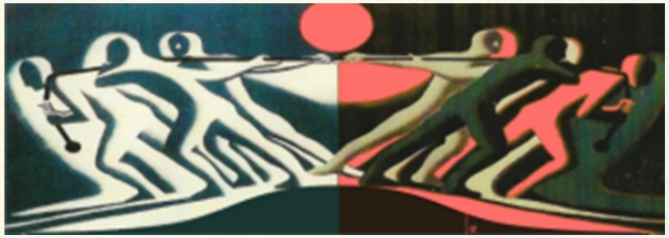
# The Statistics Wars and Their Casualties



**Deborah G Mayo**
Dept of Philosophy, Virginia Tech
Research Associate, LSE
September 22, 2022

**Workshop: The Statistics Wars and Their Casualties**
LSE (CPNSS)

# Begin with a question: From what perspective should we view the statistics wars?

- The standpoint of the ordinary, skeptical consumer of statistics
- Minimal requirement for evidence

# Requirement of the skeptical statistical consumer

- We have evidence for a claim *C* only to the extent *C* has been subjected to and passes a test that would probably have found it flawed or specifiably false, just if it is.

- This probability is the stringency or severity with which it has passed the test.

# Applies to any methods now in use

- Whether for testing, estimation, prediction—or solving a problem (formal or informal)

# 2. Statistical Significance test battles & their ironies

- Often fingered as the culprit of the replication crisis


- It's too easy to get small P-values—critics say
- Replication crisis: It's too hard to get small P-values when others try to replicate with stricter controls

- R.A. Fisher: it's easy to lie with statistics by selective reporting, ("political principle that anything can be proved by statistics" (1955, 75))

- Sufficient finagling—cherry-picking, data-dredging, multiple testing, optional stopping—may result in a claim *C* appearing supported, even if it's unwarranted—**biasing selection effects**

# Error Statistics

This underwrites the key aim of statistical significance tests:

- To bound the probabilities of erroneous interpretations of data: *error probabilities*

A small part of a general methodology which I call *error statistics*

(statistical tests, confidence intervals, resampling, randomization)

# Fraud-busting and non-replication based on P-values

- Remember when fraud-busters (Uri Simonsohn and colleagues) used statistical significance tests to expose fraud?
  - data too good to be true, or
  - inexplicable under sampling variation (Smeesters, Sanna)

"Fabricated Data Detected by Statistics Alone" Simonsohn 2013

- How is it that tools relied on to show fraud, QRPs, lack of replication are said to be tools we can't trust?

("P-values can't be trusted unless they are used to show P-values can't be trusted")

- Simmons et al. recommend "a 21 word solution" to state stopping rules, hypotheses, etc. in advance (Simmons et al., 2012)
  - Yet some reforms are at odds with this

# 3. Simple significance (Fisherian) tests

"…to test the conformity of the particular data under analysis with $H_0$ in some respect…." (Mayo and Cox 2006, p. 81)

…the **P-value:** the probability the test would yield an even larger (or more extreme) value of a test statistic T assuming chance variability or noise

NOT Pr(data|$H_0$ )

# Testing reasoning

- Small P-values *indicate* some* underlying discrepancy from $H_0$ because **very probably (1- P) you would have seen a less impressive** difference were $H_0$ true.

- This still isn't evidence of a genuine statistical effect $H_1$ yet alone a scientific conclusion $H^*$—only abuses of tests (NHST?) commit these howlers

*(until an audit is conducted testing assumptions, I use "indicate")

# Neyman and Pearson tests (1933) put Fisherian tests on firmer ground:

Introduces alternative hypotheses $H_0$, $H_1$

$H_0$: $\mu \leq 0$ vs. $H_1$: $\mu > 0$

- Trade-off between Type I errors and Type II errors
- Restricts the inference to the statistical alternative—no jumps to $H^*$ (within a model)

Tests of Statistical Hypotheses, statistical decision-making

# Fisher-Neyman (pathological) battles

- The success of N-P optimal error control led to a new paradigm in statistics, overshadows Fisher
- “being in the same building at University College London brought them too close to one another”! (Cox 2006, 195)

# Contemporary casualties of Fisher-Neyman (N-P) battles

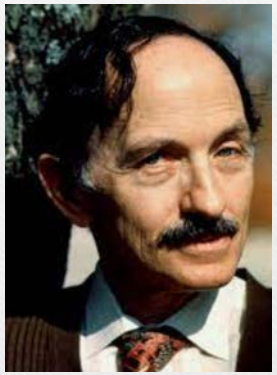
- N-P & Fisher tests claimed to be an “inconsistent hybrid” where:
- Fisherians can’t use power; N-P testers can’t report P-values (P =) but only fixed error probabilities (P <)
  - In fact, Fisher & N-P recommended both pre-data error probabilities and post-data P-value

# What really happened concerns Fisher's "fiducial probability"

The fallacy in familiar terms: Fisher claimed the confidence level measures both error control and post-data probability on statistical hypotheses without prior probabilities—in special cases

But it leads to inconsistencies

“[S]o many people assumed for so long that the [fiducial] argument was correct. They lacked the daring to question it.” (Good 1971, p. 138).



- Neyman did, develops confidence intervals (performance rationale)

# Do we need to know the history to get beyond the statistics wars?

No, we shouldn't be hamstrung by battles from 70, 80 or 90 years ago, or to what some of today's discussants think they were about.

"It's the methods, stupid" (Mayo 2018, 164)

- Key question remains (from the fiducial battle): how to have a post data quantification of epistemic warrant (but not a posterior probability)?

- Severity? Calibration?

# Sir David Cox's statistical philosophy

- We need to **calibrate** methods: how would they behave in (actual or hypothetical) repeated sampling? (performance)
  - *Weak repeated sampling*: "any proposed method of analysis that in repeated application would mostly give misleading answers is fatally flawed" (Cox 2006, 198)

# Good performance not sufficient for an inference measure (post data)

Cox's "weighing machine" example in 1958

How can we ensure the calibration is relevant, (taking account of *how the data were obtained)* without leading to the unique case, precluding error probabilities? (Cox 2006, 198)

"Objectivity and Conditionality" (Cox and Mayo 2010)

# 4. Rivals to error statistical accounts condition on the unique case

All the evidence is via *likelihood ratios (LR)* of hypotheses

$$\Pr(\boldsymbol{x}_0;H_1)/\Pr(\boldsymbol{x}_0;H_0)$$

The data $\boldsymbol{x}_0$ are fixed, while the hypotheses vary

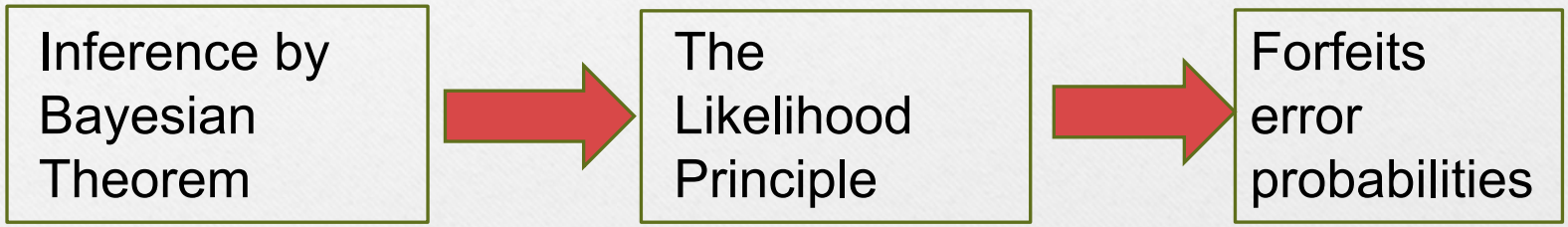
- Any hypothesis that perfectly fits the data is maximally likely

# All error probabilities violate the LP

- “Sampling distributions, significance levels, power, all depend on something more [than the likelihood function]–something that is irrelevant in Bayesian inference–namely the sample space.” (Lindley 1971, 436)

Inference by Bayes Theorem

The Likelihood Principle


Inference by Bayesian Theorem
The Likelihood Principle
Forfeits error probabilities

# Many "reforms" offered as alternative to significance tests, follow the LP

- "Bayes factors can be used in the complete absence of a sampling plan…" (Bayarri et al. 2016, 100)

- "It seems very strange that a frequentist could not analyze a given set of data…if the stopping rule is not given….***Data should be able to speak for itself***". (Berger and Wolpert 1988, 78)

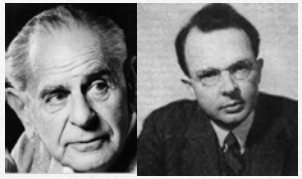
(Stopping Rule Principle)

**Table 1.1** The effect of repeated significance tests (the "try and try again" method)

| Number of trials $n$ | Probability of rejecting $H_0$ with a result nominally significant at the 0.05 level at or before $n$ trials, given $H_0$ is true |
|---|---|
| 1 | 0.05 |
| 2 | 0.083 |
| 10 | 0.193 |
| 20 | 0.238 |
| 30 | 0.280 |
| 40 | 0.303 |
| 50 | 0.320 |
| 60 | 0.334 |
| 80 | 0.357 |
| 100 | 0.375 |
| 200 | 0.425 |
| 500 | 0.487 |
| 750 | 0.512 |
| 1000 | 0.531 |
| Infinity | 1.000 |

In testing the mean of a standard normal distribution

# The LP parallels the holy grail of logics of induction *C*(*h*,*e*)



I was brought up on *C*(*h*,*e*), but it doesn't work.

Popper (a falsificationist): "we shall simply deceive ourselves if we think we can interpret *C*(*h*,*e*) as degree of corroboration, or anything like it." (Popper 1959, 418).

He never fleshed out severity

# Fisher, Neyman, Pearson were allergic to the idea of a single rule for ideally rational inference

- Their philosophy of statistics was pragmatic: to control human biases.
  (design, planning, RCTs, predesignated power)

- But there's a link to formal statistics: the biases directly alter the method's error probabilities

- Not automatic, requires background knowledge

# 5. Bayesians: we can block inferences based on biasing selection effects with prior beliefs

(without error probabilities)

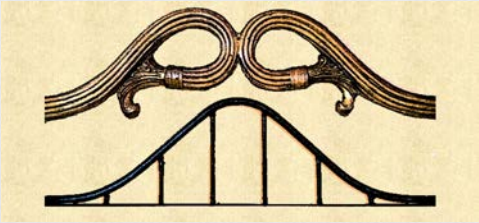
# Casualties

- Doesn’t show what researchers had done wrong—battle of beliefs
- The believability of data-dredged hypotheses is what makes them so seductive
- Additional source of flexibility, priors as well as biasing selection effects

# Peace Treaty (J. Berger 2003, 2006): "default" ("objective") priors

- Elicitation problems: "[V]irtually never would different experts give prior distributions that even overlapped" (J. Berger 2006, 392)

- Default priors are to prevent prior beliefs from influencing the posteriors–data dominant



# Casualties

- “The priors are not to be considered expressions of uncertainty, …may not even be probabilities…” (Cox and Mayo 2010, 299)
- No agreement on rival systems* for default/non-subjective priors
- The reconciliation leads to violations of the LP, forfeiting Bayesian coherence while not fully error statistical (casualty for Bayesians?)

*Invariance, maximum entropy, frequentist matching

# 6. A key battle in the statistics wars (old and new): P-values vs posteriors

- P-value can be small, but Pr($H_0|\boldsymbol{x}$) not small, or even large.
- To a Bayesian this shows P-values exaggerate evidence against.

- “[T]he reason that Bayesians can regard P-values as overstating the evidence against the null is simply a reflection of the fact that Bayesians can disagree *sharply* with each other“ (Stephen Senn 2002, 2442)

# Some regard this as a Bayesian family feud ("spike and smear")

- Whether to test a point null hypothesis, a lump of prior probability on $H_0$

$X_i \sim N(\mu, \sigma^2)$

$H_0: \mu = 0$ vs. $H_1: \mu \neq 0$.

- Depending on how you spike and how you smear, an α significant result can even correspond to

$\Pr(H_0|\boldsymbol{x}) = (1 - \alpha)!$ (e.g., 0.95)

- A deeper casualty is assuming there ought to be agreement between quantities measuring different things

# 7. Battles between officials, agencies, journal editors—and their (unintended) consequences

# ASA (President's) Task Force on Statistical Significance and Replicability (2019-2021)

The Task Force (1 page) states:

"P-values and significance testing, properly applied and interpreted, are important tools that should not be abandoned."

"Much of the controversy surrounding statistical significance can be dispelled through a better appreciation of uncertainty, variability, multiplicity, and replicability". (Benjamini et al. 2021)

## The ASA President's Task Force:

**Linda Young,** National Agric Stats, U of Florida (Co-Chair)
**Xuming He,** University of Michigan (Co-Chair)
**Yoav Benjamini,** Tel Aviv University
**Dick De Veaux,** Williams College (ASA Vice President)
**Bradley Efron,** Stanford University
**Scott Evans,** George Washington U (ASA Pubs Rep)
**Mark Glickman**, Harvard University (ASA Section Rep)
**Barry Graubard,** National Cancer Institute
**Xiao-Li Meng,** Harvard University
**Vijay Nair,** Wells Fargo and University of Michigan
**Nancy Reid,** University of Toronto
**Stephen Stigler,** The University of Chicago
**Stephen Vardeman,** Iowa State University
**Chris Wikle,** University of Missouri

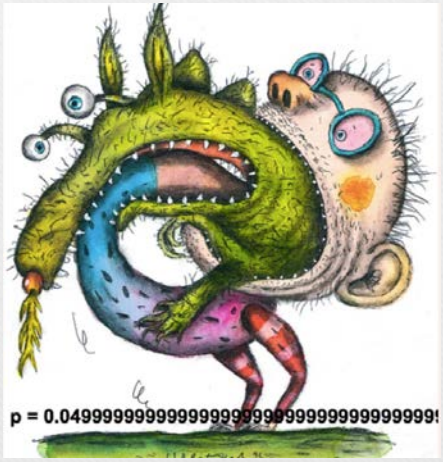
# The task force was created to stem casualties of an ASA Director's editorial (2019)*

- "declarations of 'statistical significance' be abandoned" (Wasserstein, Schirm & Lazar 2019)
- You may use P-values, but don't assess them by preset thresholds (e.g., .05, .01,.005): **No significance/no threshold view**

*2022 disclaimer

# Some (unintended) casualties

- Appearance that statistics is withdrawing tools for a major task to which scientists look to statistics: to distinguish genuine effects from noise.
- And even that this is ASA policy, which it's not



Most serious casualty: Researchers lost little time:

“Given the recent discussions to abandon significance testing it may be useful to move away from controlling type I error entirely in trial designs.” (Ryan et al. 2020, radiation oncology)

Useful for whom?

# Not for our skeptical consumer of statistics

- To evaluate a researcher's claim of benefits of a radiation treatment, she wants to know: How many chances did they give themselves to find benefit even if spurious (data dredging, optional stopping)

- Not enough that their informative prior favors the intervention—"trust us, we're Bayesians"

# No tests, no falsification

- If you cannot say about any results, ahead of time, they will not be allowed to count in favor of a claim *C* — if you deny any threshold — then you do not have a test of *C*

- Most would balk at methods with error probabilities over 50% — violating Cox's weak repeated sampling principle

- N-P had an undecidable region

# Some say: We do not worry about Type I error control: All null hypotheses are false?

1. The claim “We know all nulls are false” boils down to all models are strictly idealizations—but it does not follow you know all effects are real
2. Not just Type I errors go, all error probabilities, Type II, magnitude, sign depend on the sampling distribution

# Reformulate tests

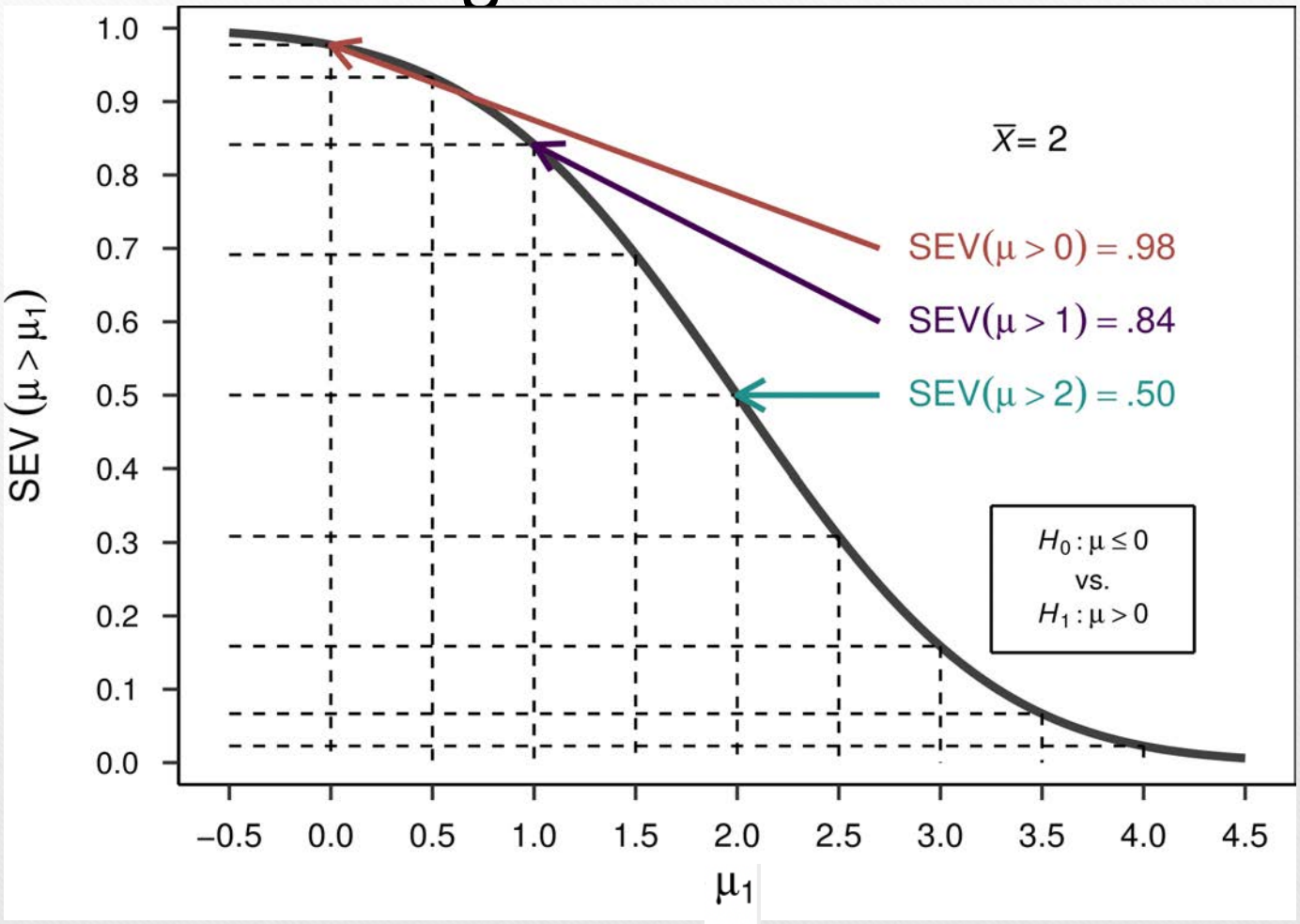
- I've long argued against misuses of significance tests

I introduce a reformulation of tests in terms of discrepancies (effect sizes) that are and are not severely-tested SEV(Test T, data *x*, claim *C*)

- In a nutshell: one tests several discrepancies from a test hypothesis and infers those well or poorly warranted

Mayo1991-2018; Mayo and Spanos (2006); Mayo and Cox (2006); Mayo and Hand (2022)

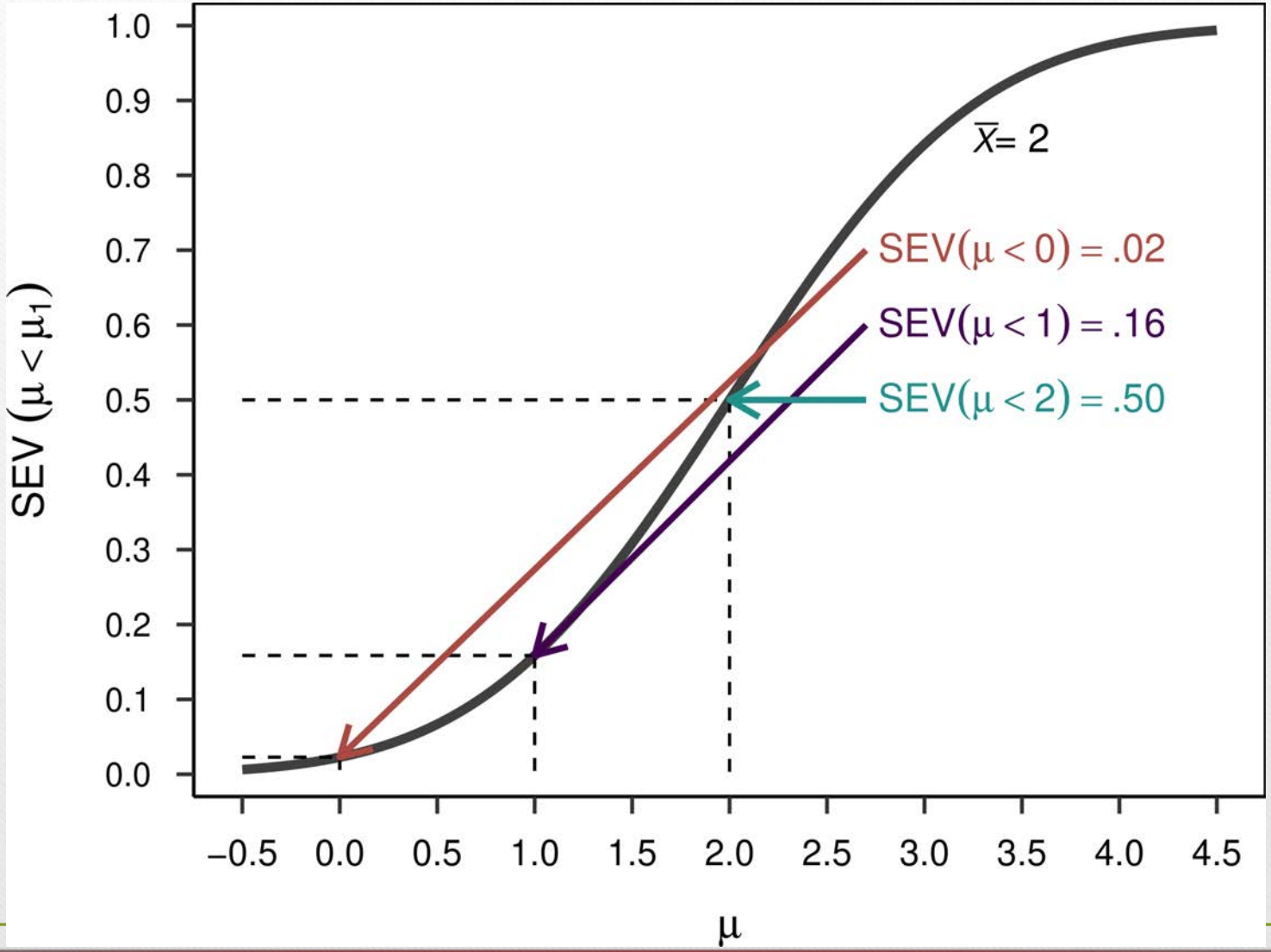
# Avoid misinterpreting a 2SE significant result



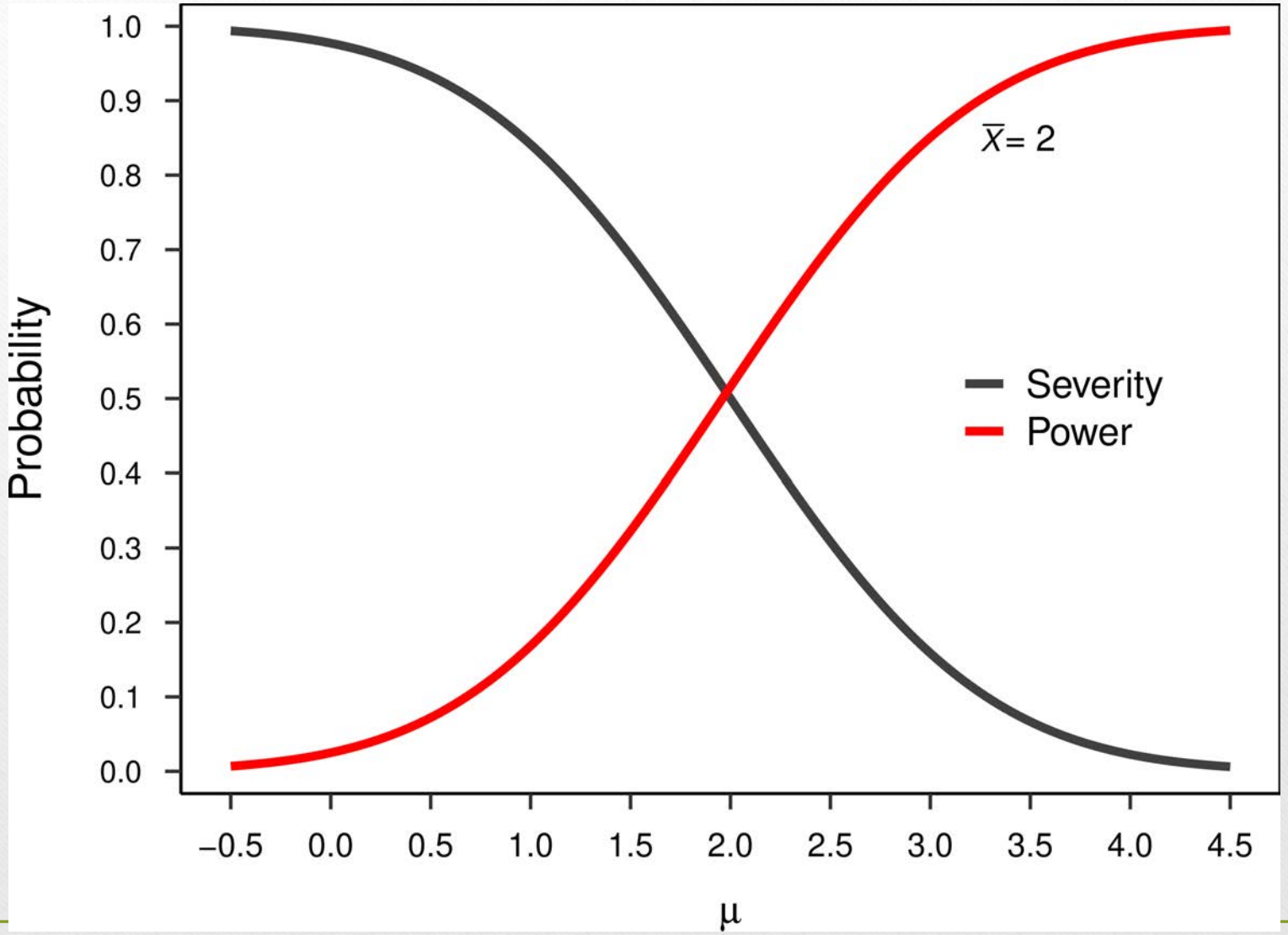
# What about fallacies of non-significant results?

- Not evidence of no discrepancy, but not uninformative even for simple significance tests—
- Minimally: Test wasn't capable of distinguishing the effect from sampling variability
- May also be able to find upper bounds $\mu_1$

# Setting upper bounds



# Severity vs Power



**Why do some accounts say a result just significant at level α is stronger evidence of ($\mu > \mu_1$) as POW($\mu_1$) increases?**

One explanation is the following comparative analysis:

Let $x$ = Test T rejects $H_0$ *just at* α = .02

$$\frac{\Pr(x;\mu_1)}{\Pr(x;\mu_0)} = \frac{\text{POW}(\mu_1)}{\alpha}$$

POW($\mu_1$) = Pr(Test T rejects $H_0$; $\mu_1$)—the numerator.

As $\mu_1$ increases, POW($\mu_1$) in numerator increases, so the more evidence ($\mu > \mu_1$)—but this is wrong!

# Recap: Mayo

The skeptical consumer of statistics: show me what you've done to rule out ways you can be wrong.

- Biasing selection effects alter a method's error probing capacities

These endanger all methods, but many methods lack the antenna to pick up on them.

Fisherian and N-P tests can block threats to error control, but pathological battles result in their being viewed as an inconsistent hybrid

- Where Fisherians can't use power, N-P can't report attained P-values—forfeits features they each need

Can keep the best from both Fisher and N-P: Use error probabilities inferentially

- What alters error probabilities
- alters error probing capabilities
- alters well testedness

Rivals to error statistical accounts **condition on the data**: import of data is through likelihood ratios (LP) (e.g., Bayes factors, likelihood ratios)

So error probabilities drop out

- To the LP holder: what could have happened but didn't is to consider "imaginary data"
- To the severe tester, probabilists are robbed from a main way to block spurious results

The error statistician and LP holders talk past each other

Bayesians may block inferences based on biasing selection effects without appealing to error probabilities:

- high prior belief probabilities to $H_0$ (no effect) can result in a high posterior probability to $H_0$:

**Casualties:**

- Puts blame in wrong place
- How to obtain and interpret them
- Increased flexibility

Recent feuds among statistical thought-leaders lead some to recommend “abandoning” significance & P-value thresholds

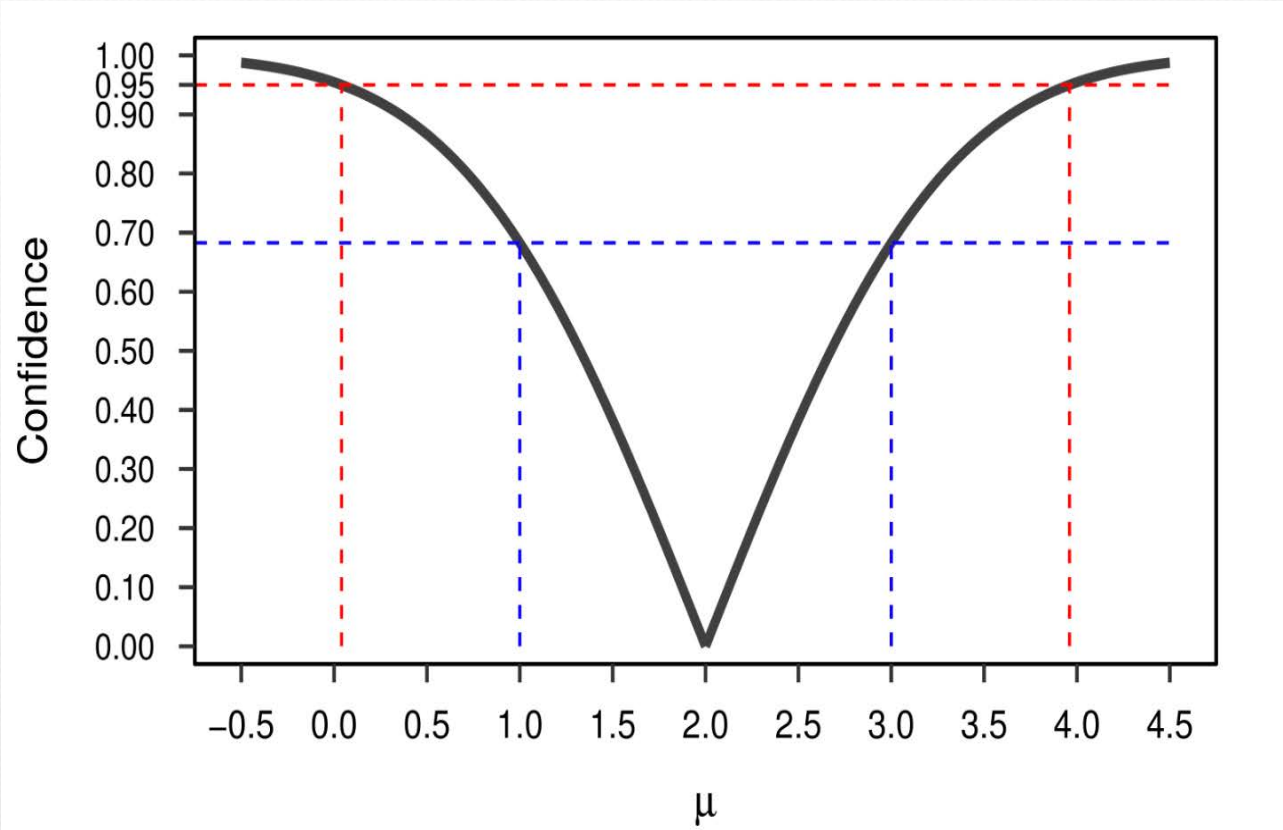
**Casualties:**

- A bad argument: don’t use a method because it may be used badly.
- No thresholds, no tests, no falsification
- Harder to hold researchers accountable for biasing selection effects
- No tests of assumptions

Mayo and Hand (2022)

- We *reformulate tests* to report the extent of discrepancies that are and are not indicated with severity
- Avoids fallacies
- Reveals casualties of equating concepts from schools with different aims
- If time:

confidence intervals are also improved;
with CIs it's "the CI only" movement that's the casualty


Confidence
1.00
0.95
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
μ

In appraising statistical reforms ask:

- what's their notion of probability?*
- What's their account of statistical evidence (LP?)

*If the parameter has a genuine frequentist distribution, frequentists can use it too—deductive updating

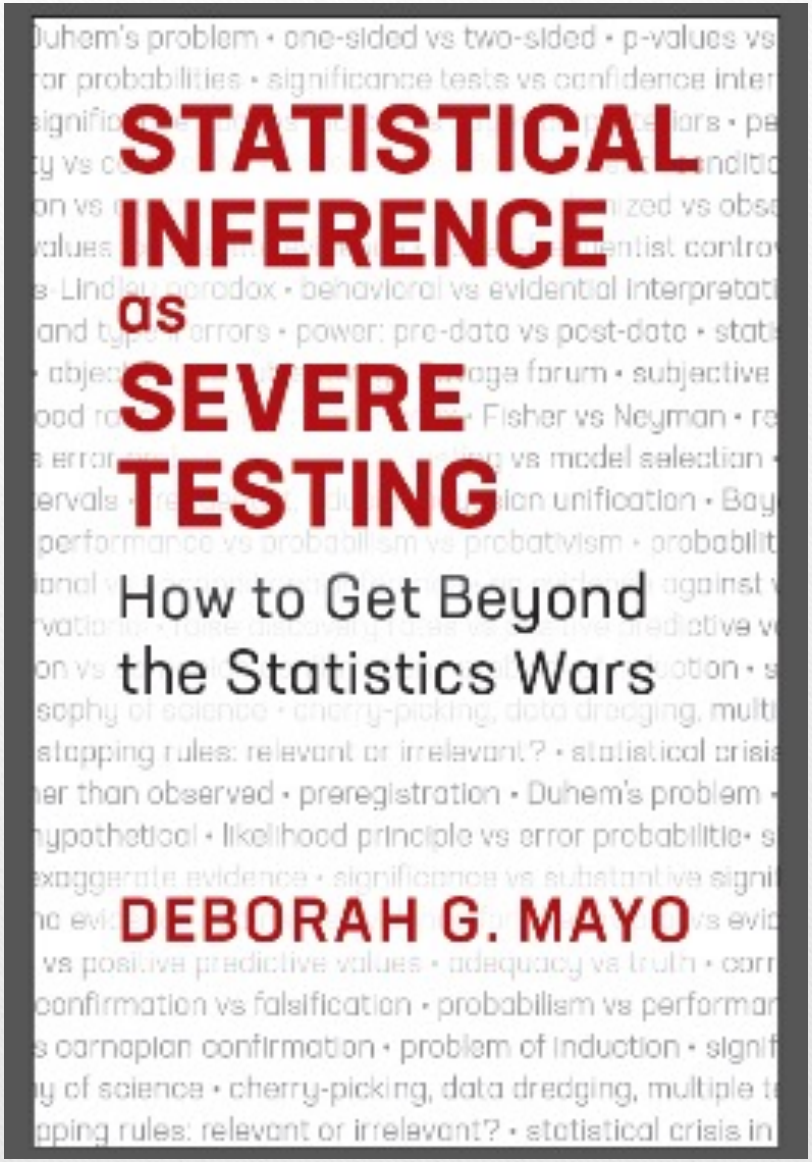
- A silver lining to distinguishing rival concepts–can use different methods for different contexts
- Some Bayesians may find their foundations for science in error statistics
    - Stop refighting the stat wars (by 2034?)

- Attempts to “reconcile” tools with different aims lead to increased conceptual confusion.

# In the context of the skeptical consumer of statistics, methods should be:

- *directly altered* by biasing selection effects
- able to *falsify* claims statistically,
- able to *test statistical model* assumptions.
- able to *block inferences* that violate minimal severity

# For those contexts: we shouldn't throw out the error control baby with the bad statistics bathwater


STATISTICAL INFERENCE as SEVERE TESTING
How to Get Beyond the Statistics Wars
DEBORAH G. MAYO

## References


- Amrhein, V., Greenland, S. and McShane B. (2019). “Comment: Retire Statistical Significance”, *Nature* 567: 305-7. (Online, 20 March 2019).
- Bayarri, M., Benjamin, D., Berger, J., Sellke, T. (2016). “Rejection Odds and Rejection Ratios: A Proposal for Statistical Practice in Testing Hypotheses." *Journal of Mathematical Psychology* 72: 90-103.
- Benjamin, D., Berger, J., Johannesson, M., et al. (*2017). “Redefine Statistical Significance”, Nature Human Behaviour 2, 6–10*
- Benjamini, Y., De Veaux, R. D., Efron, B., Evans, S., Glickman, M., Graubard, B. I., He, X., Meng, X.-L., Reid, N., & Stigler, S. M. (2021). The asa president's task force statement on statistical significance and replicability. *Annals of Applied Statistics*, 15(3), 1084–1085.
- Berger, J. O. (2003). ‘Could Fisher, Jeffreys and Neyman Have Agreed on Testing?’ and Rejoinder’, Statistical Science 18(1), 1–12; 28–32.
- Berger, J. O. (2006). “The Case for Objective Bayesian Analysis.” *Bayesian Analysis* 1 (3): 385–402.
- Berger, J. O. and Wolpert, R. (1988). *The Likelihood Principle*, 2nd ed. Vol. 6 Lecture Notes-Monograph Series. Hayward, CA: Institute of Mathematical Statistics.
- Casella, G. and Berger, R. (1987). “Reconciling Bayesian and Frequentist Evidence in the One-sided Testing Problem”, *Journal of the American Statistical Association* 82(397), 106-11.
- Cox, D. R. (1958). “Some Problems Connected with Statistical Inference”, *Annals of Mathematical Statistics* 29(2), 357-72.
- Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge: Cambridge University Press.
- Cox, D. R., and Mayo, D. G. (2010). “Objectivity and Conditionality in Frequentist Inference.” *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science.* Mayo and Spanos (eds.), 276–304. CUP.
- Fisher, R. A. (1930). “Inverse Probability”. *Mathematical Proceedings of the Cambridge Philosophical Society* 26(4), 528-35.
- Fisher, R. A. (1947). *The Design of Experiments* 4th ed., Edinburgh: Oliver and Boyd.
- Fisher, R. A. (1955). “Statistical Methods and Scientific Induction”, *Journal of the Royal Statistica; Society: Series B* 17(1): 69-78.

- Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd. Reprinted in R. A. Fisher (1990).
- Gigerenzer, G. (2004). "Mindless Statistics", *Journal of Socio-Economics.* 33(5): 587-606.
- Good, I. J. (1971). "The Probabilistic Explication of Information, Evidence, Surprise Causality, Explanation and Utility" and "Reply" in Godambe, V. and Sprott, D. (eds.), *Foundations of Statistical Inference* pp. 108-22, 131-41. Toronto: Holt, Rinehart and Winston.
- Lindley, D. V. (1971). "The Estimation of Many Parameters." in Godambe, V. and Sprott, D. (eds.), *Foundations of Statistical Inference* 435–455. Toronto: Holt, Rinehart and Winston.
- Mayo, D. G. (1991). "Novel Evidence and Severe Tests," *Philosophy of Science*, 58 (4): 523-552. Reprinted (1991) in *The Philosopher's Annual* XIV: 203-232.
- Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Science and Its Conceptual Foundation. Chicago: University of Chicago Press.
- Mayo, D. (2016). 'Don't Throw Out the Error Control Baby with the Bad Statistics Bathwater: A Commentary on Wasserstein, R. L. and Lazar, N. A. 2016, "The ASA's Statement on p-Values: Context, Process, and Purpose"', *The American Statistician* 70(2) (supplemental materials).
- Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars,* Cambridge: Cambridge University Press.
- Mayo, D. G. (2022). The statistics wars and intellectual conflicts of interest. *Conservation Biology*, *36*(1).
- Mayo, D. G. and Cox, D. R. (2006). "Frequentist Statistics as a Theory of Inductive Inference" in Rojo, J. (ed.) *The Second Erich L. Lehmann Symposium: Optimality*, 2006, Lecture Notes-Monograph Series, Volume 49, Institute of Mathematical Statistics: 247-275.
- Mayo, D. G. and Hand, D. (2022). Statistical significance and its critics: practicing damaging science, or damaging scientific practice*?. Synthese* 200, 220. https://doi.org/10.1007/s11229-022-03692-0
- Mayo, D. G., and A. Spanos. (2006). "Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction." *British Journal for the Philosophy of Science* 57 (2) (June 1): 323–357.

- Mayo, D. G., and A. Spanos (2011). “Error Statistics.” In *Philosophy of Statistics*, edited by Prasanta S. Bandyopadhyay and Malcolm R. Forster, 7:152–198. Handbook of the Philosophy of Science. The Netherlands: Elsevier.
- Neyman, J. & Pearson, E. (1933). “On the Problem of the Most Efficient Tests of Statistical Hypotheses”,, *Philosophical Transactions of the Royal Society of London Series A* 231: 289-33 7. Reprinted in *Joint Statistical Papers,* 1-66.
- Open Science Collaboration (2015). “Estimating the Reproducibility of Psychological Science”, *Science* 349(6251), 943–51.
- Pearson, E. S. & Neyman, J. (1967). “On the problem of two samples”, *Joint Statistical Papers* by J. Neyman & E.S. Pearson, 99-115 (Berkeley: U. of Calif. Press).
- Popper, K. (1959). *The Logic of Scientific Discovery*. London, New York: Routledge.
- Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Boca Raton FL: Chapman and Hall, CRC press.
- Ryan, E. G., Brock, K., Gates, S., & Slade, D. (2020). Do we need to adjust for interim analyses in a Bayesian adaptive trial design? *BMC Medical Research Methodology*, *20*(1).
- Selvin, H. (1970). “A critique of tests of significance in survey research. In *The significance test controversy*, edited by D. Morrison and R. Henkel, 94-106. Chicago: Aldine De Gruyter.
- Senn, S. (2002). A comment on replication, p-values and evidence, s.n.goodman, statistics in medicine 1992; 11:875-879. *Statistics in Medicine*, *21*(16), 2437–44.
- Simmons, J. Nelson, L. and Simonsohn, U. (2012) “A 21 word solution”, *Dialogue:* 26(2), 4–7.
- Simonsohn, U. (2013). Just post it: the lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*(10), 1875–1888.
- Wasserstein, R. and Lazar, N. (2016). “The ASA’s Statement on P-values: Context, Process and Purpose”, *The American Statistician* 70(2), 129–33.
- Wasserstein, R., Schirm, A. and Lazar, N. (2019) Editorial: “Moving to a World Beyond ‘$p < 0.05$’”, *The American Statistician* 73(S1): 1-19. (Disclaimer added June 2022.)