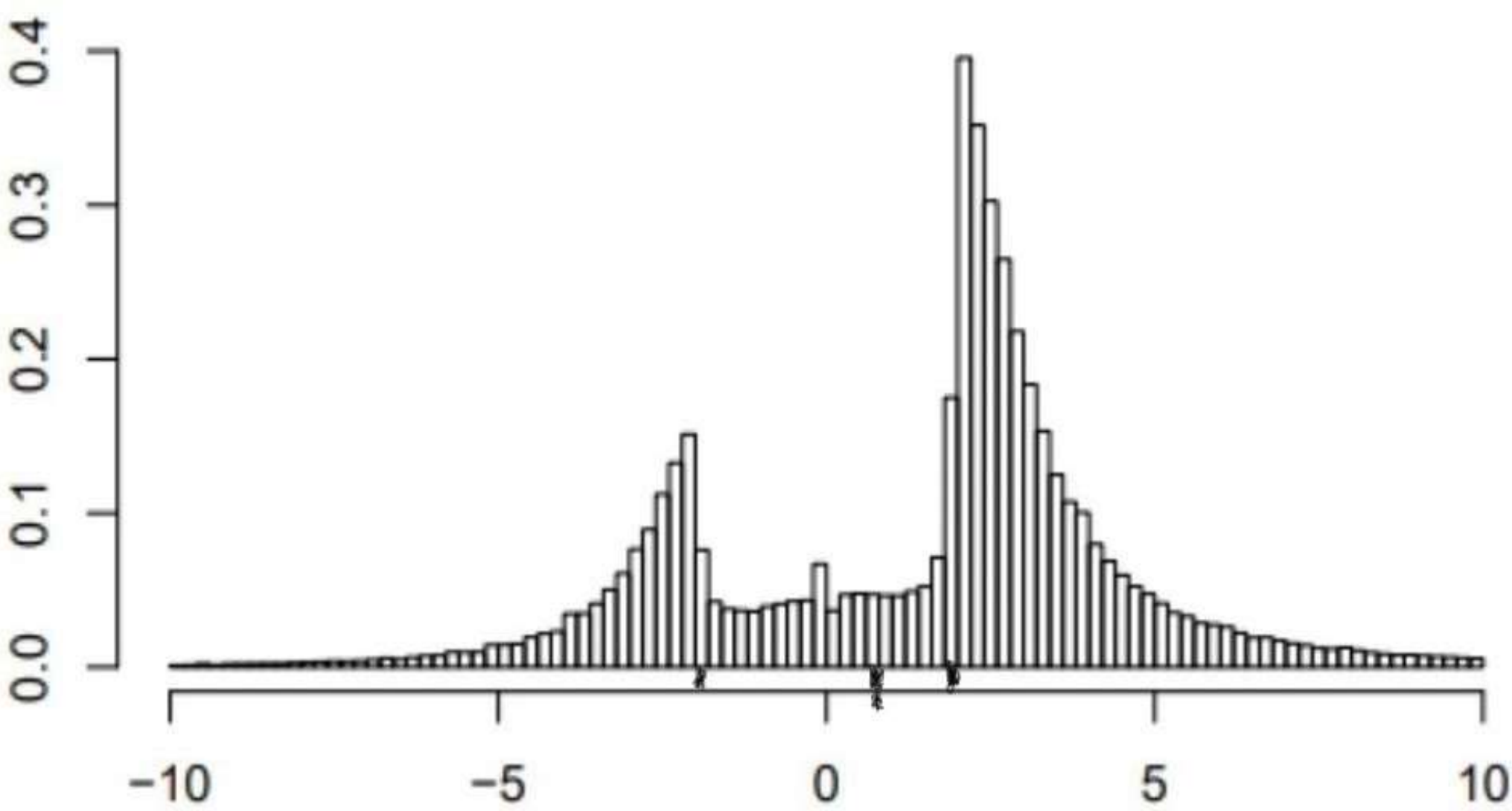
statistical training, traditions, and conventions are leading causes of cognitive biases and misreporting in research:

- McShane BB, Gal D. Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science* 2016; 62(6): 1707-18.
- McShane BB, Gal D. Statistical Significance and dichotomization of evidence (w discussion). *Journal of the American Statistical Association* 2017; 112: 885-908.
- McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. *The American Statistician* 2019;73:235-45.

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

- Past causes: **What caused** ("explains") **our observations?** which asks 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 encourages selective reporting based on study **results** 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**: Distortion of our perceptions by our need for parsimony and desire for simplicity.
- **Dichotomania**: Distortion of our perceptions by our innate compulsions toward decisiveness and black-or-white or qualitative thinking.
- **Reification**: Confusion of our models with reality – including faith that **formal methods** for reasoning, inference, and decision suffice for **real-world** reasoning, inference, and decision.

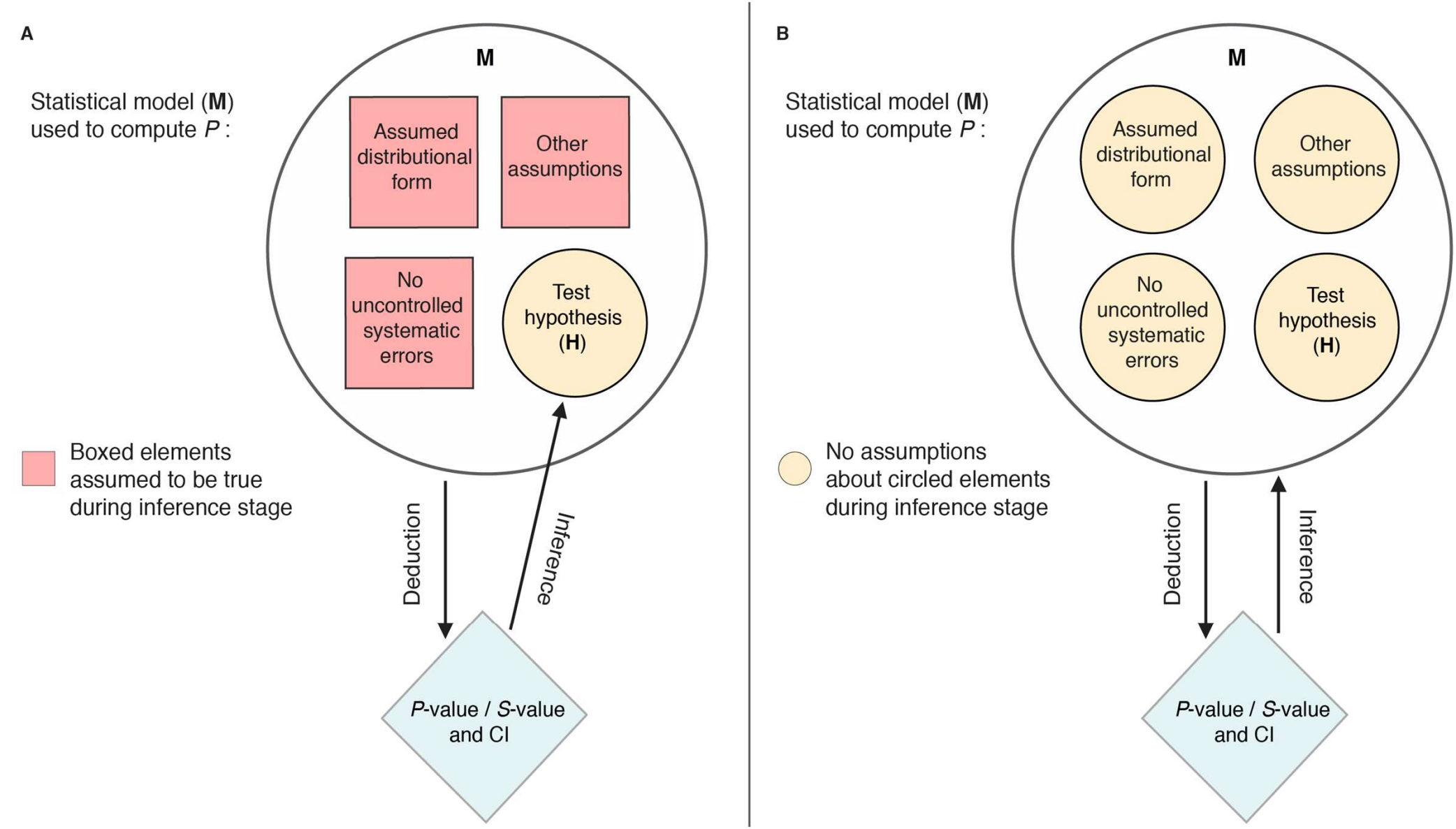
- **Against Nullism**: Reality is under no obligation to be parsimonious or simple.
- **Against Dichotomania: Many if not most important decisions are not or should not be binary**: Where do you set your oven? Your thermostat? **Your medication level**?
- **Against Reification**: Researchers routinely publish "inferences" that ignore vast model uncertainties – they usually aren't aware of most of the simplifications in their models and have no scientific rationale for using them.

To fight nullism, replace significance and hypothesis testing with P-values with unconditional descriptions of P-values

- 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 implicit**.
- Instead **bring the assumptions forward, as in**

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

from Greenland, Rafi, Matthews, Higgs http://arxiv.org/abs/1909.08583 :



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

- **Replace** "**significance**" (Edgeworth 1885) and "**confidence**" (Neyman 1934) with **compatibility** as measured by the P-value p, which ranges from
  0 = no compatibility to 1 = full compatibility of the data with the test model used to compute p, in the direction measured by the test statistic.

- **Why? Because typical modern users depend on words – for them, mathematics is simply symbolic incantation that 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 encouraged by those who place mathematics and deduction above analogical processes, ignoring the role of analogy in connecting theory to reality.**

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 recall 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 specified 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 shows results for every model that has $p > 0.05$ along the specified dimension. This is a region of "high compatibility" when translated into a simple coin-tossing experiment, as described below.

Some background and further readings on general methodology

(should be open access where links are given)

Lash TL, Heuristic thinking and inference from observational epidemiology. *Epidemiology* 2007;18:67–72.

Greenland S. Transparency and disclosure, neutrality and balance: shared values or just shared words? *J Epidemiol Comm Health* 2012;66:967-970.

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. 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 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 (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

Some educational readings for authors, reviewers, editors, students and instructors on reducing statistical misinterpretations

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

Greenland S, Mansournia M. Joffe, M. To curb research misreporting, replace significance and confidence by compatibility. *Prev Med* 2022;164, https://www.sciencedirect.com/science/article/pii/S0091743522001761.

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

Greenland S. Connecting simple and precise P-values to complex and ambiguous realities. Scand J Statist 2023;50:99-914 https://onlinelibrary.wiley.com/doi/10.1111/sjos.12645