Toward restoring realism in statistical training and practice: Preparing students for the harsh realities of research environments in which “statistical inference” is often a device to manufacture desirable conclusions or: There are lies, damn lies, and statistics!

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

Some titles of past 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

‘There’s not much science in science’

Key observations in those talks:

- **In health and medical journals, statistical analyses omit sources of uncertainty, thus causing overconfident conclusions.**
- **Many statistics primers and study reports display overt misinterpretations of already unrealistic statistical analyses.**
- **The misinterpretations are amplified in conclusions, reviews, and press coverage.**
- **Often, *motivated reasoning* determines the direction of error and misinterpretation.**

- **Various cognitive biases are produced or enhanced by statistical training and traditions, and then used to produce or support false claims or “inferences”.**

Hence,

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

- **These biases are large, pervasive, and societally important, yet overlooked by most methodologic texts and literature.**
- **Their coverage should displace many fine points of statistical methodology, which is itself is a major source of cognitive biases.**
- **Emphasize that if we are trying to see what is going on in reality, we should develop contextually rich verbal *descriptions* of the data and the mechanisms that generated (caused) it before thinking of inference!**

- **What do statistics summarize? The data.**
- **What uncertainty does statistical theory deal with? Uncertainty about the behavior of the data-generating process.**
- **What does the researcher or client want to learn about? The target population!**
- **How is statistics traditionally taught and practiced? It confuses all 3 by starting with, focusing on, and treating as if real ideal-fantasy cases in which they all correspond "to within 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* 2023

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: When “P-value = probability of the null” gets corrected to “P-value = probability chance alone produced the association” …

They are the same: “chance alone” *is* the null!

Daniel Kahneman:

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

And most relevant to statistics in soft sciences:

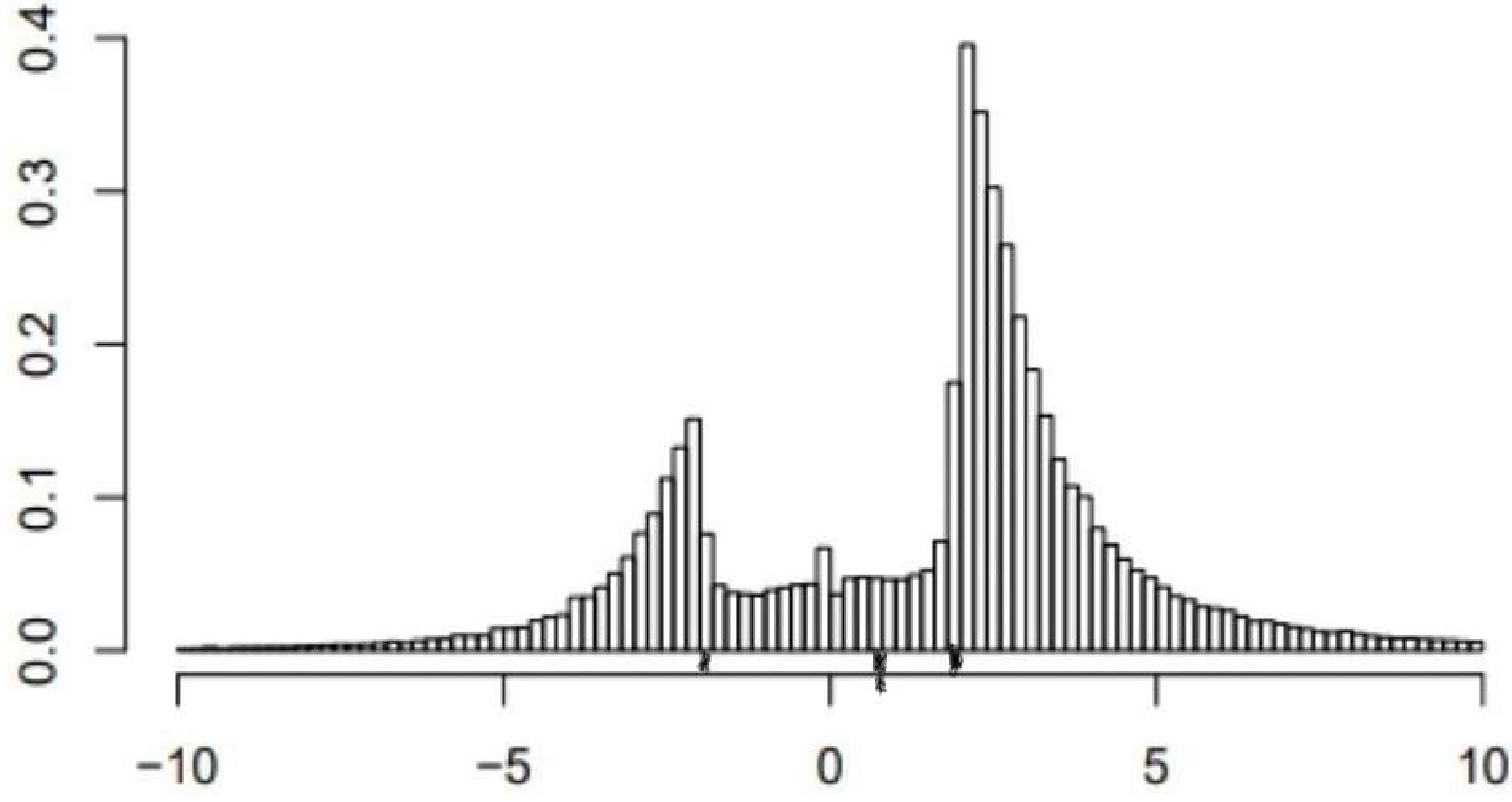
- “…**illusions of validity and skill are supported by a powerful professional culture. We know that people can maintain an unshakeable faith in any proposition, however absurd, when they are sustained by a community of like-minded believers.**”

- For examples, see most any defense of null hypothesis significance testing (NHST) as a general decision heuristic. Here is one:

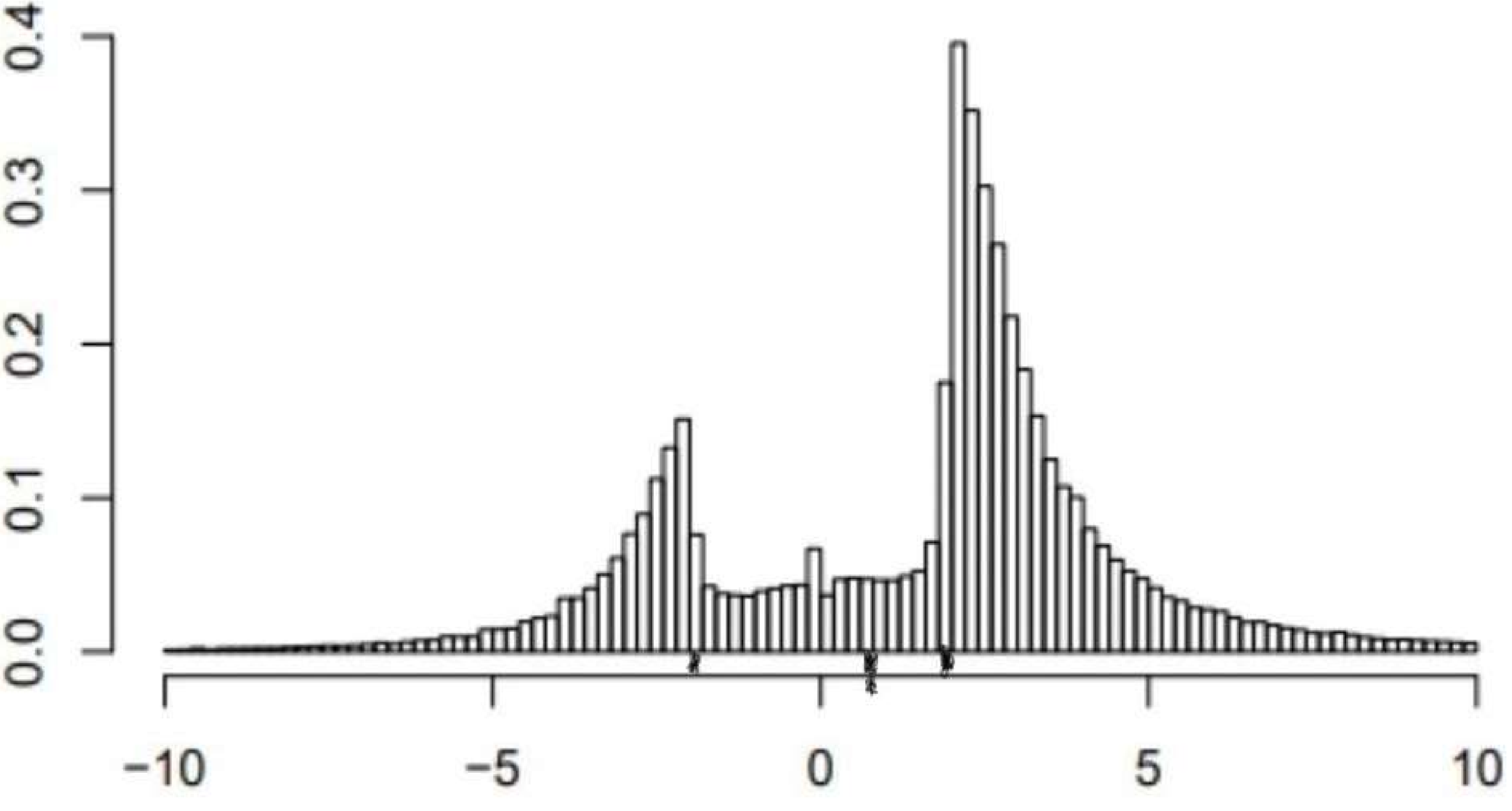
"If the p-value for the effect is greater than the journal's threshold p-value, then the editor can immediately reject the paper, which saves the journal from spending any more time on the (unconvincing) paper."

- 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



Yet more Kahneman: “**People assign much higher probability to the truth of their opinions than is warranted**.”

- **Bayesian methods open statistics to as much abuse as NHST via informative priors, and especially prior via spikes: Pr(null)=0.5 is *not* “indifference”, it is a massive null bias!**
- **Elicited priors: Summary expressions of literature biases, misunderstandings, misconceptions, and personal prejudices by overconfident and ill-informed “experts”**.

Frequentist methods use 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, and

DATA SAY NOTHING AT ALL!

Data are just markings on paper or bits on media that just sit there.

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

- **Much if not most “study conflict” arises from analysis differences, because:**
- **There are many analysis choices which must be made that are not dictated by universally accepted methods, guidelines, or rules.**
- **Example: 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.**

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. Unfortunately,**
- **Clinging to “statistical inference” leads 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.**

What then 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!**

“Statistical inference” became a distorted caricature of scientific inference

- **It has degenerated into taking output from data-processing programs** (machine-learning algorithms) **and generating “inferences” from those via rigid, decontextualized rules**.
- **It converts oversimplified models of *causes* of the data (data-generating mechanisms) into decontextualized probability functions**.
- **The semantic void it leaves causes inferential errors and enables self- and other-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"** (theorems within narrow formal systems), **without effective safeguards to prevent their harms to actual research environments and to public information**.

Statistics education should cover essential if uncomfortable features of scientific 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 and affect cognition, thus**
- **cognitive biases and social forces affect actual inferences and decisions, and**
- **every statistical analysis should be viewed as a small point in a vast sea of sensitivity.**

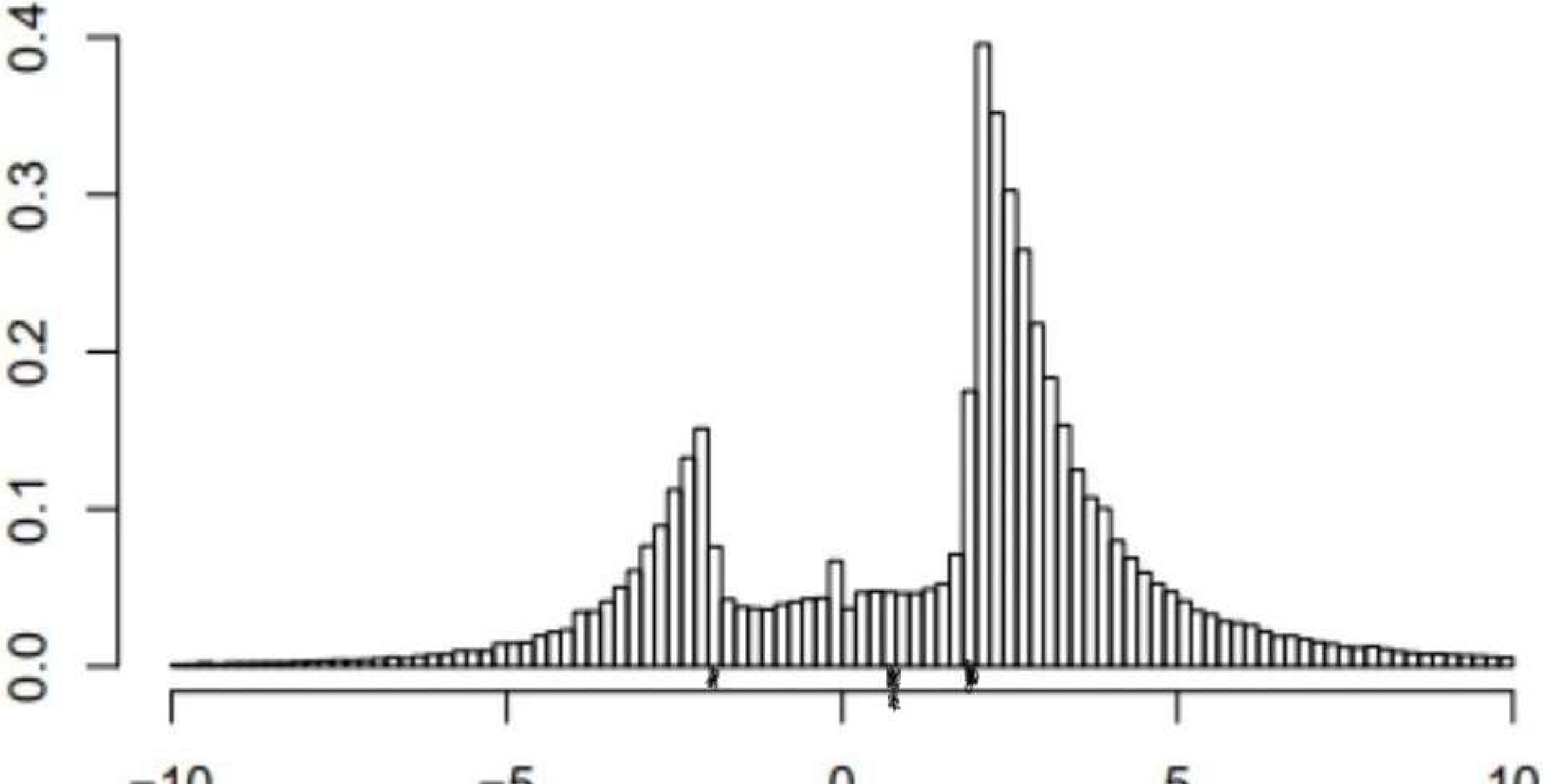
We need accurate, honest coverage of history and methods. Example: **NHST**

- 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”.**

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



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 as a final arbiter.
- **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.
- Both regarded cutoffs (α-levels) as needing contextual justification, e.g., through **decision costs (loss functions)**.

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 limitations force us to present dichotomizations such as CI, 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.**

- **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 RCTs tend to be due their expense) get p>0.05.**

Result: headline-grabbing *false* claims that

- **most results "fail to replicate"**
- **most observational studies are "refuted" by RCTs.**

Ex. **Novelty bias**: Seliger et al. EJE 2016.

- **"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:

- **Statins were inversely associated with 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 OR 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.**

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

- Abstract: “nonsignificant” OR=.60 (.28,1.30) among <12ng/ml baseline “requires further study”. **OK, but p = 0.096 for OR≥1, hence**

➢**reference posterior odds are 9:1 for OR<1**

- **Newsletters misinterpreted this as usual**, e.g. ConsumerLabs: **supplements "did not reduce risk in [those] who were vitamin D deficient (<12ng/ml) at baseline.” FALSE!**

- **Most of what I see reported as "the study found no association" is in fact misreporting of ambiguous results that lean in a direction.**
- **This could be mitigated by requiring students and research reports to tabulate or graph P-values across a range of possible parameter values**, e.g., give P-values for RR = 1, 1.5, 2, 3 or 1, 2/3, 1/2, 1/3.
- Easily implemented with Wald statistics (Z-scores) $Z(\beta) = (b-\beta)/SE(b)$, which will show

➢ point estimate b has p=1,

➢ 95% CL have p=0.05,

➢ **many β have a higher P-value than does β=0**

Scientific statistics – that is, **statistics grounded in causal thinking – requires accepting that**

- Without randomization, **“chance” does not explain anything**, and “could be due to chance” is jargon for “something unknown might have produced this association”.
- ***Investigator bias* and *social pressure* are among candidate explanations for conclusions in reports and articles**.
- **Targeted effects and uncontrolled biases operate together;** it is not “one or the other”.

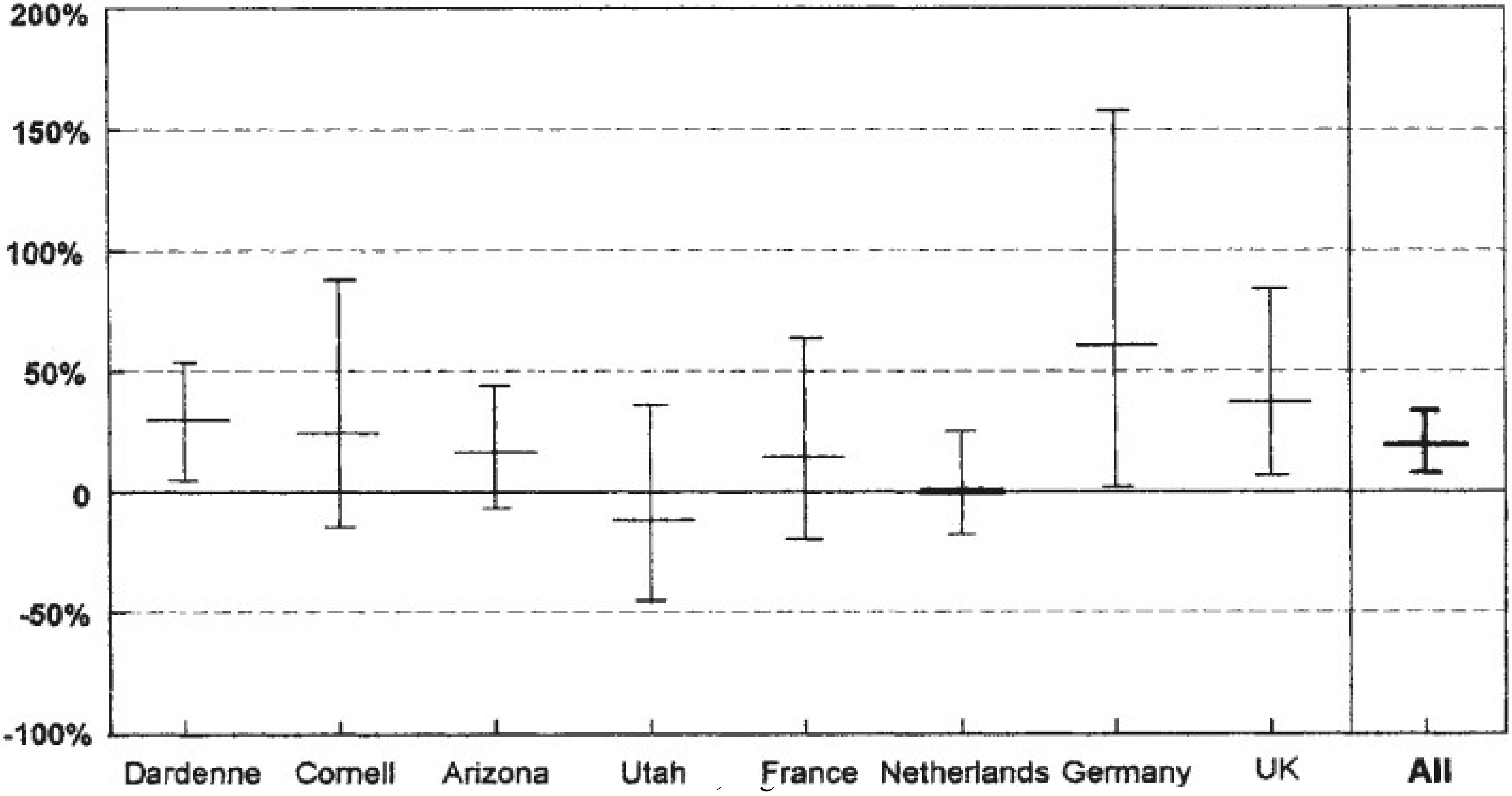
- **Reasoning motivated by legal and financial stakes feeds resistance to serious reform**

- Consider this mandatory disclaimer on U.S. dairy products labeled

> "*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**"

- there, a special-interest lobby forced a statement of fact to be accompanied by the misleading claim in red to defend rBST use…

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



Ex. **Upward P-selection (null hacking)**: 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 IPTW HDPS, the association was not significant (HR, **1.61** [95% CL **0.997, 2.59**]). **Not given**: **p = 0.0505** and more…
- Its conclusion: "**in utero exposure was not associated with autism spectrum disorder**"!!
- **Their earlier meta-analysis: 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, the association was the same, HR 1.61, p = 0.05, but all HR from 1.00 to 2.59 had p > 0.05.**
- **The association was also about the same as in our earlier meta-analysis. It may however be due to residual confounding or other uncontrolled bias** [- they did say that].

- 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 and thus encourage **overconfidence bias** when as usual we can't be certain about the physical and **human factors** that caused the (reported) statistics.
- **Bayesian methods can worsen this bias by further narrowing the intervals and by**…

Overshrinkage: As with NHST and multiple-comparisons adjustments, I have learned to distrust Bayesian analyses in the medical literature, because

- **Priors are usually pre-specified with unjustified overloading toward the null, and then misreport overshrunk posterior results as if they were empirical findings.**
- Example: Hayward et al. RCT of ivermectin and covid outcomes (J Infection 2024) reports posterior **without giving the prior**…

- **Posterior probability of benefit > .9999 >prespecified superiority threshold of .99 !**
- prespecified minimum meaningful difference = **1.2**, which they say corresponds to a ~**1.5 day** reduction in self-reported recovery time
- **HR = 1.15, 95% posterior limits 1.07,1.23**
- **Post Pr(HR≥1.2)** = **.192**, median recovery-time difference = **2.06** days (of ~14), **95% posterior limits 1.00, 3.06**
- "COVID-19-related" hospitalizations+deaths OR=**1.02 (.63,1.62)**.

- **Stated conclusion**: "Ivermectin for covid-19 **is unlikely to provide clinically meaningful improvement in recovery**…"
- "unlikely" apparently refers to the 19% post probability of HR≥1.2
- Numbers in Fig. 2 yield **HR MLE=1.19,** 95% CL = 1.12, 1.26, P-val for HR≥1.2: 0.41
- **Hence reference post Pr(HR≥1.2)** > **40%**
- **Estimated mean-zero normal prior needed to shrink 1.12,1.26 to 1.07,1.23 has 95% limits 0.89,1.13 and Pr(HR≥1.2) = 0.1%**

- **Estimated mean-ln(1.1) normal prior needed to shrink 1.12,1.26 to 1.07,1.23 has 95% limits 1.03,1.18, Pr(HR≥1.2) = 0.5%**
- **Such strong priors were not justified by previous trials; instead they reflect social pressure to discredit ivermectin.**
- **Analogous frequentist methods prevent unwanted small P-values by deploying multiple-comparisons “adjustments”, which have Bayesian justification if prior is strongly concentrated at the joint null.**

- **We thus should require Bayesian results be accompanied by reference results from the same sampling model, *without the prior*, which can be tabulations of**
- **frequentist estimates and P-values; or**
- **posterior estimates and probabilities from reference priors (e.g. Jeffreys or maxent), which in our field are close to frequentist estimates and to 1-sided P-values, respectively.**

- **Reference results reveal how much the posterior was driven by the prior rather than actual study-data information.**
- **Typical priors represent opinions whose certainty far exceeds anything derivable from actual data (such as a 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.**

- **Thus, priors need to and can be subject to diagnostics, e.g. via translation into data.**
- **"Bayesian hypothesis tests" using a prior mass of ½ for a point null are an even worse deception than 0.05-level NHST:**
- **By any sensible measure, the information in a 0.5 point mass is far beyond what can be justified by medical literature:**
- **The mass translates to a likelihood function from an infinitely large experiment!**
- **That empirical absurdity is then given massive weight in computing the posterior.**

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* of the data with the statistical model used to compute p, where that statistical 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,**
- **Confidence 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 repeated draws from a **hypothetical** data-generating algorithm, **not** to a known data generator.

In contrast, compatibility is only 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 model space.

So: 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 a criterion like P > 0.05, which represents $-\log_2(.05) \approx 4$ coin-flips (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

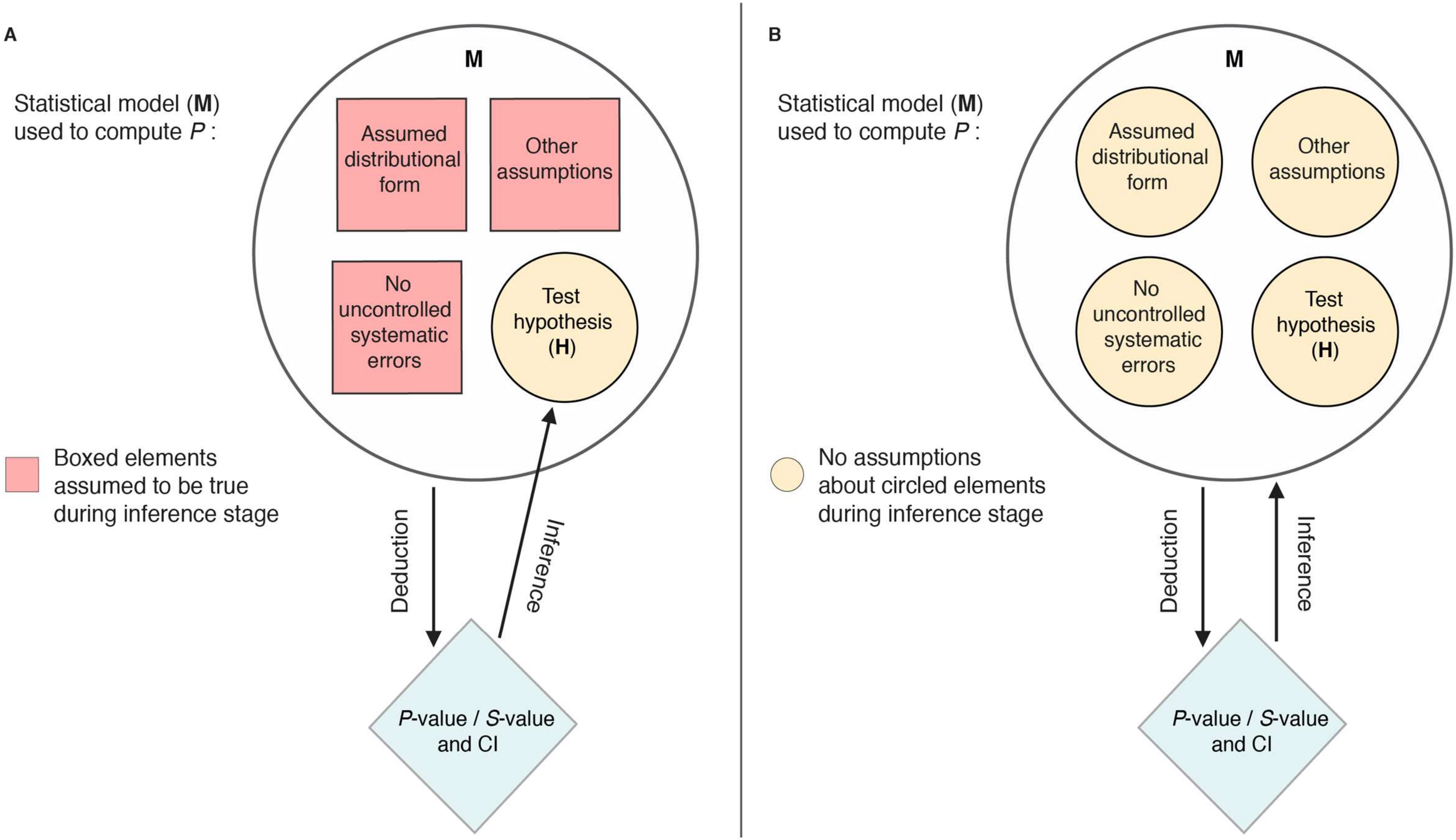
- **The norm: "The P-value is the probability of getting a test statistic as or more extreme *if* H is correct"**
- **"The CI is an interval with 95% probability of covering the true value"**.
- **Both leave the background assumptions implicit – and their uncertainty ignored.**
- **Those assumptions compose the statistical model from which p and CI are computed**.

Instead, **make the assumptions explicit in all definitions and descriptions of satistics, 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 failings of other model assumptions.**

- **In parallel, a CI only guarantees coverage of the "true parameter value" at the stated rate when the model assumptions hold. Otherwise…**
- **A 95% CI only displays the parameter values that have p>0.05 and thus, when inserted in the statistical model, produce model predictions "close" to the data *according this p-measure*.**
- **The CI may be far from the "true value" due to assumption failures.**

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


A
M
Statistical model (**M**) used to compute *P* :
Assumed distributional form
Other assumptions
No uncontrolled systematic errors
Test hypothesis (**H**)
🟥 Boxed elements assumed to be true during inference stage
Deduction
Inference
P-value / *S*-value and CI
B
M
Statistical model (**M**) used to compute *P* :
Assumed distributional form
Other assumptions
No uncontrolled systematic errors
Test hypothesis (**H**)
● No assumptions about circled elements during inference stage
Deduction
Inference
P-value / *S*-value and CI

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