Today's title is…
Meta-statistics:
Seeking treatments for the philosophical and psychosocial diseases of statistical science

Please send comments and corrections to
Sander Greenland
Department of Epidemiology and Department of Statistics
University of California, Los Angeles,
lesdomes@ucla.edu

Last year's title slide…

Toward restoring realism in statistical training and practice:

Preparing for the harsh realities of a "scientific" world in which "statistical inference" is often a device to manufacture desired conclusions,

or:

There are lies, damn lies, and statistics!

- Mark Twain did not originate that; among many others it was used by Leonard Courtney, President of the Royal Statistical Society 1897-1899

Some titles of earlier overlapping talks…

‘Advancing statistics reform: How to improve statistical science in the face of resistance’,

‘Cognition and causation before probability and inference’,

‘Breaking the tyranny of statistical authority over rational cognition’,

and my favorite,

‘There’s not much science in science’

Key observations:

- **Statistical analyses are based on vastly oversimplified models that omit major sources of uncertainty and that ignore most bias in data generation;**
- **They thus encourage overconfident "inferences" and conclusions;**
- **Statistics primers and study reports display overt misinterpretations of these unrealistic statistical analyses;**

- **The misinterpretations are amplified in reviews and popular "science communication".**
- **When the primary authors have stakes on the conclusions, *motivated reasoning* strongly determines the direction of misinterpretation:**
- **Favorable errors are overlooked or denied, unfavorable errors are vigorously sought out and eliminated.**

- **Cognitive biases are ritualized and institutionalized by statistical training, then are used to produce or support false claims or "inferences".**

Hence

- **We need to learn to admit and teach about investigator and cognitive biases, just as we do with mechanical biases like confounding, selection biases, and mismeasurement.**

- **Investigator and cognitive biases are large, pervasive, and important to society, yet overlooked by most methodologic texts and literature.**
- **In basic teaching and applied statistics, their coverage should displace many fine points of statistical methodology, *which is itself is a major contributor to cognitive biases*, especially nullism, dichotomania, and overconfidence,**

Cognitive and psychosocial biases are assumed absent or obscured by formalizations, and then ignored and institutionalized by conventional frequentist and Bayesian statistics…

Ex. A form of nullism: The bias of assuming all incentives are to "discover" rather than to refute effects. This meta-bias is rampant in the "replication crisis" literature, which ignores differences in incentives across topics and authors. For example,

- **Those invested in a treatment are biased toward reporting no excess of adverse side effects (ASEs)** (see SSRI example).

There are many other settings in which statistical norms will aggravate null bias…

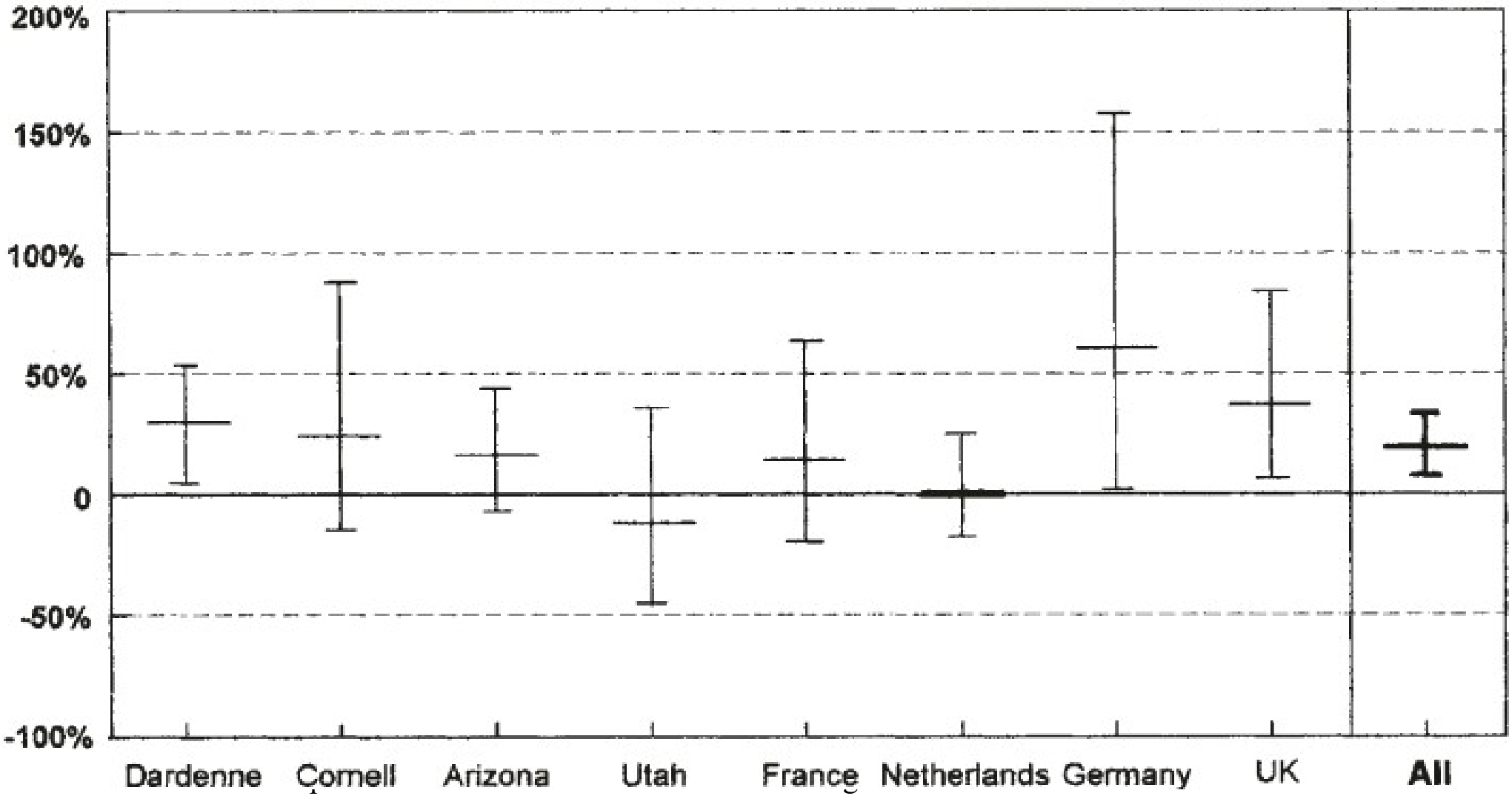
- ***Investigator bias*** **(motivated reasoning) and *social pressure* are among candidate explanations for conclusions in abstract, reports, and articles.**
- **Publicized "results" are a mix of targeted effects and biases; hence they are never fully unbiased or "reliable".**
- **Reasoning motivated by personal, legal, political and financial stakes fuels resistance to serious reform of statistical training and reporting.**

Example of political bias: A special-interest lobby forced a statement of fact to be followed by the false claim in red:

U.S. dairy products labeled
MILK from cows not treated with rBST* **are required to add this disclaimer:**

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

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



- **Reporting ambiguous statistical results as "negative"or "no association" generates spurious claims of conflict or refutation even when studies agree, as when**
  - **initial studies get p<0.05 and later, often smaller studies (as clinical trials tend to be due their expense) get p>0.05.**
- **Result: headline-grabbing *false* claims stemming from invalid uses of NHST, e.g.**
- **"most results fail to replicate" and**
- **"most observational studies get refuted"**

Ex. Claiming refutation from perfect agreement: Seliger et al. EJE 2016…

- "**use of statins was not associated with risk of glioma**."
- **Odds ratio (OR) for ≥90 prescriptions = 0.75; 95% CL 0.48, 1.17."**
- **"Our findings do not support previous sparse evidence of a possible inverse association."**
- **"[Our] study revealed a null association between statin use and risk of glioma."**

- **The reality: Their results agreed perfectly with the previous studies they cited, which reported ORs of 0.72 (0.52, 1.00) and 0.76 (0.59, 0.98).**
- **The summary OR across all 3 studies is 0.75 (0.62, 0.90), p = 0.0016.**
- **Hence there *is* an inverse association across studies, regardless of whether it reflects prevention by statin use, bias, or some combination of those.**

Example: Upward P-selection / CI widening ("null hacking") in Brown et al. "Association between serotonergic anti-depressant use during pregnancy and autism spectrum disorder in children" JAMA 2017:

- Cox-model adjusted HR = 1.59, 95% CL 1.17, 2.17.
- After switching to HDPS adjustment, HR = 1.61, 95% CL 0.997, 2.59.

- **They noted their 2017 meta-analysis of 6 studies reported HR 1.7 (1.1, 2.6).**
- **Not noted: Over a dozen other studies with summary HR as high or higher (Healy et al. 2016) DOI 10.3233/JRS-160726**
- **Their HDPS P-value for HR=1 was 0.0505, and all HR from 1.01 to 2.58 had P-values larger than 0.051.**

Instead, the paper and press concluded

- **"*in utero* exposure was not associated with autism spectrum disorder"!!**

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 of ambiguous results as "null" (nullistic bias) continue, even enforced by some medical journals?

Answer: "Human factors". Stat practice is plagued by researcher biases such as

- **Dichotomania: Even when a continuous picture is needed, practical limits force us to present dichotomizations (such as interval estimates) which are then mistaken for truth indicators or behavioral directives.**
- **Nullism: Even when there is insufficient evidence to reject even an *effect direction*, we will misinterpret ambiguous evidence as supporting no association or no effect.**

Ex. Incompetent press reporting: RCT of vitamin D (2,000 IU/day) and upper respiratory infection (Camargo et al. Clin Inf Dis 2024)

- **Abstract: "nonsignificant" OR=.60 (.28, 1.30) among <12ng/ml baseline group "requires further study".**
- ***p* = 0.096 for OR ≥ 1, hence a reference-posterior odds that OR<1 (preventive effect) will exceed 9 to 1**

- **Medical and consumer newsletters misinterpreted the trial as demonstrating no effect**, e.g.
- **ConsumerLabs: supplements "did not reduce risk in [those] who were vitamin D deficient (<12ng/ml) at baseline." FALSE!:**
- **The results are not at all definitive, but favor an effect among the deficient!**

- **Most of what I see reported as "the study found no association" is instead misreporting of ambiguous results, often pointing in a clear direction.**
- **This distortion could be mitigated by requiring students and research reports to tabulate or graph P-values across a range of parameter values,**

For example, **give P-values *p* for**

HR = 1, 1.25, 1.5, 2, 3 and / or for

HR = 1, 0.8, 2/3, 1/2, 1/3

- **That is easy!** With point estimate b and its standard error SE, P-values p for multiple values for β are easily computed from the Wald statistics (Z-scores) $Z(\beta)=(b-\beta)/SE(b)$.
- **Computing alternative P-values shows that, when using p to gauge compatibility with the data** (as in Pearson and Fisher),

➢ **many alternative β will have a higher p than does β=0 and so are more compatible with the data than the null hypothesis.**

➢ **taking β=b has p=1, so b is the value for β most compatible with the data.**

Overconfidence bias:

- **"People assign much higher probability to the truth of their opinions than is warranted."** – Kahneman
- **Statistical version: People assign much higher credibility to their own interpretation of data, models, and statistical results (or that of "trusted sources") than is warranted.**
- **Informative priors worsen the problem.**

- **As early critics feared, Bayesian methods open statistics to as much abuse as NHST via informative priors and (especially) via prior spikes:**
- **Informative priors can worsen overconfidence and null bias by narrowing and shifting interval estimates.**
- **Pr(null)=0.5 is *not* "indifference", it is a massive null bias!**

- **Elicited priors function as summary expressions of literature biases, misunderstandings, misconceptions, and group prejudices by overconfident and ill-informed "experts".**
- **They thus channel motivated reasoning, illusions of validity, and deep literature biases into statistical results.**

- As with NHST and multiple-comparisons adjustments, I distrust Bayesian reports in the medical literature because

- **their priors typically have unjustified overloading toward the null, and then they misreport their overshrunk and overprecise posterior results as if they were results from the study data…**

"Bayesian statistics is wonderful, until other people start doing it"

- **Example: Hayward et al. RCT of ivermectin and covid outcomes (J Infection 2024) reported posterior without giving the prior or likelihood…**
- **"Prespecified minimum meaningful difference" Hazard Ratio = 1.2, which they stated corresponds to a ~1.5 day reduction in self-reported recovery time**
- **Posterior HR 1.15, 95% limits 1.07, 1.23**
- **The posterior probability of HR > 1 exceeded the superiority threshold of 0.99**

- **Posterior Pr(HR ≥ 1.2) = .192 ≈ 1:4 odds for a 20% increase in the recovery rate**
- **Estimated median recovery-time difference: 2.06 days out of about 14 days average, 95% limits of 0.999 ≈ 1.00, 3.06**
- **Stated conclusion: "Ivermectin for covid-19 is unlikely to provide clinically meaningful improvement in recovery"**
- **Why is 1:4 odds for 20% or more increase in the recovery rate called "unlikely" and why is 2 days faster recovery not "meaningful"?**

- **Nowhere in the main text do they give the prior used to get their results.**
- **Nor do they give the analogous frequentist results to provide a sense of how much the prior influenced the results.**
- **That lack led me to incorrectly impute from the numbers in their Fig. 2 an MLE for the recovery HR of 1.19, 95% CI = 1.12, 1.26, p=0.41 for HR ≥ 1.2, and that a Jeffreys reference prior would yield a posterior Pr(HR ≥ 1.2) > 40%.**

- **A normal(0,SD^2) prior for β = ln(HR) that shrinks 1.12,1.26 to 1.07,1.23 has a prior SD of 0.061, 95% limits of 0.89,1.13.**
- **Under a Cox model it would take a perfect balanced randomized trial with an estimated HR=1 and about 4/0.061^2 = 1,075 observed events to produce an interval of 0.89,1.13. This is pure fiction; fortunately…**
- **that it is not what they used for β: From the supplement, the prior was normal(0, 0.3^2), described as "weakly informative" for β …**

- **A normal(0, 0.3^2) prior yields a 95% interval for HR = exp(β) of (0.56,1.80).**
- **It would take a perfect balanced randomized trial with estimated HR MLE of 1 and 4/0.09+1 = 45 events to get 95% CLs of 0.56, 1.80, slight compared to the 3640 observed recovery events.**
- **The same prior would however be *very* informative if it were applied to their hospitalizations+deaths analysis:**
- **That reported an odds ratio of 1.02 with 95% posterior limits of 0.63, 1.62 from 61 events.**

- **Thus the same "weakly informative" normal(0,0.3^2) prior would contain 45/61 > 74% of the information in the hospitalizations + deaths data.**
- **For either outcome, such a null-centered prior is not justified by previous trials; instead it seems to reflect societal bias against finding ivermectin may be effective.**
- **More generally: A prior is "weak" only relative to the analysis data.**
- **True reference priors represent only about 1 or 2 events worth of information.**

We thus should require Bayesian results from informative priors to provide both

- **summaries of the prior, *and***
- **summary results from the same sampling model, *without the prior*, which could be**

➢ **frequentist intervals and P-values; or**

➢ **posterior intervals and tail probabilities from a reference prior (e.g., Jeffreys, maxent); in our field those are very close to frequentist intervals and P-values.**

- **Reference results reveal how much the posterior was driven by the prior rather than actual study-data information.**
- **Typical informative priors represent opinions whose certainty far exceeds what is derivable from actual data (e.g., fair meta-analysis).**
- **This stems in part from biases being reinforced by social feedback loops.**
- **In the medical literature, these loops form an *echo-chamber effect*, exaggerating the content of "authoritative" opinions far beyond anything traceable to actual data.**

"Bayesian hypothesis tests" using a prior mass of ½ for a point null are an even worse deception than 0.05-level NHST:

- That mass translates to data with a likelihood function from an infinitely large experiment, which is then given massive (50%) weight in the prior.
- By any sensible measure the information in a point mass of ½ is far beyond what can be justified by medical literature.

- **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 distorted to see incentives for positive reporting while ignoring incentives for negative reporting…**

- **Again, the "replication crisis" is constantly portrayed as one of perverse incentives to make discoveries by searching out "statistical significance".**
- **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 outcome-based 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 the Brown et al. study of SSRIs and autism-spectrum disorder, JAMA 2017;317:1544-52).

From the 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"…**

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 the 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) doi:10.3233/JRS-160726**

- Why no discussion of the consistent summary association of ~70% higher risk when exposed?

One reason: the field was motivated to conclude that this highly replicated association was pure confounding (not merely that confounding contributed to the association):

- *Medscape* 2017: "**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 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 in 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 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.**

- **Again, 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 ignores differences in incentives across topics and authors, as well as the null bias of NHST.**
- **It also ignores CI-hacking to increase width by adjusting until the CI finally includes 1, even when adjustments beyond the initial model become overadjustments, inflating CI width and P-values without removing bias.**

Reforms and tools to aid statistical inference covered in detail far below:

- **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 nonequivalence hypotheses** (1970s) **and for several alternative hypotheses** (1960s)

There is a vast diversity of cognitive biases that we can only mitigate, not escape…

Empirical observation:

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: **When "The P-value is the probability of the null hypothesis" gets corrected to "The P-value is the probability chance alone produced the association" …**

- Those two statement are logically identical because "chance alone produced the association" *is* the null hypothesis!

Daniel Kahneman on "The Illusion of Validity" in *Thinking, Fast and Slow*:

- **"We can be blind to the obvious, and we are also blind to our blindness."**
- **"the person who acquires more knowledge develops an enhanced illusion of her skill and becomes unrealistically overconfident."**
- **I add: They may also see how "the experts" suffer from that illusion!**

Most relevant to statistics:

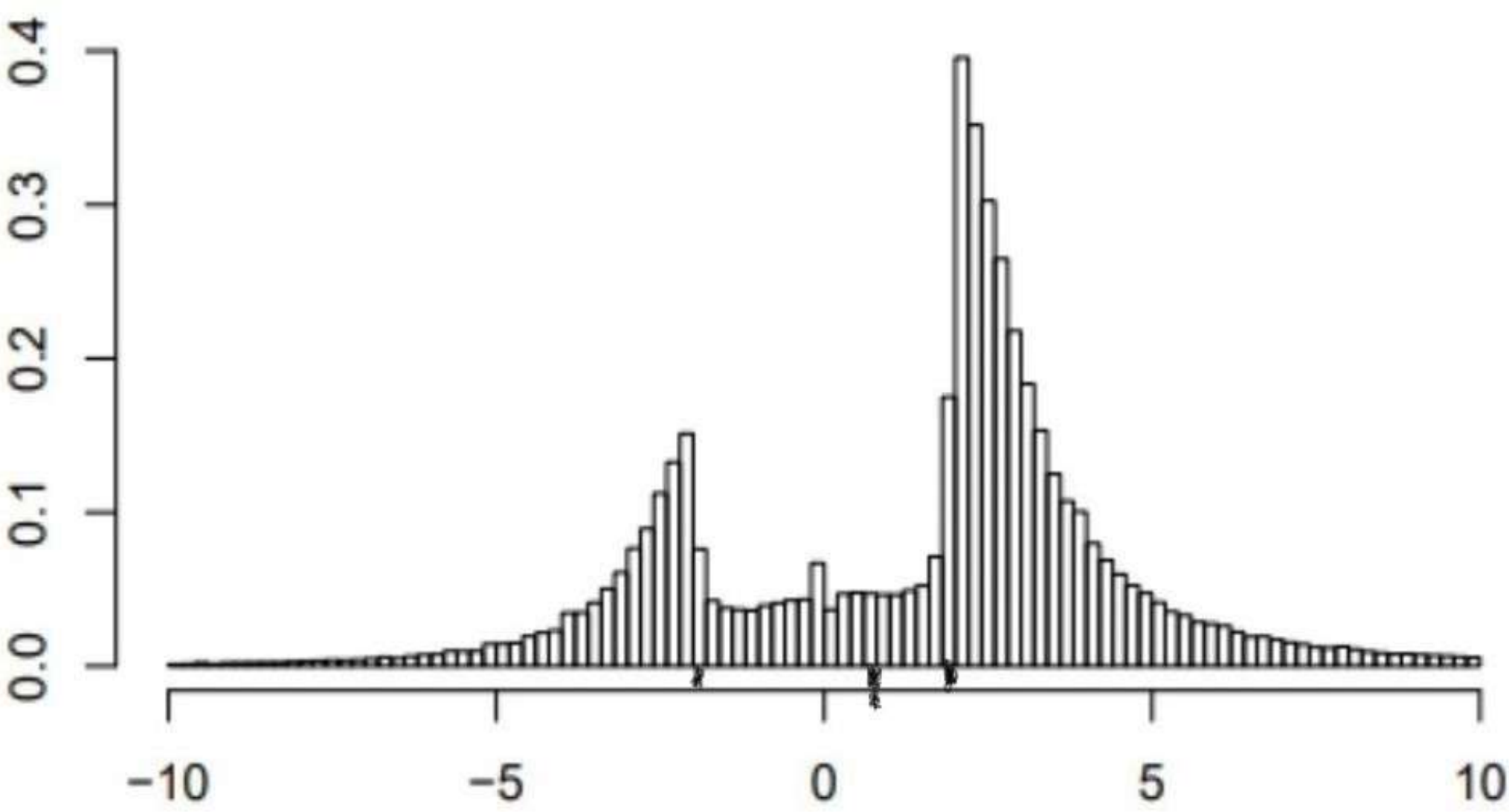
- "…**illusions of validity and skill are supported by a powerful professional culture. We know that people can maintain an unshakeable faith in any proposition, however absurd, when they are sustained by a community of like-minded believers.**"
- **For examples, see most any defense of null-hypothesis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

Here is one defense of NHST:

"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."

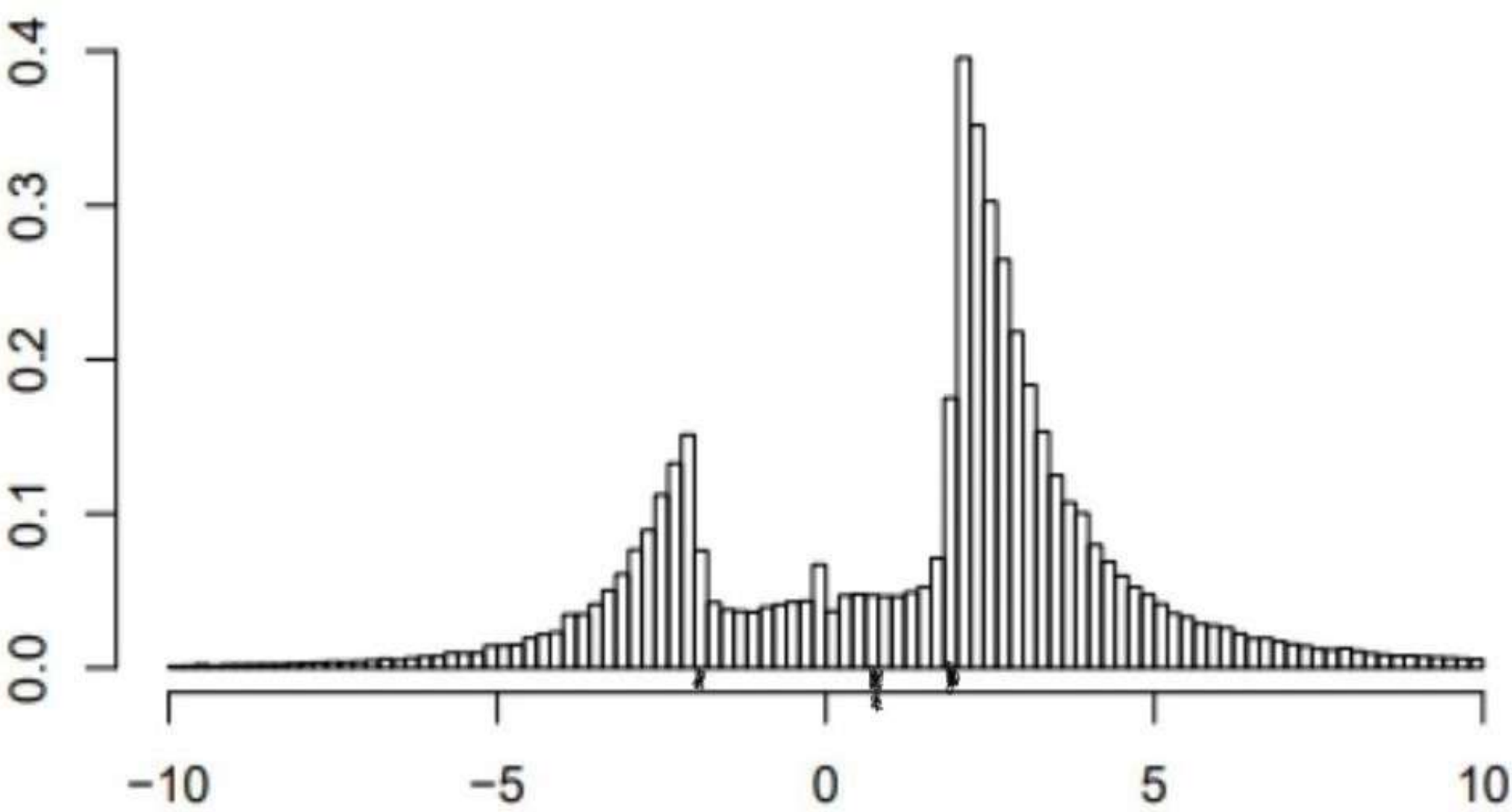
- Fisher 1920s? No, Mcnaughton 2021, *The War on Statistical Significance*.

Ignores that **any selective reporting based on study outcomes will distort the distribution of actual outcomes relative to the total:**



The information damage from NHST:

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



Frequentist "inference" uses no *explicit* prior, and so some claim the methods are "objective" or "let the data speak for themselves."

That is pure delusion because frequentist methods are filled with *implicit* priors, e.g., that every source of bias has been adequately accounted for and "controlled" so any remaining uncertainty is from "random error"…

Frequentism is a hotbed of mind-projection fallacies (reification):

- Imbuing inert quantities with attitudes, opinions, values, inferences, judgments, and decisions.
- Confusing math (logic, syntax) with interpretation (semantics).

- Rampant in discussions of results, thanks to using value descriptors like "significance", "confidence" and "severity" for narrow math concepts that don't capture ordinary meanings of those words.

Ex: "P-values overstate evidence" No!

- **P-values only provide the position of a statistic in a reference distribution (like a chi-squared) derived from a model. Any evidence overstatement is by the viewer.**

Ex: "the data speak for themselves" No!

DATA SAY NOTHING AT ALL!

Data are markings or bits that just sit there

- If you hear the data speaking, seek psychiatric care immediately!

Many if not most "study conflicts" arise from analysis and interpretation differences, because:

- **Many if not most analysis choices are *not* dictated by universally accepted methods, guidelines, or rules, and**
- **Those choices, including "pre-specified" design choices (such as priors), are influenced by group misconceptions and biases!**

Overconfidence bias: Kahneman –

- **"People assign much higher probability to the truth of their opinions than is warranted."**
- **Statistical version: People assign much higher credibility to their own interpretations of background literature, data, and statistics (or that of "trusted sources") than is warranted.**

- **Evidence: Experiments in which the same study data was given to different teams have resulted in a vast spectrum of results** (e.g., Silberzahn et al. 2018; Kummerfeld and Jones 2023)

Consequently,

- **Single analyses show nothing at all! …**
- **All analyses should be viewed as one very small and biased part of a vastly incomplete sensitivity analysis**.

Basic statistics should cover these essential if uncomfortable considerations for rational inference:

- **causal mechanisms *including bias sources* (not probabilities) are what produce data**
- **motivations, goals, and *valuations* (subjective costs and benefits) are implicit in all methodologies, perceptions, and reports,**
- **thus, cognitive biases and valuations affect actual inferences and decisions.**

- **Many cognitive biases contribute to design, analysis, reporting, and publication biases**

  https://en.wikipedia.org/wiki/List_of_cognitive_biases
- 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**.
- **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: uncritically emphasize desired evidence, while attacking or ignoring undesired evidence.**
- **Courtesy bias: Tendency to be obscure about criticisms that might cause offense.**
- **Failure to test alternatives** ("congruence bias")
- **Selective criticism** of undesirable evidence.

- **Selective reasoning to desired conclusions via selective assumptions, explanations, and data.**
- **Reification of mathematical validity: The tendency to think methods or judgments are as accurate about the world as they are in the mathematics used to derive them.**
- **Dunning–Kruger effects: The less expertise, the more the overestimation of one's competence (as when researchers, reviewers, and editors overestimate their own statistical expertise).**

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.**

- **Familiarity bias –** over-reliance on familiar methods, ignoring alternative approaches ("gets me grants and papers, so no need to change"). Leads into…
- **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 and group reinforcement effects causing overcount of evidence).**

And the most unpleasant and denied of all:

- **Value bias: We are all biased if not corrupted by our values, ideologies, and conflicts of interest…**
- Consider "saving lives". As wars and some pastimes show, not everyone wants "saving lives" as the chief goal or value of society.

Statistics as “the sick man of research methodology” and the simple ways we can move it towards health

Key Points thus far:

- **We need to learn how to systematically deal with and teach about psychosocial and cognitive biases, as we have done with mechanical biases like confounding.**
- **These biases are larger, more pervasive and socially more important than commonly recognized or acknowledged.**
- **They are easily amplified by statistics.**
- **Hence their coverage should displace many finer points of statistical methodology.**

- **Statistical methods are usually presented as logics that compel us to proceed from assumptions to conclusions, as surely as a computer program proceeds from inputs to outputs.**
- **Their theory focuses almost entirely on long trains of mathematical deductions.**
- **This approach is well suited for settings in which "human factors" are under complete control - but that is unrealistic in modern research environments.**

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 supplied us with vivid real-world examples – and no agreement on what those examples are.**

- **Statistical inferences only apply to the immediate mechanisms that generated the analysis data (including all the procedural and human biases).**
- **In typical controversies, inferences beyond that data-generating process to some underlying treatment effect requires assumptions that are not known to hold and are often known to be false.**

In the face of such unaccounted for uncertainties, clinging to "statistical inference" leads to reification:

- **Presenting deductions from a model as "findings", forgetting their sensitivity to uncertain assumptions.**
- **Mitigation: Replace "statistical inferences" and "statistical decisions" with *unconditional descriptions* of statistics, as in**
- **Greenland S (2025). Statistical methods: Basic concepts, interpretations, and cautions. In Ahrens W, Pigeot I, eds. *Handbook of Epidemiology*, 3rd edn. https://arxiv.org/abs/2508.10168**

- **Classical "science" purports that its over-arching goal is this: We are trying to see what is going on in reality, *regardless of the consequences*.**
- **That goal requires that we should develop contextually rich verbal *descriptions* of the data and the mechanisms that generated (caused) it before drawing inferences!**

- **What does a *neutral* researcher or reader want to learn about? Causal effects in a target population.**
- **What do statistics summarize? The data and its relations to mathematical models for the data-generating process (DGP).**
- **In controversial settings, the real DGP is unknown in detail and is NOT random sampling from the target population.**
- **So what uncertainty does statistical theory actually measure? Uncertainty about the behavior of the unknown DGP, not uncertainty about the target effect!**

- **How is statistics traditionally taught and practiced? It confuses the DGP with the target by starting with, focusing on, and treating as if real ideal-fantasy cases in which all uncertainties stem from "random variation". And so**
- **"…we ended up with an absurd dogs breakfast of an inference system that even Fisher or Neyman would have found ridiculous…"**

"…If I've learned nothing else from my research on cultural evolution and iterated learning, it's that a collection of perfectly-rational learners can ratchet themselves into believing foolish things, and that the agents with most extreme biases tend to dominate how the system evolves." – Danielle Navarro,

A personal essay on Bayes factors 2020/2023
https://blog.djnavarro.net/posts/2023-04-12_bayes-factors/

- **Because of their deductive form, statistical methods get treated as if oracles of truth, instead of the thought experiments they are.**
- **The truth they are claimed to reveal is supposedly cautioned by interval estimates. But those are too narrow!**
- **Statistics thus encourages overconfidence when, as usual, we can't be certain about the explanations for the reported statistics.**

Teaching that claims to cover the basics of inference needs to include cognitive science to deal with delusions and biases such as

- **Nullism: Distortion of perceptions by our need for parsimony and desire for simplicity.**
- **Dichotomania: Distortion of perceptions and inferences by our compulsions toward decisiveness and black-or-white thinking.**
- **Reification: Confusion of our models with reality – including treatments of formal methods as if they always aid or (worse) suffice for real-world 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 simplifications in statistical models and have no valid rationale for using them.**

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,** with this among central assumptions:

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

My perception ← Reality → Your perception

- **Thus, *valid inference* needs cognitive science!**

Explanatory ("inferential") statistics – that is, **statistics grounded in causal thinking – requires accepting that**

- **Without unbroken randomization, "chance" does not explain anything,**

and that

- **"Could be due to chance" is jargon for "some uncontrolled or unimagined or unmentionable force might have produced this association".**

Instead, "statistical inference" became a distorted caricature of scientific inference

- **It degenerated into taking output from data-processing programs** (machine-learning algorithms) **and generating "inferences" from those via rigid, decontextualized rules**.
- **It converted oversimplified models of *causes* of the data (data-generating mechanisms) into decontextualized probability functions**.
- **The semantic void it left produced rampant misinterpretations and enabled deception.**

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, academic statistics has focused on generating context-free "methodologies"** (formal systems), **without effective safeguards to prevent their harms to actual research environments and to public information.**

We need accurate, honest coverage of history and methods. Example: NHST

- **A worst-choice hybridization of Fisherian and Neymanian ideas, with elements that one or both would condemn.**
- **Pretends that mechanical decision rules derived from uncertain assumptions and hidden loss functions are an oracle for binary declarations of detecting or denying associations; it thus hides the continuous, subjective nature of uncertainty and loss.**

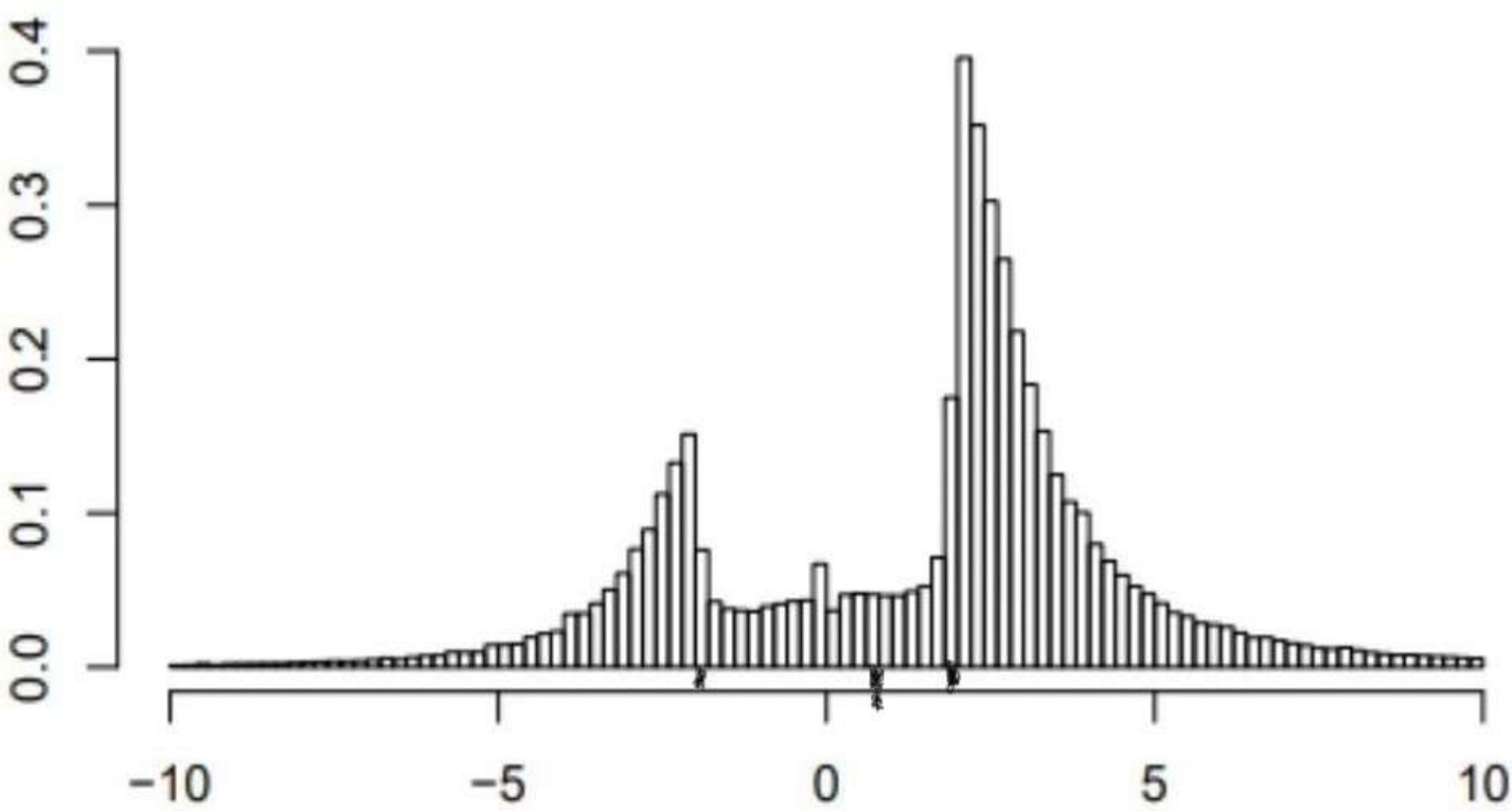
- **Extensive data and documentation that NHST and "statistical significance" is grossly misunderstood and misused by most researchers, and as a result:**
- **NHST has warped the research literature.**
- **Yet NHST has become a religious institution of "science" defended by the highest authorities (who have taught and relied on it throughout their careers), always with the empirically refuted defense that "we just need to teach it better".**

Items ignored in conventional NHST:

- **Fisher maintained that a "significance level" (his P-value) should serve only as advisory input for inference and decision, not a final arbiter. Egon Pearson agreed.**
- **Neyman maintained that the tested hypothesis should be the one most costly to falsely reject, *not* defaulted to the null of no association or no effect.**
- **All regarded cutoffs (α-levels) as needing contextual justification (Neyman: through explicit decision costs or loss functions).**

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



With all the misconceptions and abuse afoot, why focus on terminology? Because:

- **We depend on verbal descriptions to connect mathematics to the application.**
- **In "soft sciences", the math is always an oversimplified description that gets confused with reality (*reification*).**
- **Bad terminology creates misconceptions that synergize with *wish bias* to inflate and perpetuate bad practices – as in confusing "statistically nonsignificant" with no effect.**

- **"That's just semantics" irresponsibly fails to grasp the essential mapping of statistics to reality encoded in the semantics (words).**
- **Such irresponsibility is encouraged by prioritizing mathematics and deduction over valid mapping between our *unrealistically precise* abstract theory and the *messy reality* that generated the data.**
- **Again: statistical analyses are only thought experiments of the form "under these assumptions, we get these probabilities…"**

- **Yet statistics has ignored 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 apparent successes in narrow automated environments induced destructive feedback loops in teaching and research, since

- **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 rules to ease reviewing and publication decisions.**

The prevailing semantics became especially popular and destructive via enforced dichotomization of inference

- **Dichotomies satisfy human drives for definitive conclusions, because 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.

Null preference is a cognitive and value bias, NOT a statistical or philosophical principle!

- **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" – even from ambiguous results, *which most results are*.**
- **This vice is synergized by lower publication prospects for honestly reported ambiguity.**

Nullism endures as a norm because

- **it enables an illusion of simplicity when reality is too complex to model credibly, forgetting how "nature is under no obligation to be understandable to you"**
- **it creates an illusion of learning and certainty based on study results that are ambiguous (convey limited information).**
- **it allows the imposition of the values and preferences of those who believe in or have stakes on the null, without having to recognize or reveal those values or stakes.**

So: Stop repeating Fisher's error of using "null hypothesis" for any test hypothesis H

(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, following Neyman, **use *tested* or *targeted* hypothesis, and from the start discuss non-null, directional, and interval H instead of only point null H.**

More generally: Overthrow misleading traditional jargon (Statspeak) to realign statistical terms with ordinary language

- **Rescue the P-value from "statistical testing" by reframing it as an ordinal index of compatibility with data, applicable to *any* hypothesis H or model (not only nulls!).**
- **If a study reports "there was no significant difference", require it also report the P-value for a small but important non-null difference (e.g. a 10% survival difference).**

- **Replace "statistical significance"** (Edgeworth 1885) **and "confidence"** (Neyman 1934) **by *compatibility*** (Pearson 1900) **of the data with the statistical model used to compute p, where that model is composed of *every* assumption made in the computation, not just the targeted H.**
- **"CI" now means "compatibility interval".**
- **Small p now indicates incompatibility of the data with the model along a specific direction defined by conflict with H.**

"Compatibility" is far more cautious and logically much weaker than "confidence":

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

Compatibility is no basis for confidence:

- **False stories (models) can be compatible with data *and* lead to effective interventions. But,**
- **<u>Confidence</u> in a story will eventually *mis*lead**
- **Ex.: "Malaria is caused by bad air that collects near the ground around swamps."**

➢ **The story (model) implies effective solutions: its hypothesized cause (bad air) and the actual cause (mosquitos) are both reduced by raising dwellings and draining swamps;**

➢ **Yet the story misleads us about bed-net use.**

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

- **"Confidence" requires us to know with certainty the actual frequency with which the interval covers the "true value"** (eg 95%).
- **But when uncertain assumptions are used (as usual) the *actual* frequencies are unknown, so no such confidence is warranted.**
- **The stated coverage thus refers only to draws from a hypothetical probability model, not to a known data generator.**

In contrast, compatibility is an observed relation between the data and the model

- **Compatibility only means the data set is "not far" from where it would be expected if its generating mechanism followed the model being used or evaluated.**
- **A 95% compatibility interval shows results for every model in a family that has $p > 0.05$ along a specific parametric direction.**
- **The interval thus defines a range of models "highly compatible" with the data along a parametric direction in the space of models.**

So: Get rid of Neyman's "confidence trick"

- **Assigning high "confidence" is not distinct from assigning high probability.**
- **Rename and reconceptualize "CI" as compatibility intervals showing parameter values found most compatible with the data under a criterion like $P > 0.05$, which represents $-\log_2(.05) \approx 4$ bits or less information against the parameter value.**
- **This involves no computational or numeric change! It's about changing perceptions…**

To recap the problems being addressed:

- **Medical research always involves uncertainty about the data generator.**
- **Statistical methods always assume that the generator is known *with certainty* to follow strong assumptions like random selection, assignment, loss, and measurement error given adjustment covariates.**
- **When an assumption is uncertain, the statistical results will fail to reflect this source of uncertainty.**

- **In the face of uncertainty, use of the label "inferential statistics" is a deception, for then**
- **Statistical inferences only follow under conditions that are not known to hold and are often known to be false.**
- **Clinging to "statistical inference" has led to reification: Presenting deductions from a model as "findings", forgetting their sensitivity to our uncertain assumptions.**
- **Mitigation: Replace statistical decisions and statistical inferences with *unconditional descriptions* of statistics.**

Replace statistical testing and estimation with *unconditional descriptions* of statistics

The norm:

- **A P-value is the probability of getting a test statistic as or more extreme *if* H is correct"**
- **"A CI is an interval with 95% probability of covering the true value".**

➢ **Both leave the background assumptions implicit – and their uncertainty ignored.**

➢ **Those assumptions compose the statistical model from which p and CI are computed.**

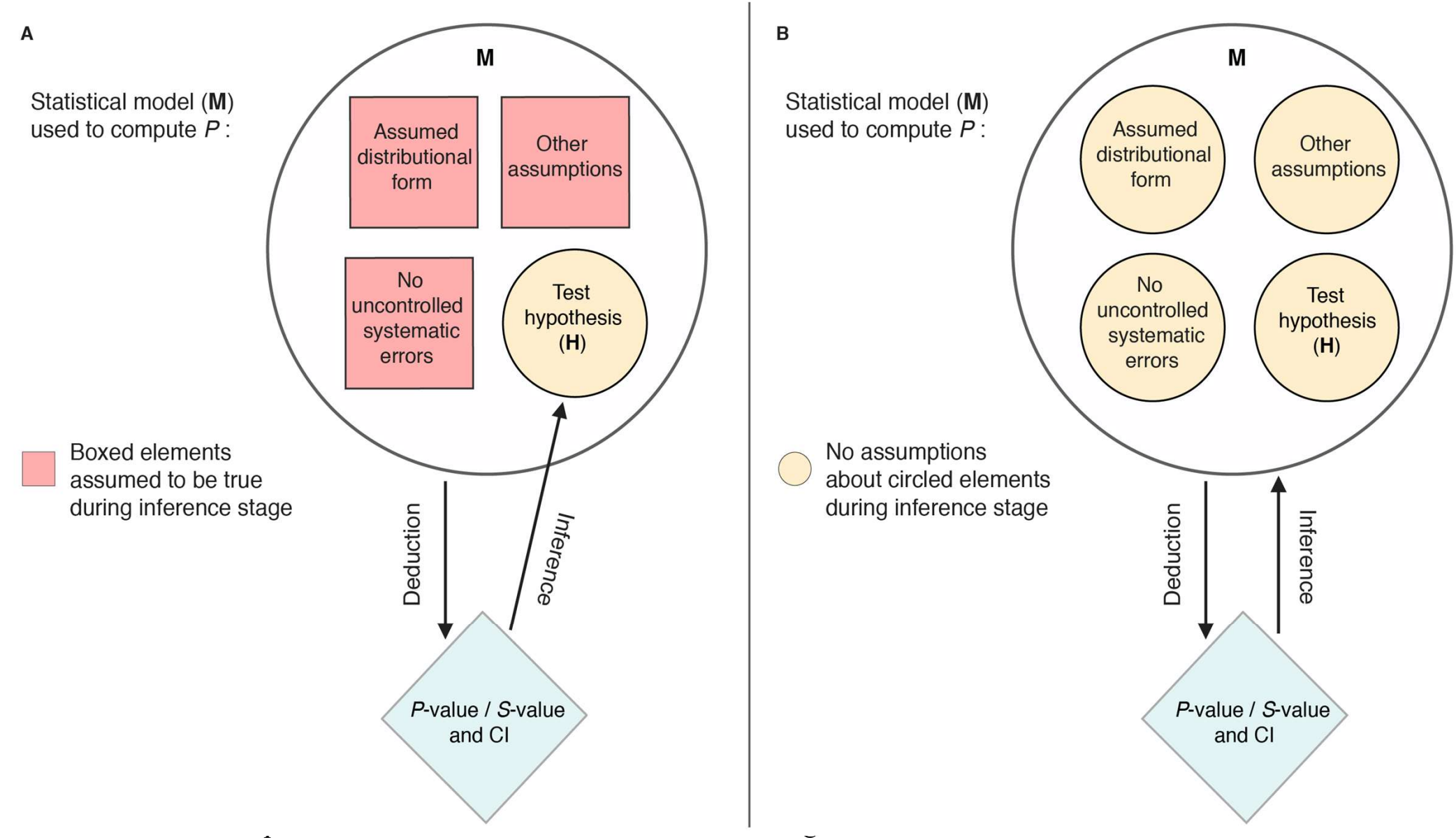
Instead, **make the assumptions explicit in all definitions and descriptions of statistics, as in**

- **The statistical model evaluated by a P-value is the hypothesis H *and* all other assumptions used to compute p.**
- **A P-value can "test" H only when those other assumptions hold. Otherwise…**
- **p is but one measure (*among many*) of how close the data are to the model predictions.**
- ***Regardless of H* being true or false, p may be small or large due to failure of other model assumptions.**

Translation to interval estimates (CI):

- **A 95% CI only displays the parameter values that have p>0.05 according to some method for computing P-values.**
- **These are parameter values that, when inserted in the statistical model, produce model predictions "close" to the data *according the method for measuring "close".***
- **Due to assumption failures and model uncertainties, a CI may be far from the "true value" and will be too narrow to capture warranted uncertainty.**

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



Some background and further readings on general methodology

(should be open access where links are given)

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

McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. *The American Statistician* 2019;73:235–245.

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;50:899-914 https://onlinelibrary.wiley.com/doi/10.1111/sjos.12645

Amrhein V, Greenland, S. Discuss practical importance of results based on interval estimates and p-value functions, not only on point estimates and null p-values. *J Inf Technol* 2022;37:316-320 https://journals.sagepub.com/doi/full/10.1177/02683962221105904

Supplement:
Statistics reform -
An assortment of slides about cognitive, psychosocial, and foundational issues from a *causal information* perspective

- **Classical methods for reducing cognitive biases include playing "devil's advocate" (arguing for disdained positions) and "alien observer" (arguing as if indifferent to the welfare of humans or other species).**
- **More detailed methods have long existed in applied psychology, but much work will be needed to adapt them to our fields.**
- **More essential than theory and abstractions are case studies – which the pandemic has supplied an abundance!**

Some lessons case studies can provide:

- **Examine and present absolute proportions and rates when discussing importance.**
- **Beliefs are based on webs of trust; no one has the capacity to validate all their sources.**
- **Those who do attempt verification will catch falsehoods, which will destroy trust.**
- **We should not dismiss criticisms or treatments simply because their proponents are biased or use faulty data, logic, or methodology to support their views.**

An alien viewpoint: Imagine conclusions, claims, and actions labeled as "science" and "following the science" from the perspective of an alien space probe with no stake in any aspect of the events such as human welfare.

Its goal could be to map the entire causal system in which these events emerge, just as biologists try to delineate systems in which species emerge and disappear, or historians or sociologists try to delineate systems in which societies form and disintegrate…

An emergentist thesis: The alien would find that all humans and their groups fall far short of comprehending the social subsystems they are in, due to their profoundly limited data acquisition and processing capacities – much like all individual animals and colonies fall far short of comprehending the ecology they are in.

In particular: **All scientists and philosophers of science fall far short of comprehending the science system they are in; their asserted demarcations of science are wishful thinking**.

- those limits apply to me, you, everyone.

In the view of some, we are apes with brains enlarged to coordinate action in large groups.

Evidence for this view: Those who study and debate cognitive biases exhibit the same cognitive errors as everyone else – even or especially when discussing these problems.

Accessible discussions of this topic can be found in blogs (perhaps more illuminating those that found in magazines and journals) …

Example: posts by Gelman highly critical of popular advocates of "behavioral economics" even if not critical of the topic of cognitive biases, describe how the advocates suffer from (apparently inescapable) biases – including those called "hubris" in ordinary language:

https://statmodeling.stat.columbia.edu/2020/03/31/stasis-back-in-town-my-last-post-on-cass-sunstein-and-richard-epstein/

https://statmodeling.stat.columbia.edu/2021/02/07/nudgelords/

My revision of Gelman's 2021 list:

1: **Cognitive biases are pervasive - People cannot reason neutrally when (as usual) they have investments in conclusions.**

2: **Uncertainty triggers emotional reactions. People get upset by uncertainty when the issue is perceived as affecting their ego, status, wealth, etc. They then deny uncertainty, or offer downwardly biased evaluations of uncertainties, often aided by warped statistical conventions.**

3: **1 and 2 together almost guarantee avoidable mistakes will be made when making decisions under uncertainty, including by experts and agencies**.

- The problems have been studied in depth at least since the 1960s, yet methods for dealing with them remain controversial.

- Francis Bacon discussed cognitive problems at length in his "Idols" in *Novum Organum* (1620), a founding work of modern science

https://sirbacon.org/the-four-idols-of-sir-francis-bacon/

Our "modern" methodologies for dealing with uncertainty – statistics, sensitivity and bias analysis – are based on idealizations that ignore these problems and use assumptions that are patently false in some of the most important health and medical research:

- That the analyst is unprejudiced, neutral, free of bias; all COIs have been "managed".
- That all important sources of uncertainty have been captured by formulas, or else can be managed intuitively in light of outputs.

This idealization meshes well with the prevalent individualist or heroic story in which

- Science is a reliable system for finding "facts" or "truths" independently of what others claim, and
- Scientists can comprehend the science system in a complete and reliable manner.

In this story, individuals achieve official recognition as a "scientist" via familiar social mechanisms: degrees, grants, publications, etc.

Contrast to: Science is a social subsystem self-identified as making "scientific inferences", offering explanations, predictions and recommendations through merging of selected (curated) data with

- incompletely explicated (partially implicit and often unrecognized) assumptions,
- incomplete and often fallacious logic, and
- evocative semantics (e.g., "significance") that confuses precise abstract theoretical entities with distantly related ordinary concepts.

Bayes vs. frequentist: Two sides of the same cognitively biased (probability-fixated) coin.

- The frequentist vs. Bayes conflict is merely a product of parties treating differently focused algorithmic (programmable) toolkits as if they were sufficient for all tasks.
- As often said, adhering to only one is crippling, like refusing to use nails or refusing to use screws in wooden framing.
- Often both are and sometimes neither are needed for analyzing data (or joining wood).

- As Box often stated, data models *are* priors: priors for data given a model.
- In information theory, **data models can also be seen as decoders of data information**.
- Common information measures involve expected log inverse (negative log) probabilities or their hessians (Fisher information).
- For such measures, **Bayesian updating becomes information summation**:

➢posterior information =
data information + prior information, where
data information = "frequentist" information

- In one view, **"Bayesian data analysis" (BDA) is an oxymoron** when "the data" and "the prior" are independent (data model and prior are both separate and prespecified) and the prior information is nonzero:
- The resulting posterior is **not** a data analysis, but is **a meta-analysis of the information in the likelihood function and the prior**.
- **The prior information can be checked by converting it into "prior data"**, e.g., data from a perfect RCT.

- In another view, only Bayes is logically coherent for inference or decision, with frequentist results a limiting case:
- The **total model** is the data model times the prior model, treated as a random-parameter model for the data
- The data model may inverted to parameter probabilities using an improper prior.
- Unfortunately, logically coherent deductions may correspond poorly with reality when the premises (models) they use are inaccurate.
- **Compatibility P-values are reality checks!**

Causation of data patterns: circumstance and design vs "random variation" ("chance")

- Concerns about "random error" dominate statistical methodology, yet capture only a small portion of concerns in real research evaluation – even in designed experiments!
- Overemphasis of "chance" explanation diverts effort away from examining physical explanations (causal hypotheses) for data patterns, focusing instead on abstract probability theory.

Some concepts of chance :

- An intrinsic, physical information measure, as in quantum and communication theory.
- An intervention that blocks all unknown causes of a variable (e.g., randomization).
- **A wastebasket term for the combined effect on our observations of all factors we failed to control (and may not even know about).** In nonrandomized medical studies, a distribution for this combined effect is a subjective prior with no physical basis.

Standard confusion and deception of statistics when applied to observational studies:

- Treat the wastebasket term as if it were an unmeasured physically random variable with a simple, identifiable parametric distribution.
- This is what is done when a model is taken as given, as in all conventional regression.
- In practice, the model is not given! It must be **caused** by known real-world mechanisms, or else our error calibration is unknown and **frequentist error claims are deceptions**.

Example: In a cohort of patients of size N, the number A of patients classified as developing a disease over a specified risk period.

➢Among requirements for A~binomial(π,N):

- Causal (physical) independence of case-defining events (outcome measurements); eg, no transmission **and** no re-evaluating patients based on findings in other patients.
- Causal independence of loss (censoring events) *across* case-defining events; eg, no enhanced follow-up of some patients based on findings in other patients.

- **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 bias portrayal of facts for propaganda and to bias inferences and policies to favor 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 as nullism) 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.**

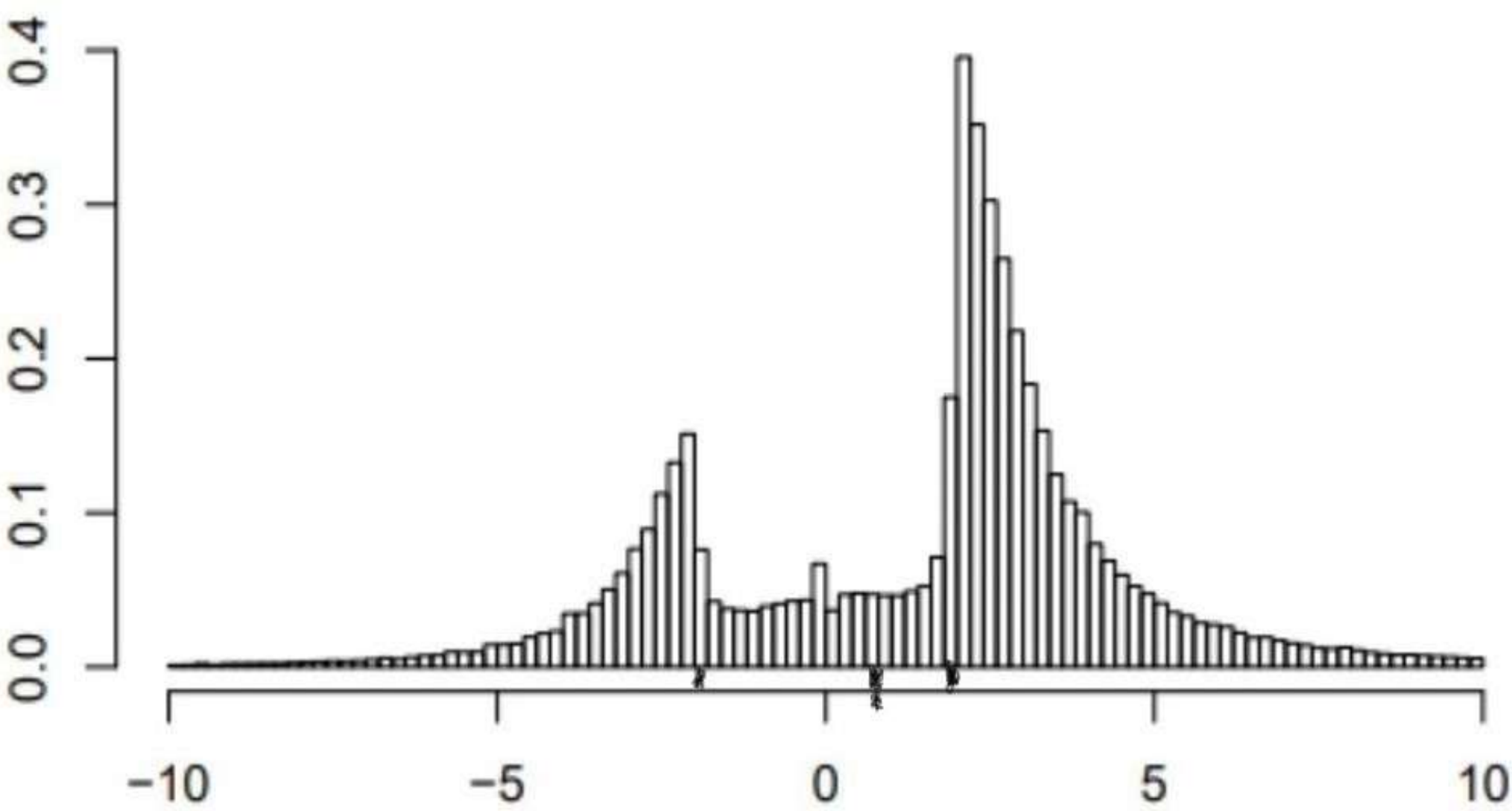
In the radical Bayesianism of DeFinetti, all probability is "subjective" – describing only properties of observer's minds. In that view

- 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".

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 leads to selective reporting based on study results will distort the distribution of available results relative to all results



Some tools for rational 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.