The Importance of Philosophy of Science for Statistical Science and Vice Versa

Deborah G Mayo

Dept of Philosophy, Virginia Tech

Chapman University conference:

"Is Philosophy Useful for Science, and/or Vice Versa?"

Feb 2, 2024

Statistics has long been the subject of philosophical debate marked by unusual heights of passion and controversy

Statisticians themselves are well aware of this

"The statistics wars"

“confusion about the foundations of the subject is responsible, in my opinion, for much of the misuse of the statistics that one meets in fields of application such as medicine, psychology, sociology, economics, and so forth.
(George Barnard 1985, p. 2)

The 1980s saw the move to philosophy of science relevant to practice

Phil Stat was ahead of its time

- philosophers of science could regularly be found in statistics conferences
- less so today—but needed more than ever

(formal epistemology is different)

At one level, statisticians and philosophers of science ask similar questions:

- *What should be observed and what may justifiably be inferred from data?*

- *How well do data confirm or test a model?*

- *How can spurious relationships be distinguished from genuine regularities?*

(How to avoid being fooled by chance)

- *How can we infer more accurate and reliable observations from less accurate ones?*

Statistical methods enter when effects are neither swamped by noise, nor so clear cut that formal assessment is not needed

Two-way street

- Statistics is a kind of “applied philosophy of science” (Kempthorne, 1976).

Statistics ⟷ Philosophy

Statistics → Philosophy

Statistical accounts are used in philosophy of science to:

1) **Model Scientific Inference**—ways to arrive at evidence and inference
2) **Solve (or reconstruct) Philosophical Problems** about inference and evidence (e.g., problem of induction)
3) **Metamethodological Critique (**e.g., should we prefer novel predictions?)

Philosophy of Science → Statistics

- A central job for philosophers of science: minister to conceptual and logical problems of sciences

Today's foundations of statistics are more in turmoil than ever

- Widely used methods (e.g., statistical significance tests) are said to be causing a crisis (and should be "abandoned" or "retired")
- Members of different "schools" of statistics often talk past each other

I focus on the second direction (Phil Sci → Stat Sci)

- What are the key debates really about?
- Critically appraise the recent statistical "reforms"
- How to reformulate/improve the controversial statistical significance tests?

Role of probability: performance or probabilism?

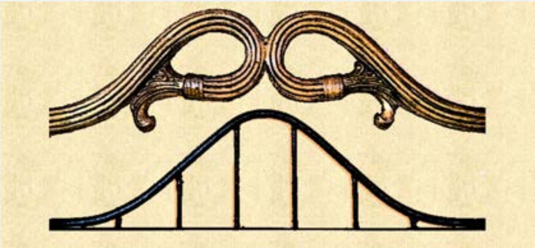
(Frequentist vs. Bayesian)

- End of foundations? (We now have "unifications")
- Long-standing battles simmer below the surface in today's "statistical (replication) crisis in science"
- What's behind it?

I set sail with a minimal principle of evidence

- We don’t have evidence for a claim *C* if little if anything has been done that would have found *C* flawed, even if it is

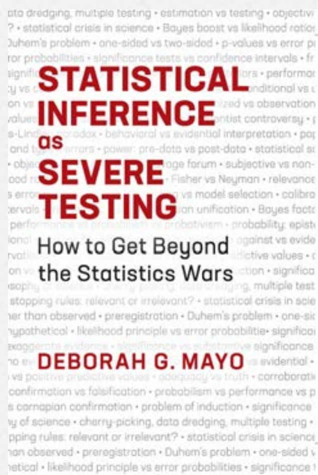
Basis to reformulate frequentist tools



Probability arises (in statistical inference) to assess and control how capable methods are at uncovering and avoiding erroneous interpretations of data

Statistical inference as severe testing

- Excavation tool getting beyond the "statistics wars" and for appraising reforms



So-called Replication crisis leads to "reforms"

Several are welcome:

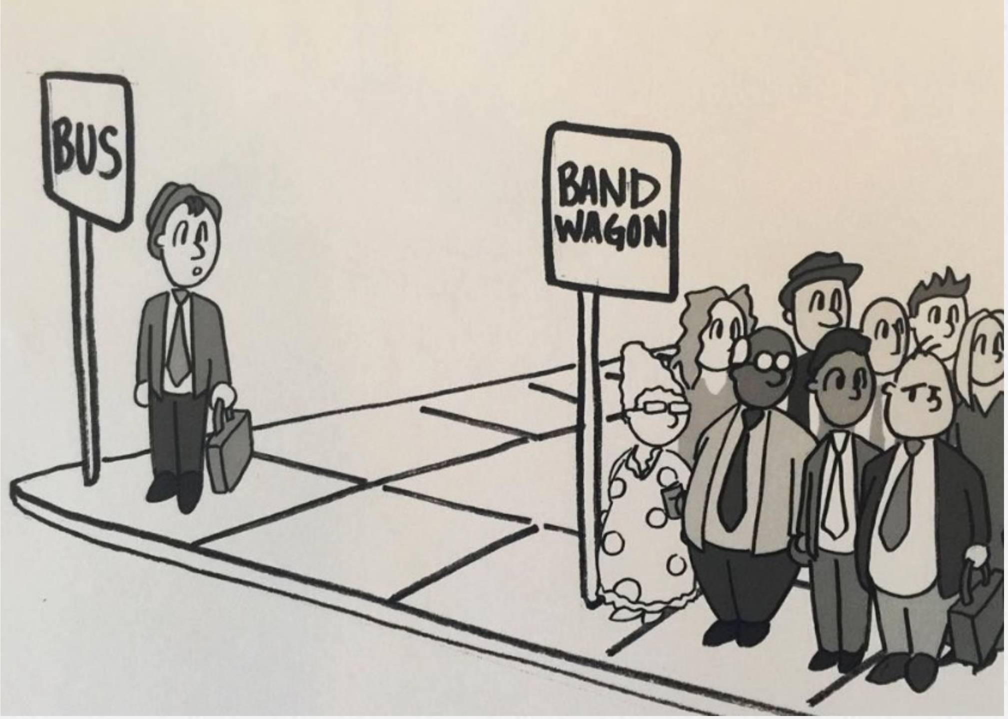
- preregistration of protocol, replication checks, avoid cookbook statistics

Others are radical

- and even lead to violating our minimal requirement for evidence

Being an outside philosopher helps

To combat paradoxical, self-defeating “reforms” requires taking on strong ideological leaders



Most often used tools are most criticized

"(T)he rate of nonreplication of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. …"

(Ioannidis 2005, 696)


OMG!

R.A. Fisher

"[W]e need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result." (Fisher 1947, 14)

Simple significance tests (Fisher)

“to test the conformity of the particular data under analysis with H_0 in some respect:

…we find a function $T = t(\boldsymbol{y})$ of the data, the **test statistic**, such that

- the larger the value of T the more inconsistent are the data with H_0;

$$p = Pr(T \geq t_{0bs};\, H_0)\text{”}$$

(Mayo and Cox 2006, 81)

Testing reasoning

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

- Still not evidence of a genuine statistical effect H_1 yet alone a scientific conclusion H^*—abuses of tests commit such howlers

[1]until an audit is conducted testing assumptions

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

Introduce alternative hypotheses H_0, H_1

$$H_0: \mu = 0 \text{ vs. } H_1: \mu > 0$$

- Trade-off between Type I errors (erroneous rejections) and Type II errors (erroneously failing to reject), power
- Restricts the inference to statistical alternatives (in a model)

Fisher-Neyman (pathological) battles (after 1935)

- The success of N-P optimal error control led to a new paradigm in statistics, overshadows Fisher.

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

N-P & Fisher tests claimed to be an "inconsistent hybrid" (Gigerenzer 2004):

- Fisherians can't use power; N-P testers can't report P-values only fixed error probabilities (e.g., $P < .05$)
  - Fisher & N-P used both pre-data error probabilities and post-data P-value

Fisher & N-P are essentially mathematically identical: their philosophy differs

- They fall under "*error statistical tools*"
- Confidence intervals, N-P and Fisherian tests, resampling, randomization

Both Fisher & N-P: it's easy to lie with biasing selection effects

- Sufficient finagling—cherry-picking, multiple testing, post-data subgroups, trying and trying again, look elsewhere effects—may practically guarantee an impressive-looking effect, even if it's unwarranted by evidence
- Violates error control and severity

Severity Requirement

- 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 *C* flawed, just if it is.

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

Beyond frequentist performance (and probabilism)

- Good long-run performance is a necessary, not a sufficient, condition for severity

Key to solving a central problem for frequentists

- Why is good performance relevant for inference (not just “acceptance sampling” in industry)?
- What bothers you with selective reporting, cherry picking, stopping when the data look good, P-hacking

We cannot say the case at hand has done a good job of avoiding the sources of misinterpreting data

Inferential construal of error probabilities

- "Our goal is to identify a key principle of evidence by which hypothetical error probabilities may be used for inductive inference." (Mayo and Cox 2006)
- There are many attempts, probabilist

A claim *C* is not warranted _______

- ***Probabilism**:* unless *C* is true or probable (gets a probability boost, made comparatively firmer, more believable)
- ***Performance***: unless it stems from a method with low long-run error
  - ***Probativism (severe testing)*** unless something (a fair amount) has been done to probe (& rule out) ways we can be wrong about *C*

Popperian falsification vs logics of induction/ confirmation

Severity is Popper's term,

"the probability of a statement . . . simply does not express an appraisal of the severity of the tests a theory has passed (Popper 1959, 394–5).

- But he never cashed it out adequately (he could have used error statistics)

Comparative logic of support

- Philosopher Ian Hacking (1965) “Law of Likelihood”:

  x support hypothesis H_0 less well than H_1 if,

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

Any hypothesis that perfectly fits the data is maximally likely (even if data-dredged)

- “there *always* is such a rival hypothesis *viz.*, that things just had to turn out the way they actually did” (Barnard 1972, 129)

Error probabilities are “one level above” a fit measure:

Pr(H_0 is less well supported than H_1; H_0) is high for some H_1 or other

"There is No Such Thing as a Logic of Statistical Inference"

- Hacking retracts his Law of Likelihood (LL), (1972, 1980),

  "I now believe that Neyman, Peirce, and Braithwaite were on the right lines to follow in the analysis of inductive arguments"

  (Hacking 1980, 141)

Likelihood Principle (LP)

A pervasive view remains: all the evidence from **x** is contained in the ratio of likelihoods:

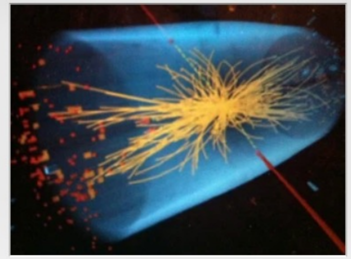
$\Pr(\mathbf{x};H_0) / \Pr(\mathbf{x};H_1)$

- Follows from inference by Bayes theorem
- (Note: the likelihood of a hypothesis is not its probability)

On the LP, error probabilities appeal to something irrelevant

"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)



Aside: The same Lindley prompts an inflammatory letter on the 5 Sigma Higgs discovery in 2012

"Why such an extreme evidence requirement? ..[Is] the particle physics community completely wedded to frequentist analysis? ...has anyone tried to explain what bad science that is?"

(Phy/Stat community did not agree)

Many Bayesian “reforms” offered as alternatives to significance tests, follow the LP

- “Bayes factors can be used in the complete absence of a sampling plan…” (Bayarri, Benjamin, Berger, Sellke 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, *The Likelihood Principle* 1988, 78)

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 Stopping Rule Principle

- "if an experimenter uses this [optional stopping] procedure, then with probability 1 he will eventually reject any sharp null hypothesis, even though it be true".

(Edwards, Lindman, and Savage 1963)

Still, from their Bayesian standpoint the stopping rule is irrelevant

Contrast this with reforms from replication research

- Replication researchers (re)discovered that data-dependent hypotheses and stopping are a major source of spurious significance levels.

- Simmons, Nelson, and Simonsohn (2011):

  "Authors must decide the rule for terminating data collection before data collection begins and report this rule in the articles" (ibid. 1362).

Current Bayesians often echo Edwards, Lindman and Savage (1962)

"[I]f the sampling plan is ignored, the researcher is able to always reject the null hypothesis, even if it is true. ..Some people feel that 'optional stopping' amounts to cheating…. This feeling is, however, contradicted by a mathematical analysis. (Eric-Jan Wagenmakers, 2007, 785)

The mathematical analysis assumes the likelihood principle

Replication Paradox

- ***Significance test critic***: It's too easy to satisfy standard statistical significance thresholds
- ***You***: Why is it so hard to replicate significance thresholds with preregistered protocols?
- ***Significance test critic***: The initial studies were guilty of P-hacking, cherry-picking, data-dredging (QRPs)
- ***You***: So, the replication researchers want methods that pick up on these biasing selection effects.
- ***Significance test* critic**: Actually, "reforms" recommend methods with no need to adjust P-values due to multiplicity

Bayesian clinical trialists say they are are in a quandary

- “The [regulatory] requirement of type I error control for Bayesian adaptive designs causes them to lose many of their philosophical advantages, such as compliance with the likelihood principle]” (Ryan et al. 2020, radiation oncology:
- (A session on “why do we disagree?”).

Probabilists may (indirectly) block intuitively unwarranted inferences
(without error probabilities)

- Likelihoods + prior probabilities
- Give high prior probability to “no effect”

(spike prior)

Problems

- It 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

Most Bayesians (last 15 years) use Conventional or “objective” priors

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

- Berger 2006: “objective Bayesianism”

How should we interpret them?

- "Conventional priors may not even be probabilities…" (Cox and Mayo 2010, 299)

- No agreement on rival systems for default/non-subjective priors, no uninformative priors

(maximum entropy, invariance, maximizing missing information, coverage matching.)

“Default” Bayesian tests are based on the spike prior to the null of no effect

- The posterior probability $\Pr(H_0|\boldsymbol{x})$ can be high while the P-value is low (2-sided test)

The Bayes/Fisher Disagreement or Jeffreys-Lindley Paradox

With a lump of prior to a point null, and the rest spread over the alternative [spike and smear], an α significant result can be high

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

with large *n.*

2-sided H_0: $\mu = 0$ vs. H_1: $\mu \neq 0$.

- To the Bayesian, the P-value exaggerates the evidence against H_0
- The significance tester objects to taking low p-values as no evidence against, or even evidence for, H_0

A Popular Reform: “Redefine Statistical Significance”

“Spike and smear” is the basis for the move to lower the P-value threshold to .005 (Benjamin et al. 2018)

Opposing megateam: Lakens et al. (2018)

- The problem isn’t lowering the probability of type I errors (if that is a chosen balance)

- The problem is assuming there should be agreement between quantities measuring different things

- A silver lining to distinguishing highly probable and highly probed–can use different methods for different contexts

- Trying to reconcile often creates confusion

“A Bayesian Perspective on Severity” van Dongen, Sprenger, Wagenmakers (2022**):**

“As Mayo emphasizes, the Bayes factor is insensitive to variations in the sampling protocol that affect the error rates, i.e., optional stopping”

Bayesians can satisfy severity “regardless of whether the test has been conducted in a severe or less severe fashion”.

What they mean is that even if error statistical severity is violated, data can be much more probable on hypothesis H_1 than on H_0

There's a difference in goals.

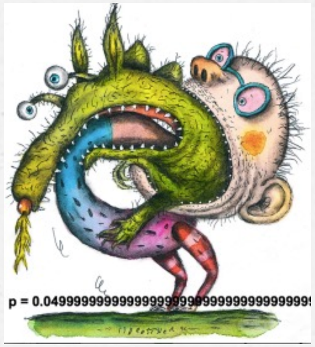
Some Bayesians reject probabilism (Gelman: Falsificationist Bayesian; Shalizi: error statistician)

"[C]rucial parts of Bayesian data analysis, … can be understood as 'error probes' in Mayo's sense"

"[W]hat we are advocating, then, is what Cox and Hinkley (1974) call 'pure significance testing', in which certain of the model's implications are compared directly to the data." (Gelman and Shalizi 2013).

You can't also champion "abandoning statistical significance"—*as he now appears to*

Controversy at the American Statistical Association: 2016 ASA Statement on P-values



- P-values are not measures of effect size, are not posterior probabilities; are invalidated with selection effects; should not by themselves be the basis for substantive claims

2019 ASA Executive Director Editorial: Abandon 'significance'

Surprisingly, in 2019…

- "the 2016 statement "***stopped just short*** of recommending that declarations of 'statistical significance' be abandoned" and announce "***We take that step here***
- "Whether a *p-value* passes any arbitrary threshold should not be considered at all" in interpreting data (Wasserstein, Schirm & Lazar 2019)

- Many claim removing P-value thresholds, researchers lose an incentive to data dredge and multiple test

- I argue the opposite: it’s much harder to hold data-dredgers accountable

No thresholds, 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*

- Is anyone in favor of error probabilities of 50%?

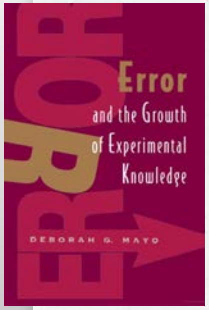
- N-P had an undecidable region

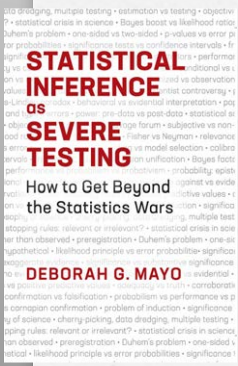
Some researchers lost little time:

“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 the skeptical consumer)

To be clear





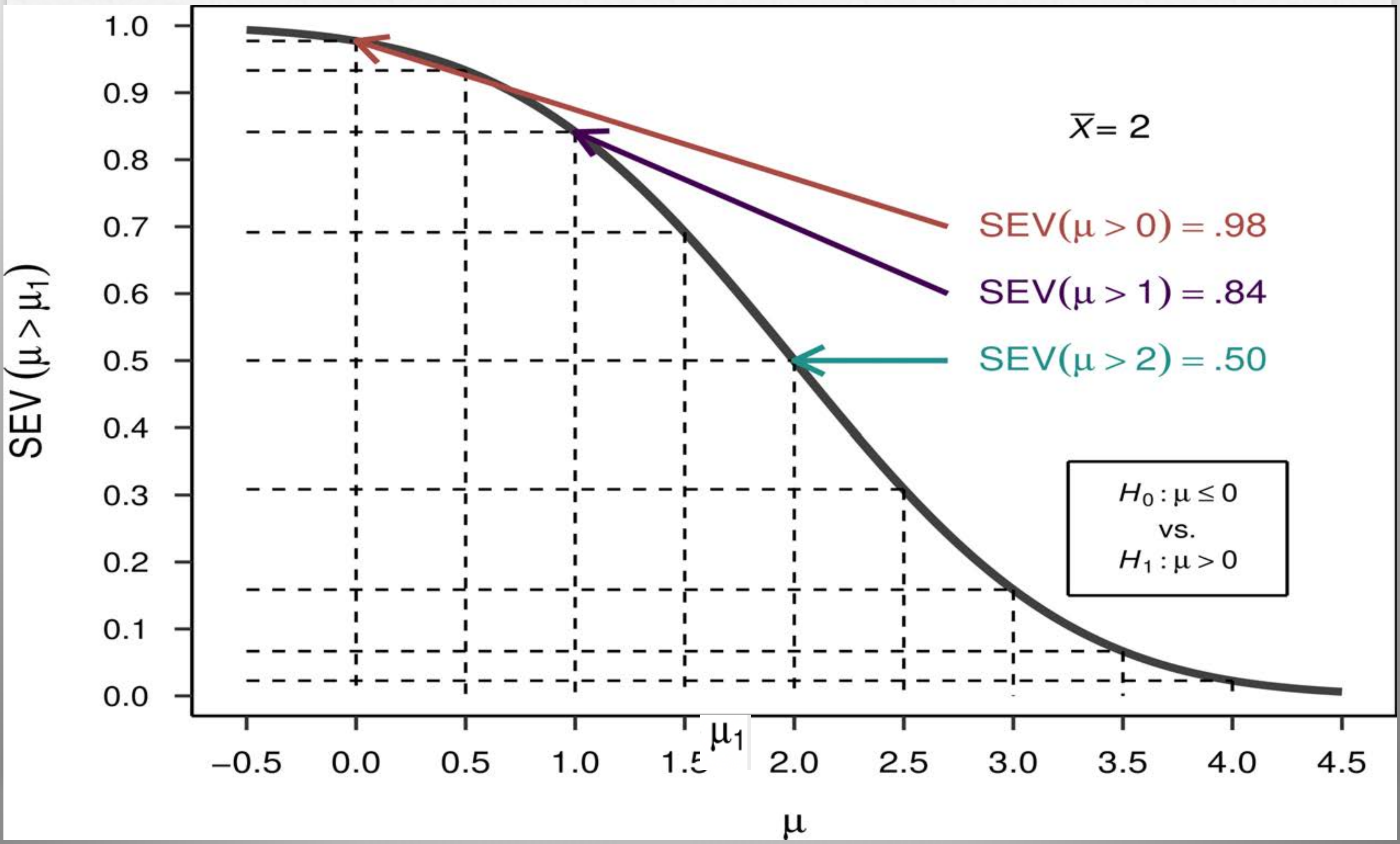
I'm very sympathetic to moving away from accept/reject uses of tests

- In my reformulation of tests*, instead of a binary cut-off (significant or not) the particular outcome is used to infer discrepancies that are or are not warranted with severity

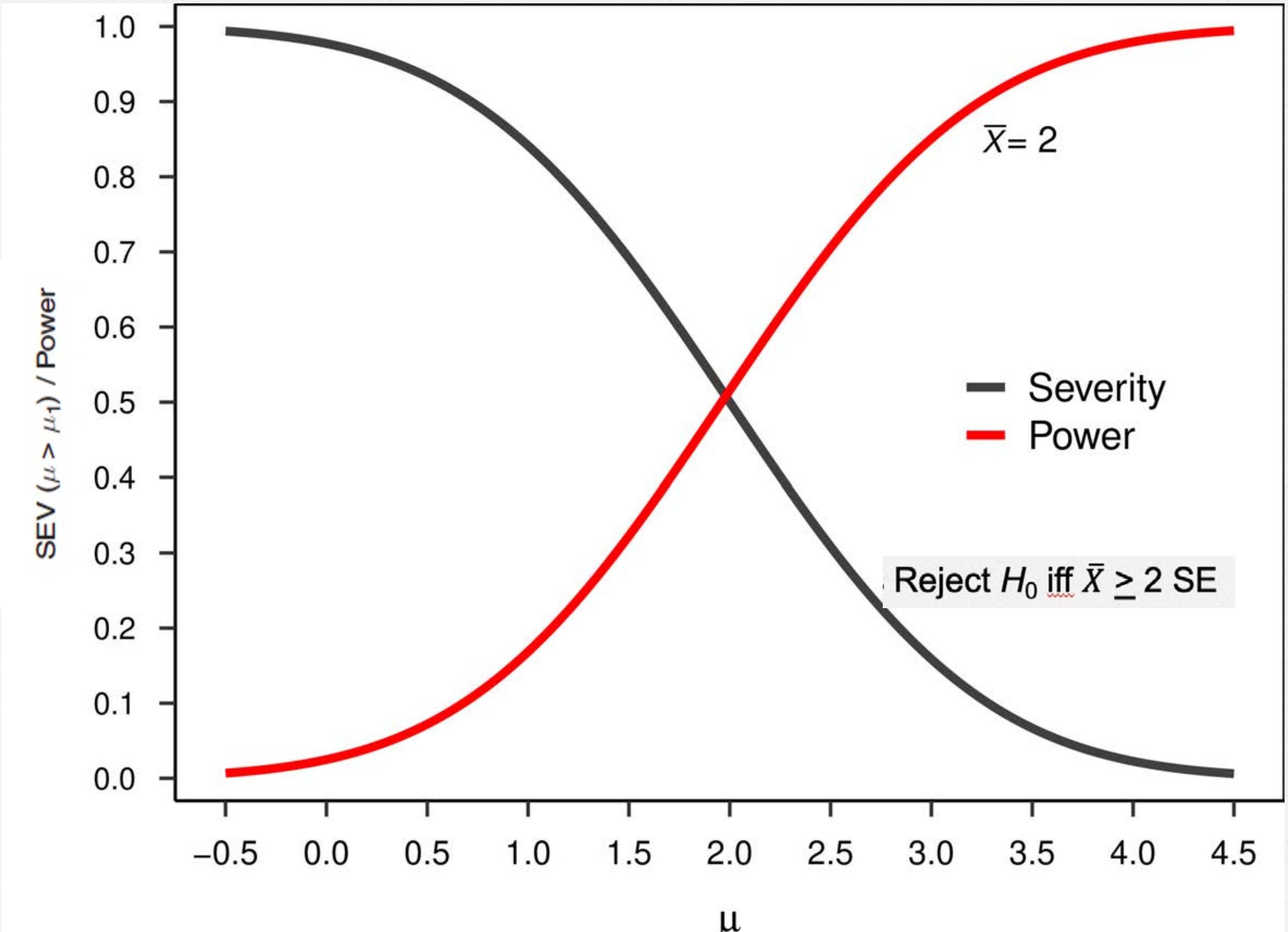
*Mayo1991-2023, developed with others: Aris Spanos, David Cox, David Hand

SEV($\mu > \mu_1$), $\mu_1 = \mu_0 + \delta$ to avoid misinterpreting low P-values

(SE =1)



Severity for $\mu > \mu_1$ vs Power



In the same way, severity avoids the "large *n*" problem

- Fixing the P-value, increasing sample size *n*, the 2SE cut-off gets smaller

- Large *n* is the basis for the Jeffreys-Lindley paradox

Severity tells us:

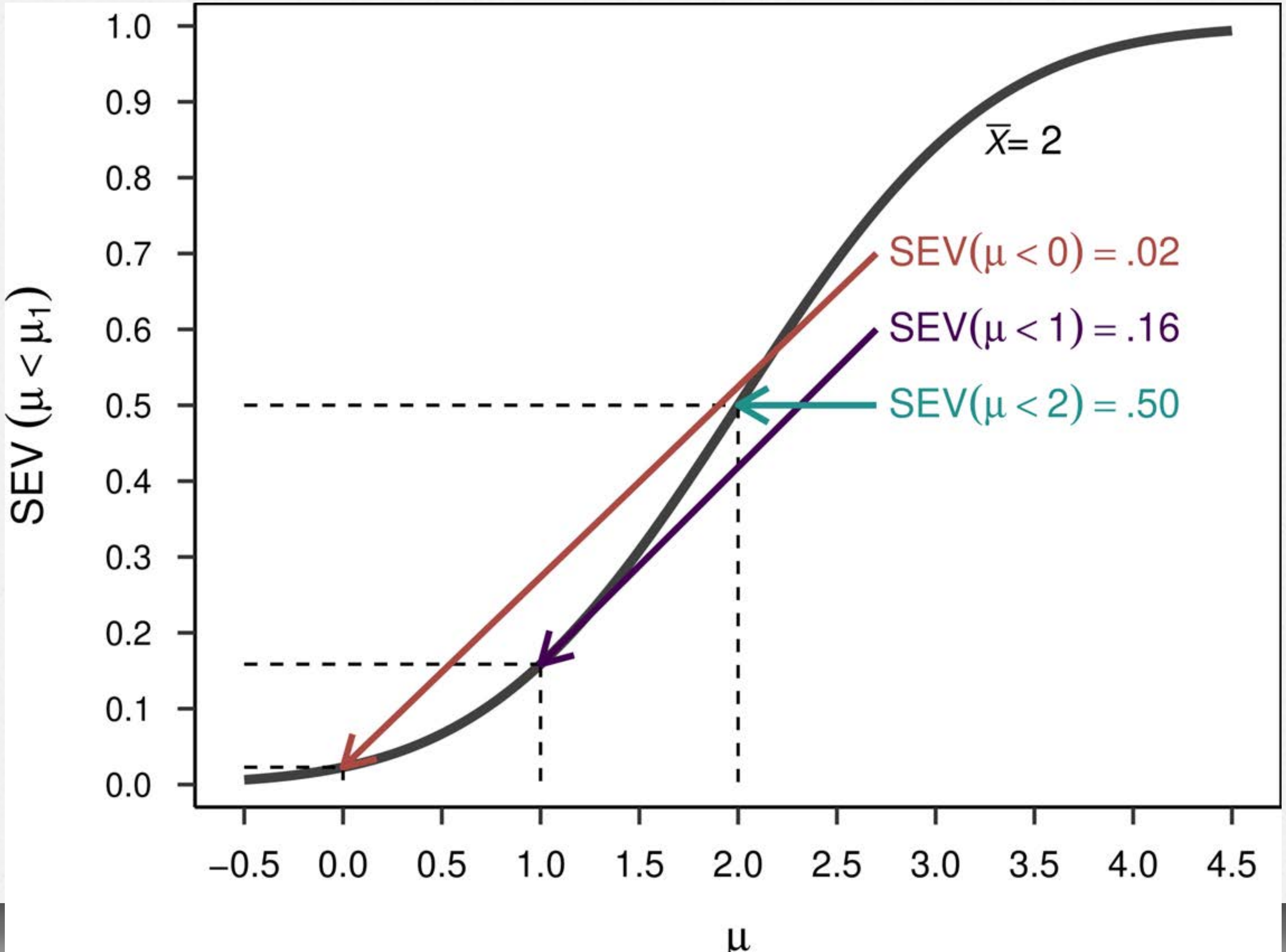
- A difference just significant at level α indicates *less* of a discrepancy from the null if it results from larger (n_1) rather than a smaller (n_2) sample size ($n_1 > n_2$)
- What's more indicative of a large effect (fire), a fire alarm that goes off with burnt toast or one that doesn't go off unless the house is fully ablaze?

- [The larger sample size is like the one that goes off with burnt toast]

What about fallacies of non-significant results?

- Not evidence of no discrepancy, but not uninformative
- Minimally: Test was incapable of distinguishing the effect from noise
- Can also use severity reasoning to rule out discrepancies

SEV($\mu < \mu_1$), to set upper bounds



To close:

The most relevant way philosophy can be relevant to science:

- Solve (or at least illuminate) logical, conceptual, value-laden issues in the field
- Not only working on the subject-matter of the field

I discussed 3 philosophical tasks I have been involved in

- elucidating the core controversies
- critically appraising ”reforms” in statistics
- reformulating statistical significance tests

Others not discussed:

- linking statistical inference to substantive scientific claims
- testing statistic assumptions (auditing)

Thank you!

I'd be glad to have questions

error@vt.edu

References

- Barnard, G. (1972). The logic of statistical inference (Review of "The Logic of Statistical Inference" by Ian Hacking). *British Journal for the Philosophy of Science 23*(2), 123–32.
- Barnard, G. (1985). *A Coherent View of Statistical Inference*, Statistics Technical Report Series. Department of Statistics & Actuarial Science, University of Waterloo, Canada.
- 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. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z
- Benjamini, Y., De Veaux, R., Efron, B., et al. (2021). The ASA President's task force statement on statistical significance and replicability. *The Annals of Applied Statistics*. https://doi.org/10.1080/09332480.2021.2003631.
- Berger, J. O. (2006). 'The Case for Objective Bayesian Analysis' and 'Rejoinder', Bayesian Analysis 1(3), 385–402; 457–64.
- Berger, J. O. and Wolpert, R. (1988). *The Likelihood Principle*, 2nd ed. Vol. 6 Lecture Notes-Monograph Series. Hayward, CA: Institute of Mathematical Statistics.

- Birnbaum, A. (1970). Statistical methods in scientific inference (letter to the Editor).” *Nature, 225* (5237) (March 14), 1033.
- 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. (2006). *Principles of statistical inference*. Cambridge University Press. https://doi.org/10.1017/CBO9780511813559
- Cox, D. and Hinkley, D. (1974). Theoretical Statistics. London: Chapman and Hall.
- Cox, D. R., and Mayo, D. G. (2010). Objectivity and conditionality in frequentist inference. In D. Mayo & A. Spanos (Eds.), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*,, pp. 276–304. Cambridge: Cambridge University Press.
- Edwards, W., Lindman, H., and Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193-242.
- Fisher, R. A. (1947). *The Design of Experiments* 4th ed., Edinburgh: Oliver and Boyd.
- Fisher, R. A., (1955), Statistical Methods and Scientific Induction, *J R Stat Soc* (B) 17: 69-78.
- Gelman, A. (2019). When we make recommendations for scientific practice, we are (at best) acting as social scientists. *European Journal of Clinical Investigation* 49(10): e13165.

- Gelman, A. and Shalizi, C. (2013). Philosophy and the Practice of Bayesian Statistics' and Rejoinder. *British Journal of Mathematical and Statistical Psychology* 66(1), 8–38; 76–80.
- Giere, R. N. (1969). Bayesian Statistics and Biased Procedures. *Synthese* 20(3), 371–87.
- Giere, R. N. (1976). Empirical probability, objective statistical methods, and scientific inquiry. In Foundations of probability theory, statistical inference and statistical theories of science, vol. 2, edited by W. 1. Harper and C. A. Hooker, 63-101. Dordrecht, The Netherlands: D. Reidel.
- Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33*(5), 587–606.
- Glymour, C. (1980). *Theory and Evidence*. Princeton: Princeton University Press.
- Godambe, V., and D. Sprott, eds. 1971. Foundations of statistical inference. Toronto: Holt, Rinehart and Winston of Canada.
- Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge: Cambridge University Press.
- Hacking, I. (1972). 'Review: Likelihood', British Journal for the Philosophy of Science 23(2), 132–7.
- Hacking, I. (1980). The theory of probable inference: Neyman, Peirce and Braithwaite. In D. Mellor (Ed.), *Science, Belief and Behavior: Essays in Honour of R. B. Braithwaite*, Cambridge: Cambridge University Press, pp. 141–60.

- Harper, W. and Hooker, C. (eds.) (1976). Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science, Volume II. Boston, MA: D. Reidel.
- Howson, C. (1984). Bayesianism and support by novel facts. *British Journal for the Philosophy of Science* 35:245-51.
- Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine 2*(8), 0696–0701.
- Kafadar, K. (2019). The year in review…And more to come. President's corner. Amstatnews, 510, 3–4.
- Kempthorne, O. (1976). 'Statistics and the Philosophers', in Harper, W. and Hooker, C. (eds.), 273–314.
- Kyburg, H. E., Jr. (1971). Probability and informative inference. In Foundations of statistical inference, edited by V. Godambe and D. Sprott, 82-103. Toronto: Holt, Rinehart and Winston of Canada.
- Kyburg, H. E., Jr. (1974). The logical foundations of statistical inference. Dordrecht, The Netherlands: D. Reidel.
- Lakens, D., et al. (2018). Justify your alpha. *Nature Human Behavior* 2, 168-71.
- Levi, I. (1980). The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Change. Cambridge, MA: MIT Press.
- Lindley, D. V. (1971). The estimation of many parameters. In V. Godambe & D. Sprott, (Eds.), *Foundations of Statistical Inference* pp. 435–455. Toronto: Holt, Rinehart and Winston.

- Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Science and Its Conceptual Foundation. Chicago: University of Chicago Press.
- Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars,* Cambridge: Cambridge University Press.
- Mayo, D. G. (2019). P-value Thresholds: Forfeit at Your Peril. *European Journal of Clinical Investigation* 49(10): e13170. (https://doi.org/10.1111/eci.13170
- Mayo, D. G. (2020). Significance tests: Vitiated or vindicated by the replication crisis in psychology? *Review of Philosophy and Psychology 12,* 101-120. DOI https://doi.org/10.1007/s13164-020-00501-w
- Mayo, D. G. (2020). P-values on trial: Selective reporting of (best practice guides against) selective reporting*. Harvard Data Science Review 2.1.*
- Mayo, D. G. (2022). The statistics wars and intellectual conflicts of interest. *Conservation Biology : The Journal of the Society for Conservation Biology*, *36*(1), 13861. https://doi.org/10.1111/cobi.13861.
- Mayo, D.G. (2023). Sir David Cox's Statistical Philosophy and its Relevance to Today's Statistical Controversies. *JSM 2023 Proceedings,* DOI: https://zenodo.org/records/10028243.
- Mayo, D. G. and Cox, D. R. (2006). Frequentist statistics as a theory of inductive inference. In J. Rojo, (Ed.) *The Second Erich L. Lehmann Symposium: Optimality*, 2006, pp. 247-275. Lecture Notes-Monograph Series, Volume 49, Institute of Mathematical Statistics.

- Mayo, D.G., Hand, D. (2022). Statistical significance and its critics: practicing damaging science, or damaging scientific practice?. *Synthese 200,* 220.
- Mayo, D. G. and Kruse, M. (2001). Principles of inference and their consequences. In D. Cornfield & J. Williamson (Eds.) *Foundations of Bayesianism, pp. 381-403.* Dordrecht: Kluwer Academic Publishes.
- 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 P. Bandyopadhyay and M. Forster (Eds.), *Philosophy of Statistics*, *7*, pp. 152–198. *Handbook of the Philosophy of Science*. The Netherlands: Elsevier.
- Neyman , J. (1956). Note on an Article by Sir Ronald Fisher, *J R Stat Soc* (B) 18: 288-294.
- Neyman, J. and 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*, 289–337. Reprinted in *Joint Statistical Papers*, 140–85.
- Neyman, J. & Pearson, E. (1967). *Joint statistical papers of J. Neyman and E. S. Pearson*. University of California Press.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science", *Science 349*(6251), 943-51.
- Pearson, E., (1955). Statistical Concepts in Their Relation to Reality, *J R Stat Soc* (B) 17: 204-207.
- 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.
- yan, E., Brock, K., Gates, S., & Slade, D. (2020). Do we need to adjust for interim analyses in a Bayesian adaptive trial design?, *BMC Medical Research Methodolog*y 20(1), 1–9.
- Savage, L. J. (1962). *The Foundations of Statistical Inference: A Discussion*. London: Methuen.
- Seidenfeld, T. (1979). Why I Am Not an Objective Bayesian; Some Reflections Prompted by Rosenkrantz. *Theory and Decision* 11(4), 413–40.
- Simmons, J. Nelson, L. and Simonsohn, U. (2011). A false-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant", *Dialogue: Psychological Science*, *22*(11), 1359-66.
- van Dongen, N., Sprenger, J. & Wagenmakers, EJ. (2022). A Bayesian perspective on severity: Risky predictions and specific hypotheses. *Psychon Bull Rev 30*, 516–533. https://doi.org/10.3758/s13423-022-02069-1
- Wagenmakers, E-J., (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review 14*(5), 779-804.
- Wasserstein, R., & Lazar, N. (2016). The ASA's statement on p-values: Context, process and purpose (and supplemental materials). *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108
- Wasserstein, R., Schirm, A., & Lazar, N. (2019). Moving to a world beyond "p < 0.05" (Editorial). *The American Statistician 73*(S1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Jimmy Savage on the LP:

"According to Bayes' theorem,…. if ***y*** is the datum of some other experiment, and ***if it happens that P(x|μ) and P(y|μ) are proportional functions of μ (that is, constant multiples of each other), then each of the two data x and y have exactly the same thing to say about the values of μ…"*** (Savage 1962, 17)