The Need for Cognitive Science 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

Does the title sound radical? Consider:

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

- **This is also true of much of "medical science"**

  - and that should scare you!

What if the major source of the problem is statistics (and its "philosophers") neglecting its own deficiencies and those of its developers, users, and consumers?

Said of physics c. 1940!:

"The public drinks in and swallows eagerly everything that tends to dispossess intelligence in favor of some technique; it can hardly wait to abdicate from intelligence and reason...[They] ask nothing better, it would seem, than to leave their destiny, their life, and all their thoughts in the hands of a few men with a gift for the exclusive manipulation of this or that technique."

- Simone Weil (1909-1943), "Wave Mechanics" in *On Science, Necessity, and the Love of God* (transl. by R. Rees 1968, p.75)

- if only this thought were applied to statistics!

Of Newtonian time and space, Roveli (*The Order of Time*, 2018) said "Don't take your intuitions and ideas to be 'natural': they are often the products of the ideas of audacious thinkers who came before us."

- Intuitions are shaped mostly and often entirely by what you **learned**. Thus
- they are products of information (often incorrect) you absorbed (often incorrectly) interacting with hard-wired preferences ("natural instincts"), and so
- they are subject to error and **bias**, both yours and the founders, books, and teachers of statistics. It matters not that the founders were "brilliant"…

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 academic 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**, as with medicines, unintended adverse effects can be severe.
- **Bayesian methods open statistics to even more abuse via prior spikes and "elicited priors" (summary expressions of biases, misreadings of literature, 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.**"

Greenland: "There's not much science in science."

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

- Successes are used to distract attention from failures.
- **Mathematics is used to distract attention from hard real-world methodologic 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 taken for granted by most reporting and are core assumptions of almost all statistics today (outside of forensic research).

The boundary between incompetence and malfeasance is blurry, because both are

- caused by **Perverse Incentives** (**PI**, incentives that violate the official scientific goal of truth-finding but serve the professional advancement and enrichment of researchers and clients) and its interactions with **Wish Bias** (thinking and seeing what one desires).
- manifested in "questionable research practices" (QRPs, a polite description of **systematic errors**).

An honor system does not work for science any more than for politics or safety. Yet law enforcement can worsen practices (e.g., requiring $p<0.05$ to report).

Can we re-establish statistics as a science of extracting information and meaning from observations? Yes, but a century of problems shows **probability is an insufficient foundation**.

- Most "statistical analysis" in open-ended research (as opposed to industrial error control) has been about applying **often-inappropriate probability rituals to data, based on incorrect or garbled understanding of the terms and outputs**.
- Meanwhile, **statistical theory has degenerated into an extension of probability theory**, with data processing taken over by information science.

Logical foundations for statistical analysis and interpretation based on explicit **and implicit** goals

- "Scientific inference" contains elements labeled "statistical," which deal with "data analysis."
- These elements need to be integrated into reasoning laden with **contextual meaning**, not treated as abstract math or computing.
- When this is done, it can be seen that even **objective-sounding goals like "accurate prediction" are in fact driven by valuations (loss functions) which may conflict across stakeholders – even within academic "science"!**

- "Statistical inference" is currently limited to data processing and **information mining** based on data summaries and assumed mechanisms generating the summaries (the **data-generating model, DGM**). Thus formal statistics is an **information** science (Efron 2005) for processing data sequences with information-extracting algorithms (some labeled "frequentist" and others "Bayesian").
- **This "information" is logical reduction of the data based on constraints in the DGM, some unjustified and some derived from causal stories about how the data were generated.**

- **Those stories are what map the reality to data and the statistics back to reality.**
- Thus, those stories must be "true enough" to achieve goals of describing reality as accurately as possible given the available data.
- **Yet stories may be wrong and still lead to effective interventions**. Example: Malaria is caused by bad air that collects near ground level around swampy areas. Implied and effective solutions: raise dwellings, drain swamps - hypothesized cause (bad air) and actual cause (mosquitos) are both reduced by the interventions.

Subjective elements play a decisive role in all statistical analyses

- **There is an illusory sense of objectivity induced when there is great overconfidence**, as generated by elaborate theory (math, biologic, etc.), strong sense of authority, and extensive social agreement.
- **Feelings of objectivity in turn feed back to create more overconfidence**. This is amply illustrated in history by scientists and even entire fields assigning near certainty to hypotheses later refuted (e.g., Fisher on smoking and lung cancer, Jeffreys on continental drift).

Health and medical examples abound: Bland diet for ulcers; low-fat diets for weight loss; numerous drugs that were aggressively promoted and then discredited (with errors sometimes encouraged by deceptive trial analysis and reporting, e.g.,Vioxx).

- There are many parallels in modern statistical practice. Among them, **the objective-frequentist hegemony produced an epidemic of discrete significance testing that in turn led to entrenched reporting distortions (publication bias, rampant misrepresentation of ambiguous results as null)**.

'**Objectivity' in statistics usually means nothing more than the following act of faith**:

- The data were derived from a study conducted in a manner that physically *forced* the assumed **data generating model** (DGM) to hold.

This meta-assumption is usually derived from independence assumptions which follow when interventions (selection or treatments) are applied to population units following a **known and perfect randomized-design protocol** (albeit perhaps depending on covariates in complex ways).

In the real world of soft sciences, however such faith (absolute certainty) in researchers and publications is unwarranted, for reasons such as

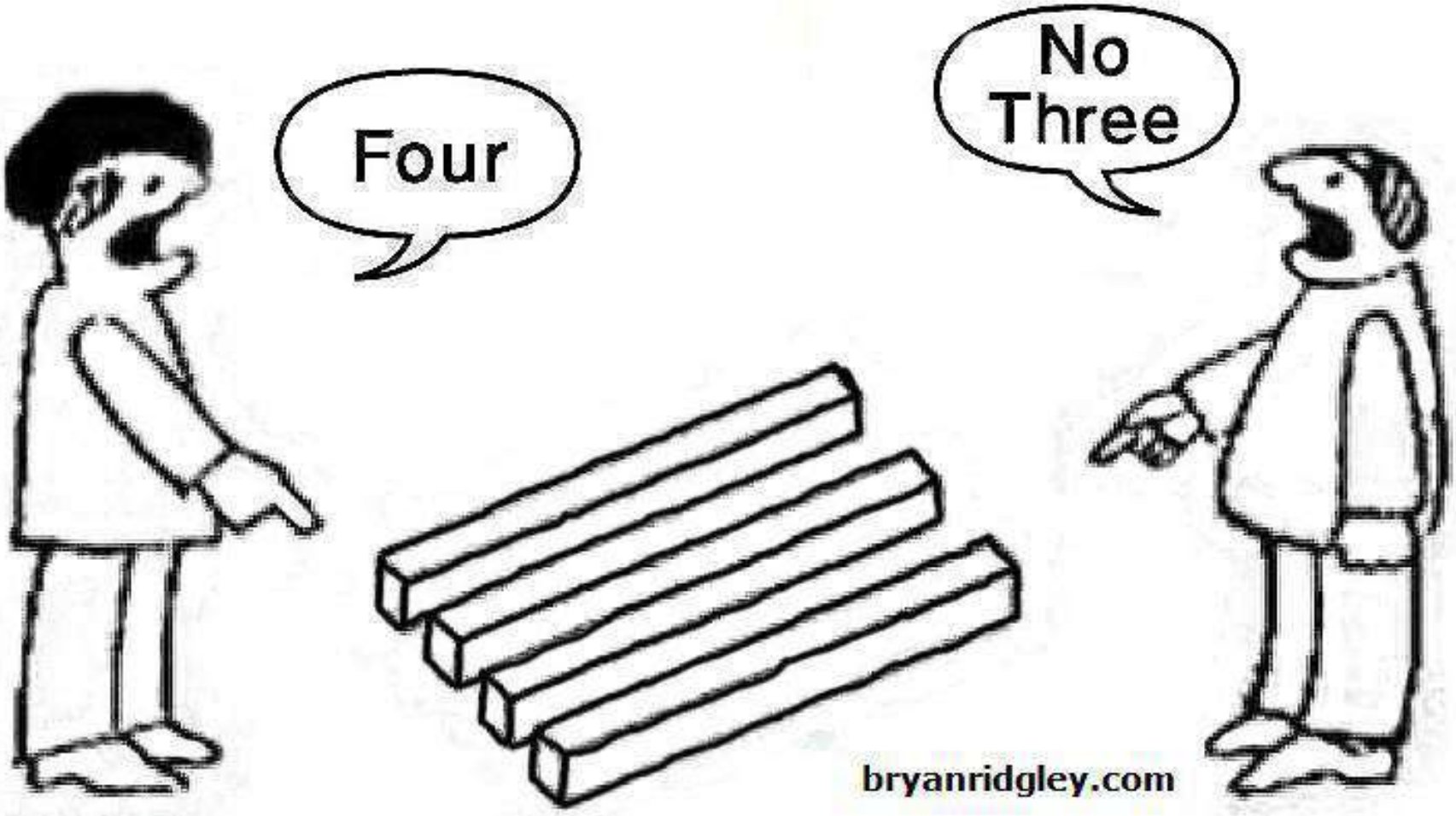
- Cheating, fraud (see WSJ ’09 on 21 faked studies)
- **Procedural error and bias -** often undocumented, but hard to deal with even if recognized.
- **Pressure to reach conclusions from data which would appear ambiguous** if interpreted soundly.
- Pressure to reach conclusions to satisfy prior prejudices (**Wish Bias**) or **sociopolitical agendas**. Example: Highly selective citation (as in Young and Karr in *Significance* Dec. 2011).

The ubiquity of error at all levels

Error (including systematic error) is inevitable – not only data and inference error but also **conceptual error extending to the highest authorities**.

- 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.
- A key to minimizing conceptual error is to **vary perspectives** by applying conceptually different approaches to assessments, and by considering a lengthy list of **cognitive biases**.

Reality can be so complex that equally valid observations from differing perspectives can appear to be contradictory.



Statistical training and practice for valid open-ended information generation needs to cover at least:

1. **Cognitive science**, to recognize, understand, and control **human bias sources (HBS)**. **Current formal training ignores and suffers from HBS**.
2. **Logic and graphical models**, to display validity threats **including human bias sources**. Only slowly being introduced, and then only for special cases of methodologic (design) biases, not HBS.
3. Design strategies to **block** biases. Current training focuses on experimental designs (where randomization and blinding block certain HBS).

4. Thorough yet compact study and data description (**information extraction and summarization**). Current training focuses most attention on programming this step, as it is by far the most mathematical, with probability models in the central role as information-extraction devices.
5. **Valid interpretation of the information**. Currently either very informal and thus hard to get right without extensive trial-and-error experience (including error correction, which academics are largely immune to), or else bound by rigid, primitive "philosophies" and religious practices.

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

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

These 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 that make it *appear* to users that definitive answers are provided by single studies.

Confidence intervals perpetuate these biases…

Example (Eur J Epid 2016;31:947-51):

- Abstract: "**use of statins was not associated with risk of glioma (OR for ≥90 prescriptions=0.75; 95% CI 0.48-1.17). Our findings do not support previous sparse evidence of a possible inverse association between statin use and glioma risk.**" [prev. studies: 0.72 (0.52-1.00); 0.76 (0.59-0.98)]
- 1st sentence of discussion: "**This matched case–control study revealed a null association between statin use and risk of glioma.**"

- so, simply banning "significance" does not stop the **nullistic fallacy** and statistical misinterpretations.

A more 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**])."

[2017 M-A, 4 cohorts, same 1st author: **1.7, 1.1-2.6**]

- Abstract and article conclusions: "…**exposure was not associated with autism spectrum disorder**..." despite observing practically the same increased risk as in earlier studies!

Yet massive cognitive problems remain even if dichotomania and nullism are cured, and all probabilities are reported to supportable numeric precision ("continuously"):

- Sound *inferential* (as opposed to mathematical) interpretations of statistics eludes most people, **including most statisticians and stat educators**.

This is a problem in **science education and practice, *not* math**. [We should want to get our math right, but should realize that answers derived with some math errors can be closer to reality than competing answers with perfect math.]

18 different misinterpretations of P-values (along with several related misinterpretations of confidence intervals and power) are catalogued in **Statistical tests, confidence intervals, and power: A guide to misinterpretations**, Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.C., Poole, C., Goodman, S.N., Altman, D.G. (2016). *The American Statistician*, 70, free at http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/utas_a_1154108_sm5368.pdf

- **Over half of the errors are inversion fallacies**, mistakenly equating

Pr(hypothesis|evidence) to Pr(evidence|hypothesis)

A major reason why most published research giving precise P-values remains false

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

Presenting P-values for multiple alternatives (P-value functions) would help combat this mistake.

Inversion mistakes include misinterpreting a *P*-value as the probability 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?) - This is typical, not exceptional!

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

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

Abstract of an article listing Kooperberg, Lumley, Psaty (Am J Epidemiol, 2007) among the authors says they "conducted a permutation test to estimate the **probability of a chance finding**":

- If the null model is correct and a 'chance finding' is a false positive, what is "**the probability of a chance finding**"? Answer: 5% when using a test with a Type-I error rate (size) of 5%.
- If the null model is *in*correct and a 'chance finding' is a false negative, what is "**the probability of a chance finding**"? Answer: the Type-II error rate β (1–power), which is usually **much** higher than 5%

Unfortunately for all such misinterpretations, if chance **alone** produced a difference, then the difference was *not* produced by any of

- Real effects
- **Biases** from uncontrolled error sources, such as errors in assumptions.

Thus, the hypothesis that **chance *alone*** produced a difference or "finding" is **logically identical** to the hypothesis that **the set of assumptions (model) used to compute the *P*-value is correct**. Hence

Pr(chance alone|evidence) = Pr(model|evidence)

- This is a posterior probability, not a *P*-value!

So, unfortunately for these misinterpretations,

- A *P*-value is the probability that the chosen **test statistic** would be **at least** as large as observed ***if* the model used to compute it were correct**.
- **A *P*-value is thus a probability of evidence given the model**, Pr(evidence|model)

where the "evidence" is an inequality about the test statistic used to compute *P*. Thus, **"chance alone" is a hypothesis**, ***not* a study or data feature**, and so

- **Equating a *P*-value to the probability of "chance alone" is another example of an inversion fallacy**!

- **Ugly fact: The main problems of P-values will extend to any statistic, because they are problems of truth-subverting ("perverse") incentives and cognitive biases, not of *P*-values.**
- **Perverse incentives create cognitive biases (wishful thinking, positive projection) to see what the incentives will reward. These biases pervade reports in fields like medicine.**
- **Incentives are often to report nulls even if those are false negatives, as when researchers want to explain away unwanted associations (nullism).**

Returning to an example…

Brown et al.:

- 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**])."
- Article and news conclusions: "**exposure was not associated with autism spectrum disorder**"

[2017 M-A of 4 cohorts, same 1st auth: 1.7, 1.1-2.6]

[uncited 2016 M-A of 16 cohorts: 1.74, 1.19-2.54]

Why no discussion of the consistent association of 60% higher risk among the exposed? Well,

the authors are convinced it's all confounding…

- They invoke **causal** arguments and indeterminate evidence to support the confounding explanation, but fail to note that **uncontrolled bias is itself a hypothesis** requiring its own critical assessment.
- That implies they should acknowledge the uncertainty left by the evidence and reach no conclusion from the data they present.
- Why then do they present unwarranted conclusions *at odds with their own data*?

Possible cause: **Investigator bias** (as should be expected if the investigators or their colleagues have been prescribing the treatment under scrutiny).

- The Brown et al. example appears to involve **analysis hacking** to get p>0.05, in this case by adjusting until the CI finally includes 1 (even though the adjustments beyond the initial model appear to be overadjustment, inflating variance without removing bias).

Note: This is opposite the current cognitive social meta-bias which talks as if all incentives are to report positives and thus hack to get p<0.05.

- **This meta-bias is rampant in the "replication crisis" literature**, which uncritically ignores the difference in incentives across topics and authors.

It is important to question conclusions about matters for which there is far from sufficient evidence for any certainty.

- To carry out this advice, we have to come up with defensible measures of uncertainty, and methodology for constructing those uncertainty measures from statements (including data).
- That turns out to be an incredibly difficult task – **perhaps too difficult for routine use by independent research groups**. But we have to see what reliable uncertainty assessment involves to appreciate how much we fall short in practice.

What is INFERENCE?

- Dictionary example: "**A conclusion reached on the basis of evidence and reasoning**."
- "**Statistical inference**," in any current formalism, "school" or toolkit, **is nothing more than decision output from a program** (learning algorithm) **for generating conclusions via deductive logic from quantities treated as known** (data; "missing data" is an oxymoron). **Distinct from…**
- "**Scientific inference**," **a complex but narrowly moderated judgment about reality** (based on the assumption that an objective reality exists).

Many cognitive biases affect inferences

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

- **Cognitive biases affect and produce design, analysis, reporting, and publication biases**.

All 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, etc.
- **Confirmation bias** –selective focus on desirable evidence (Brown et al.)
- **Courtesy bias** –tendency to be obscure about criticisms that will cause offense

- **Failure to test alternatives** (congruence bias)
- **Selective criticism** of undesired evidence
- **Selective reasoning** to desired conclusions via selection of assumptions, explanations, and data.
- **Dunning–Kruger effect** – 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 outputs of methods or judgments are as accurate about the world as they are in the thought experiments used to derive them.

Other problems rotting the core of statistical training and research practice:

- **Reification** – Treating outputs of assumption sets as if those models were correct, ignoring assumption uncertainty (**overconfidence** in formalisms and deductive conclusions).
- **Dichotomania** – Reducing quantities (variables, P-values, hypothesis) to dichotomies without providing explicit reasoning for doing so or for the boundary chosen.
- **Nullism** – privileging the null with no explicit reasoning (from a loss function) for doing so.

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

Value bias afflicts all decision-theoretic inference, most often as **nullism** (subtalk: 13 pages)

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 be based on **imagined** costs of rejecting the null (as in product surveillance), or more often, metaphysical beliefs (parsimony, ideology).

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

- "**Heavier than air flying machines are impossible**" – Lord Kelvin, 1895 (repeated 1902)
- "**Continental drift is out of the question**" because no [known] mechanism is strong enough – Sir Harold Jeffreys, geophysicist (and originator of spiked priors)
- "**Physics shows that cell phones cannot cause cancer**" because microwaves are not ionizing – Michael Shermer, *Scientific American* Oct. 2010

- In soft sciences **there is rarely any positive scientific evidence that the null is *exactly* true**, and few specialties have credible mechanistic arguments for claiming departures from the null are probably negligible.
- D.R. Cox (2001) opined that in many studies "there may be no reason for expecting the effect to be null. The issue tends more to be whether **the direction of an effect has been reasonably firmly established** and whether **the magnitude of any effect is such as to make it of [contextual] importance**."

- **This view directly indicts a good portion of the Bayesian literature**, where *null spikes* are used to represent the belief that a parameter "differs negligibly" from the null.
- In psychology, many (e.g., Cohen) have argued that null hypotheses are almost never exactly true.
- 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: medicines are pursued precisely because they interact with systems in disease processes**.

- Still, **many scientists and statisticians exhibit quite a bit of prejudice in favor of the null based on faith in oversimplified physical models of biology**. Shermer is a vivid example (cancer is caused by far more than just ionizing radiation).
- Nullism also arises from **confusion of decision rules with logical inference**, and from **adoption of simplicity or parsimony as a metaphysical principle rather than as a heuristic**.
- We might be highly certain that any effect present is small enough so that the cost of ignoring it is acceptable **- but this is a value-laden judgment**.

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

That is, H may be the non-null hypothesis!

Neyman continues: “From the point of view of the manufacturer [of a chemical A] the error in asserting the carcinogenicity of A is (or may be) more important to avoid than the error in asserting that A is harmless. Thus, **for the manufacturers of A, the 'hypothesis tested' may well be: 'A is *not* carcinogenic'. On the other hand, for the prospective user of chemical A the hypothesis tested will be unambiguously: 'A *is* carcinogenic'**. In fact, this user is likely to hope that the probability of error in rejecting this hypothesis be reduced to a very small value!”

In Neyman's example, nullism is bias *against* the consumer, *for* the manufacturer.

Multiplicity adjustments worsen nullistic bias:

- They take the **joint** null as the hypothesis more important to not reject incorrectly using the maximum tolerable Type-I error rate of 0.05 for **the entire ensemble of nulls**.
- These tests assume false-positive costs are always more that false-negative costs and their cost ratios always increase with the number of hypotheses.
- This valuation applies to drug companies monitoring adverse effects, **but not to patients**.