Statistics as a condemned building: A structural analysis and plans for demolition and reconstruction

The Need for Cognitive Science and Causality in the Foundation and Practice of Statistics

Sander Greenland

Department of Epidemiology and
Department of Statistics, UCLA

Please report errors and send comments to Sander Greenland at lesdomes@ucla.edu

"The reason social science calls itself a 'science' is because of statistics. And their statistics are practically BS everywhere. I mean, really, everywhere." – N.N. Taleb @nntaleb 1:58pm 9Feb2019

- **This is also true of much of "health and medical sciences" – that should scare you!**
- **What if a major source of the problem is pundits in statistics and "meta-research" neglecting their own cognitive deficiencies and those of developers, instructors, users, and consumers of statistics?**

Why some call "statistical science" an **oxymoron** (self-contradiction):

- In the U.S. at least, statistical training largely degenerated into statistical mathematics and computing, to become a primordial, ritualized form of machine learning.
- It treated context, meaning, and values as if those were mere abstract algorithmic inputs (e.g., "prior distributions," "loss functions").
- It denigrated semantics and clear ordinary language, favoring instead deceptive jargon.

This degeneration of statistical science into a mathematical shell left behind explication of and training in these essential components of **scientific inference**:

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

Empirical fact: **We are all stupid**

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**."

"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 of the worlds of scientific research and statistics.

"**The confidence people have in their beliefs is not a measure of the quality of evidence** but of the coherence of the story the mind has managed to construct." – Daniel Kahneman

- Few pushing reform have tested their ideas by comparing practice impacts. **As "confidence" intervals** (CI) **illustrate**, unintended adverse effects can be severe (just like with medicines).
- **Bayesian methods open statistics to even more abuse via prior spikes and "elicited priors" (summary expressions of biases, literature misreadings, and personal prejudices**).

More Kahneman: "**People assign much higher probability to the truth of their opinions than is warranted**." (see: **Bayesian statistics**)

"We can be blind to the obvious, and we are also blind to our blindness." (see: **CI examples below**)

And most relevant to statistics in the 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.**"

- Example: **Claiming Pr(null)=0.5 is "indifference"**

Deficiency number 1 to address:

The need for cognitive science in statistics

to address human psychology and its biases, e.g.,

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

Nullism has a long and glorious history among physics idolaters as **pseudo-skepticism** (empirically indefensible certainty about nulls):

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

- **Against Nullism**: Reality is under no obligation to behave simply for you.
- **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**?
- **Hidden Reification**: Researchers routinely publish "inferences" that ignore vast model uncertainties (they don't know a rationale for neglecting all those missing effects in models - they just don't think about them).

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

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

<u>All</u> of the following and more should form part of basic training for moderating inferences:

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

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

Those problems are among the major reasons that 'most published research findings are false':

- Like everyone, **stat instructors**, users, and consumers suffer from **dichotomania** and **nullism**: They crave true-or-false conclusions for null hypotheses (misapplying the excluded middle).
- **One study can never provide absolute certainty**, even if it **is** the basis of a decision.
- Yet statisticians have invented sophisticated **decision theories** which make it **appear** to users that single studies can supply definitive answers.

"Confidence intervals" perpetuate these biases…

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

- Abstract: "[Cox] 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**])."
- Abstract and article conclusions: "…**exposure was not associated with autism spectrum disorder**..." **despite reporting the same increased risk in earlier studies, citing a meta-analysis of 4 cohorts with HR 1.7 [1.1, 2.6]**

- **Ugly fact: The main problems of P-values will extend to any statistic, because they are caused by truth-subverting (perverse) incentives and cognitive biases, not P-values.**
- **Perverse incentives create cognitive biases (wishful thinking, positive projection) to see what the incentives dictate. These biases pervade reports in fields like medicine.**
- **Incentives are often to report ambiguous results as null results, as when researchers want to explain away unwanted associations -- a form of null bias.**

Value bias pervades received statistical methodology, most often in the form of **nullism**

Call a methodology **value-biased** when it incorporates assumptions about error costs that are not universally accepted (and are usually hidden).

- Example: **The consistent use of the null as the test hypothesis**, to the point of failing to distinguish the null and test hypothesis. This is an example of **nullism**, value bias toward the null.
- May stem from **imaginary universal costs** of rejecting the null (as in product surveillance), or from metaphysical beliefs (parsimony, ideology).

- **This bias afflicts a good portion of the Bayesian literature**, where **null spikes** are used to represent a belief that a parameter "differs negligibly" from the null.
- In most medical-research settings, concentration of prior probability around the null has no basis in genuine evidence. In fact **prior spikes usually contradict genuine prior information. For example, medicines are pursued *precisely* because they affect physiological systems**.

- Still, **many scientists and statisticians exhibit quite a bit of prejudice in favor of the null based on faith in oversimplified biological models**.
- Nullism also arises from **confusion of decision rules with evidence summaries**, and from **adoption of simplicity or parsimony as a metaphysical principle instead of a heuristic**
- We might be confident that any effect present is small enough so that the cost of ignoring it is acceptable **- but that's a value judgment!**

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

Why most published inferences remain false even with precise P-values

- Instructors and users want P-values to be the probability of a point hypothesis (usually, a null hypothesis of no association or no effect).
- A P-value is rarely near that probability.
- **Yet the literature encourages subtle fallacious descriptions that are equivalent to treating a P-values *as if* they were hypothesis probabilities ("P-inversion").**

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

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

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

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

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

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

- **All these problems underscore the need for coverage of reasoning errors and cognitive biases as an essential component of any specialty claiming to promote sound scientific inference from data.**
- Instead, statistics primers indulge in the **ludic fallacy** of treating all uncertainty as if it arises from games of chance – random draws from a distribution of **known** form – instead of addressing our deep uncertainty about **the form and sources of variation and bias**.

What is INFERENCE?

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

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

My perception ← Reality → Your perception

Contrast scientific inference to

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

The challenge: **Statistics (like medicine) is a technology that has become a major source of harms as well as benefits**

- Successes have distracted the field from failures.
- **Mathematics has distracted attention from hard real-world problems**, diverting research and teaching into math that solves nothing real if human cognitive problems are not addressed.
- Example: **Competence and integrity are widely compromised, yet are core assumptions of almost all statistical methods.**

Time to face the ubiquity of error at all levels

Error (including error from bias) is inevitable, not only in data and inference but also **conceptual missteps extending to the highest authorities**.

- A key to minimizing conceptual error is to **vary perspectives** by considering conceptually different approaches, and by considering a lengthy list of **cognitive biases in claims.**
- A key to minimizing average error *cost* is uncertainty assessment, to encourage well-balanced hedging: full analysis of **alternatives** to 'accepted' hypotheses or 'null' hypotheses.

Reconstruction – A simple start:
STOP perpetuating the mistakes of "great men" of statistics and the cognitive bias they reflect, create, and encourage

- Statistics education has assumed users understand mathematics well enough to see through terminology to the correct math (general) meaning. *Perhaps* true in Fisher's heyday, it became utterly false in the research explosion after WWII, when the pool of researchers exploded to fill demand.

- **Typical users now depend on words because the mathematics is for them simply symbolic incantations they must take on faith to get funded and published.**
- **"That's just semantics"**: Irresponsibly fails to grasp the essential analogical information conveyed by the semantics. That failure is common among the mathematically able, who place syntax and deduction above analogical processes, or even dismiss or miss entirely the role of analogy in mapping reality to math.

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

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

*"Consistency" is nearly equivalent but is used for too many other concepts.

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 "highly" or "moderately" compatible with the data under some test criteria like P > 0.04, 0.01 (= 96, 99% coverage given background assumptions).
- This involves no computation or numeric change! It's all about **perception**…

"Compatible" is unambitious and far more cautious than "confidence":

- There is always an infinitude of models compatible with our data – **and most are unimagined and even unimaginable given current knowledge** (recall Jeffreys' error).
- "Confidence" implies belief, encouraging the inversion fallacy that infiltrates discourse about the models themselves. Whereas
- **Compatibility is <u>no</u> basis for confidence…**

Compatible, effective – and false:

- **Causal stories are what map hypotheses to data. Compatible but false stories *may* lead to effective interventions**.
- Example: "Malaria is caused by bad air that collects near the ground around swamps." Implied, effective solutions: raise dwellings, drain swamps - the compatible cause (bad air) and actual cause (mosquitos) are both reduced by those interventions. But **confidence in the story will eventually mislead**!

Stop repeating Fisher's error of using "null hypothesis" for any test hypothesis (which openly invites nullistic bias)

"Null" in English Dictionaries:

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

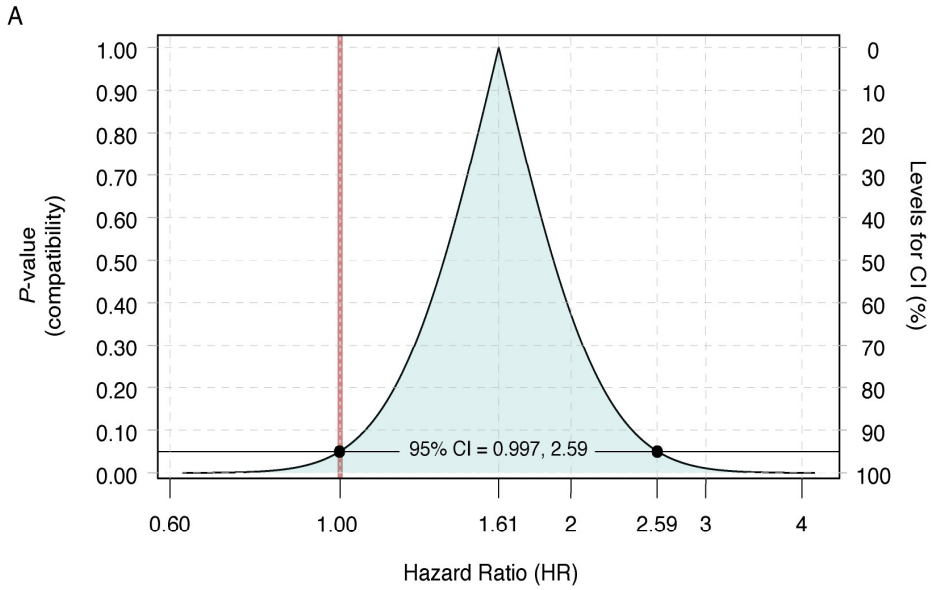
Stop repeating the massive error of NOT treating P-values as estimation tools

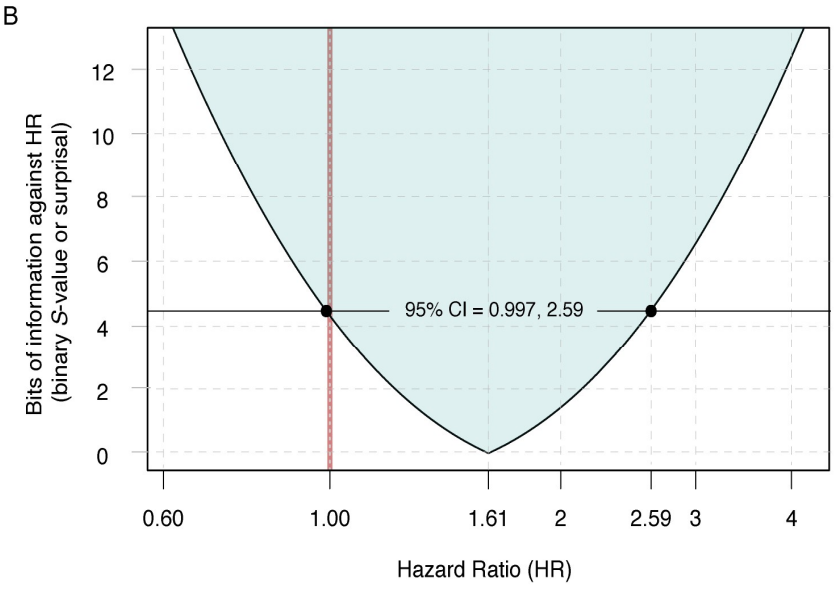
(which also openly invites nullistic bias)

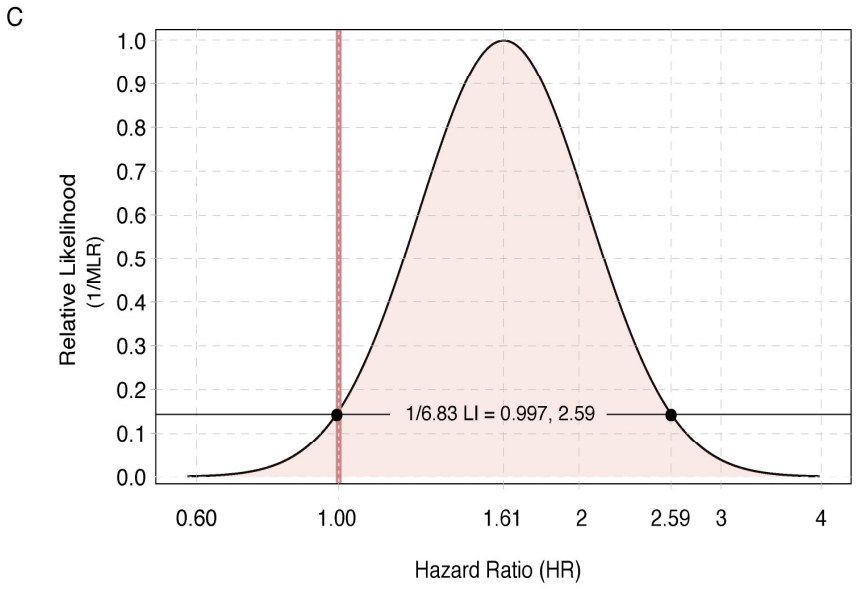
"…the distinction between significance testing and estimation is artificial…indeed of negative value if it leads to needless duplication of effort in the belief that one is solving two different problems" – Edwin Jaynes, informationalist

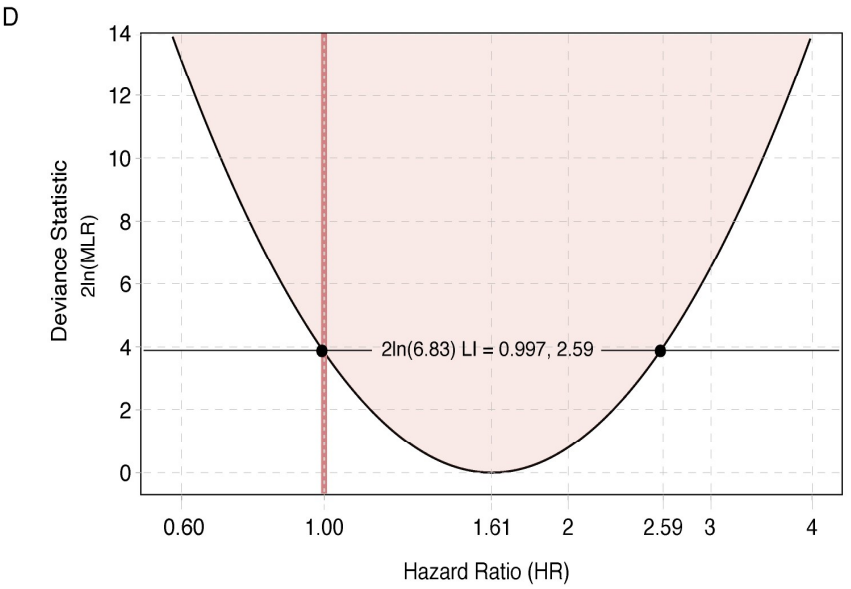
- Indeed, the distinction has been **entirely destructive** in encouraging tests and decisions to focus on just one point or model in an entire spectrum of hypotheses and models:

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





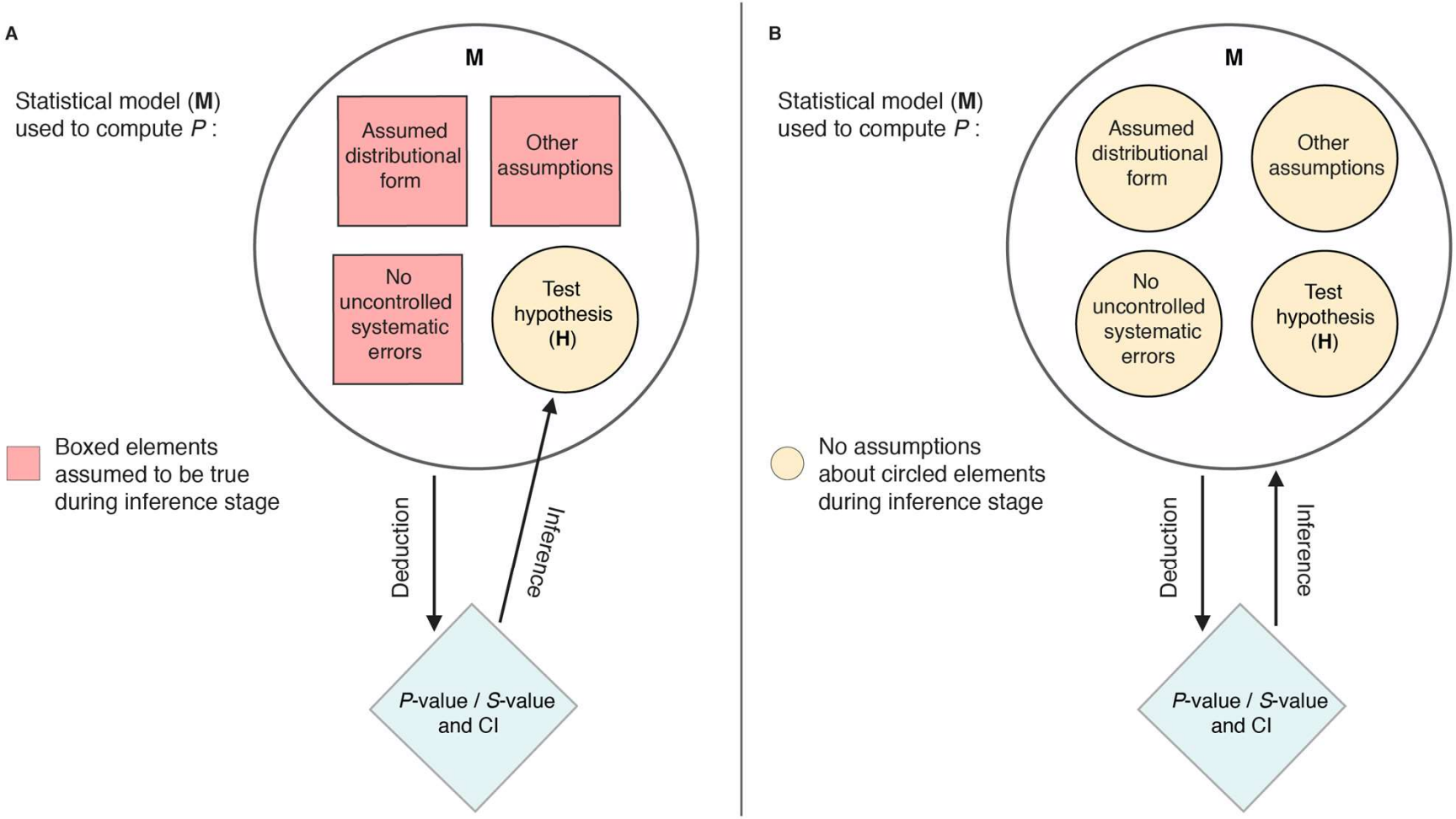




Shift emphasis away from conditional "hypothesis-testing" interpretations to unconditional descriptive interpretations

- The norm "The P-value is the probability of a statistic as or more extreme under the **tested hypothesis**" leaves the background assumptions implicit. **Use instead**
- "The P-value is the **percentile under the tested model** at which the statistic falls". That model includes the test hypothesis **and all other assumptions used to compute P!**

from Greenland & Chow http://arxiv.org/abs/1909.08583



STOP using the distortive inverse-exponential scale of P-values for gauging evidence

- Switch to the Shannon information **against** the model supplied by the test: the S-value (surprisal)

$$\mathbf{S = \log(1/P) = -\log(P)}$$

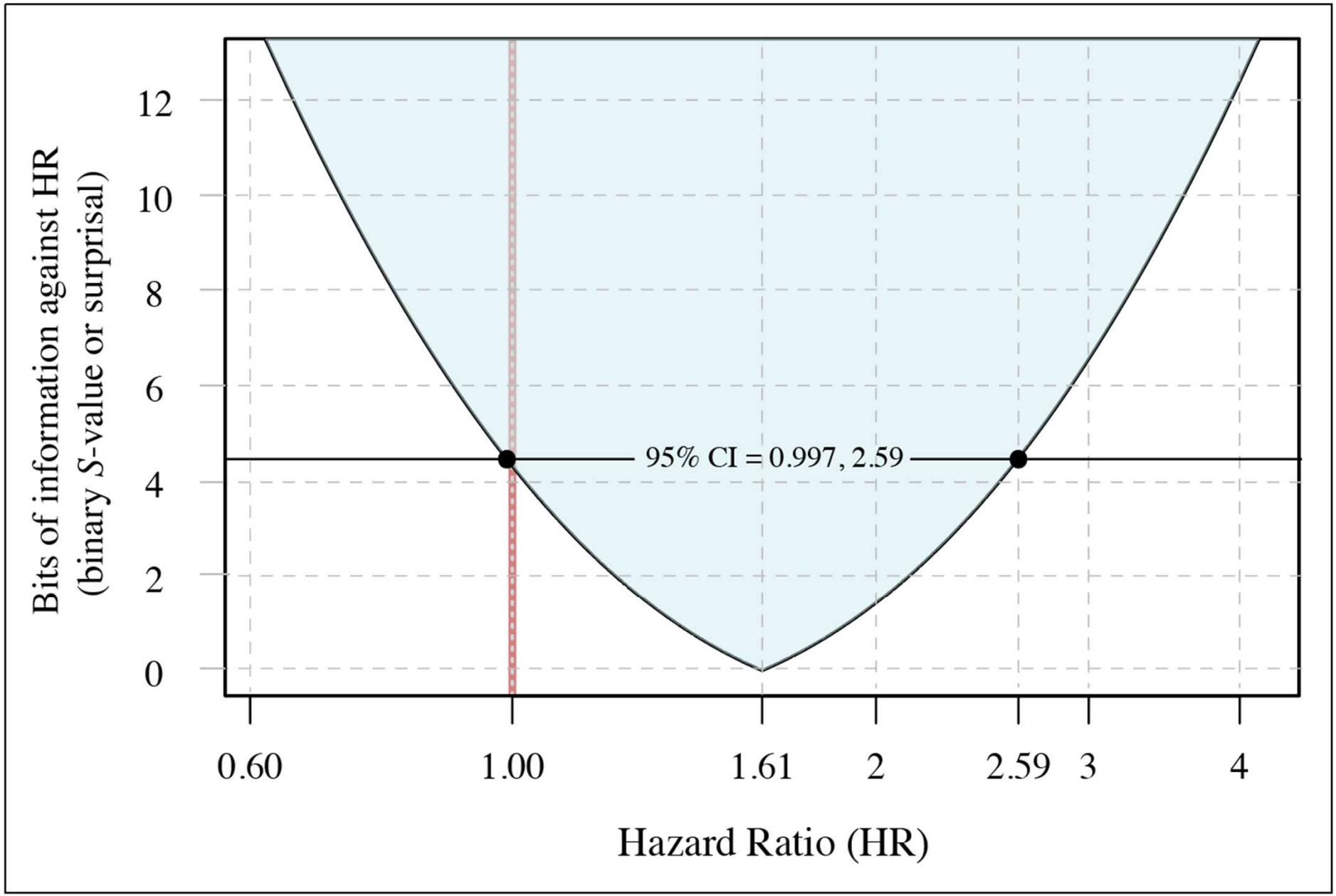
- S is a measure of data **in**compatibility with the model.
- Unlike P, S is an equal-interval scale that is additive over independent test statistics (as Fisher 1930 used for his meta-analytic test).

- This idea goes back at least to Good (1956) and has resurfaced repeatedly since, whenever theorists needed to gauge the evidence or information in a P-value.
- S is hard to confuse with a Bayesian probability because it ranges far above 1.
- S does not require a prior distribution, but can use a prior by computing P as a test of fit of a compound sampling model that treats the prior as a parameter distribution in the model (a "random-effects" model).

- With base-2 logs, S is the Shannon information against the tested model, measured in bits…
- $-\log_2(.05) = 4.3$ bits, which is small in this sense: S approximates the number of binary observations needed to provide that much information against the tested model.
- −log2(.005) = 7.6 bits
- Would you say you had definitive evidence from 5 binary observations? From 8? **There is no correct answer out of context!**

- The overall idea is to stop and ask: What exactly is the entire set of assumptions (model) **checked** by the test**.**
- When all test assumptions are explicit, you then ask: How much information does the test convey **against** the entire model?
- This analysis is for "discovery" (e.g., **refutation of absence** of model violations).
- P and S do **not** provide confirmation (e.g., **evidence of absence** of violations) – that requires checking **alternative** models.

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



In closing:

- **Blind acceptance of mathematical frameworks, deification of "great men" and their conceptual errors, and neglect of cognitive problems have rotted the core of statistical training and research practice.**
- **The "replication crisis" hysteria continues the problem via its nullism and dichotomania.**
- **We must rebuild statistics as an information science, not a branch of probability theory, with cognitive science as a core component.**