Severity as a basic concept in philosophy of statistics

16 April 1894 -
5 August 1981

Deborah G Mayo

Dept of Philosophy, Virginia Tech

October 9, 2024: Neyman Seminar

Dept of Statistics, UC Berkeley

I want to begin by thanking the statistics department!

For elaborate organizational arrangements

Ryan Giordano, Amanda Coston,
Xueyin (Snow) Zhang

Panel discussants:

Ben Recht

Philip Stark

Bin Yu

Xueyin (Snow) Zhang

In a conversation with Sir David Cox

COX: Deborah, in some fields foundations do not seem very important, but we both think foundations of statistical inference are important; why do you think?

MAYO: …in statistics …we invariably cross into philosophical questions about empirical knowledge and inductive inference.

Sir David Cox: 15 July 1924 - 18 January 2022

(“A Statistical Scientist Meets a Philosopher of Science” 2011)

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?

Two-way street

Statistics ⟷ Philosophy

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

“Inductive Behavior as a Basic Concept in Philosophy of Science” (Neyman 1957)

Statistics → Philosophy

Statistical accounts are used in philosophy of science to:

1) **Model Scientific Inference**

2) **Solve (or reconstruct) Philosophical Problems** about inference and evidence (e.g., problem of induction)

Problem of Induction

- Failure to justify (enumerative) induction led to building logics of confirmation C(*H*,***e***), like deductive logic: Carnap
- Evidence ***e*** is given, calculating C(*H*,***e***) is formal.
- Scientists could come to the inductive logician for rational degree of confirmation

They did not succeed, and the program is challenged in the 80s (Popper, Lakatos, Kuhn)

Popperians

Evidence *e* is theory-laden, not given: Collected with difficulty, must start with a **problem**

Conjecture and Refutation: EI is a myth

Corroboration, not Confirmation: Those that survive stringent or severe tests

Popper never adequately cashed out the idea—tried adding requirements like novel predictive success to C(h, e)

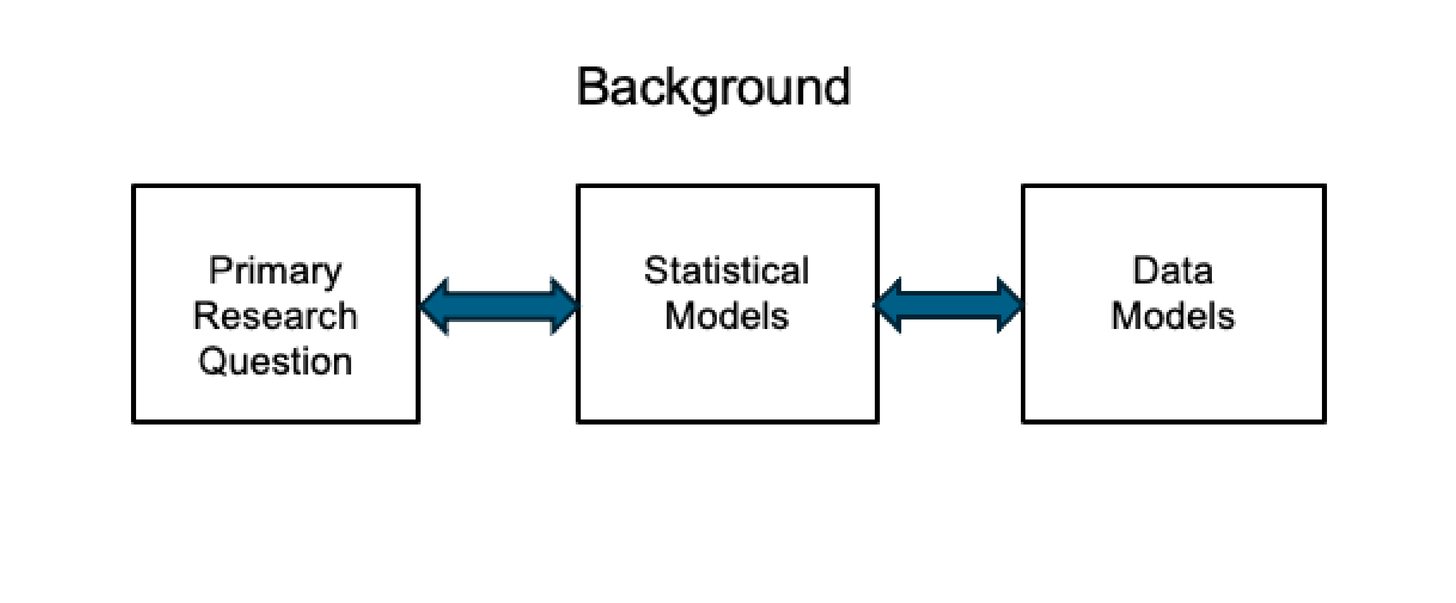
(his anti-inductivism and unfamiliarity with statistical methods)

Neyman and Pearson (N-P): Inductive performance

“We may look at the purpose of tests from another view-point...We may search for rules to govern our behavior…in following which we insure that in the long run of experience, we shall not be too often wrong” (Neyman and Pearson 1933, 141-2)

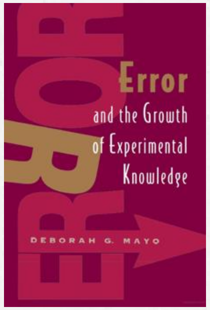
- Probability is assigned, not to hypotheses, but to methods

Series of models



It is a fundamental contribution of modern mathematical statistics to have recognized the explicit need of a model in analyzing the significance of experimental data. (Patrick Suppes 1969, 33)

I suspected that understanding how these statistical methods worked would offer up solutions to the vexing problem of how we learn about the world in the face of error. (Mayo 1996, *Error and the Growth of Experimental Knowledge*)



Philosophy of Science → Statistics

- An obstacle to appealing to a methodology is if its own foundations are problematic

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

- Especially when widely used methods are said to be causing a crisis (and should be “abandoned” or “retired”)

- A separate challenge concerns the “new (data science) paradigm” about the nature and goals of statistics

Central Issue: Role of probability: performance or probabilism?

(Frequentist vs. Bayesian)

- End of foundations? (we do what works)

- 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

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.

Error Statistics

This underwrites the 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)

- (Formal) error statistics is a small part of severe testing
- The severity concept is sufficiently general to apply to any methods, including just using data to solve an error-prone problem (formal or informal)
- But let’s go back to statistical significance tests

Paradox of replication

- Statistical significance tests are often seen 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

Fisher & Neyman and Pearson: it’s easy to lie with biasing selection effects

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

- Cherry-picking, significance seeking, multiple testing, post-data subgroups, trying and trying again—may practically guarantee an impressive-looking effect, even if it’s unwarranted

Simple statistical 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;

$$p = Pr(T \geq t_{0bs}; H_0)”$$

(Mayo and Cox 2006, 81)

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 claim C

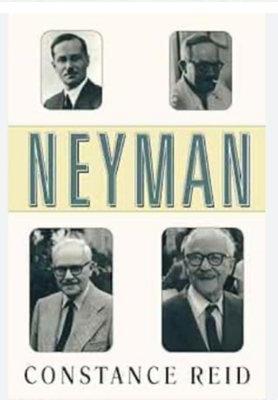
*(until an audit is conducted testing assumptions, and checking biasing selection effects I use “indicate”)

Neyman and Pearson tests (1928, 1933) put Fisherian tests on firmer ground:

- Introduces alternative hypotheses that exhaust the parameter space: H_0, H_1
- Trade-off between Type I errors and Type II errors
- Restricts the inference to the statistical alternative within a model

E. Pearson calls in Neyman “an Anglo-Polish Collaboration: 1926-1934”!

“I was torn between conflicting emotions: a. finding it difficult to understand R.A.F., b. hating [Fisher] for his attacks on my paternal ‘god,’ c. realizing that in some things at least he was right” (Constance Reid 1982, 1998).



Fisher-Neyman split: 1935

- Initially, it was Fisher, Neyman and Pearson vs the “old guard” at University College

- 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)

Main issue: Fisher’s fiducial fallacy

In short: Fisher claimed the fiducial level measures both error control and post-data probability on statistical hypotheses without prior probabilities—in special cases

Fiducial frequency distribution (to distinguish it from inverse [Bayesian] inference)

Neyman’s confidence intervals capture what he assumed Fisher means

“[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, resulting in the break-up

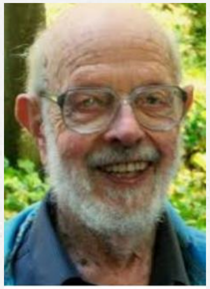
Construed as a deep philosophical difference: p’s vs α’s (Fisher vs N-P)

Fisher claimed that N-P turned “my” tests into acceptance-sampling tools that

“confuse the process appropriate for drawing correct conclusions, with those aimed rather at, let us say, speeding production, or saving money.” (Fisher 1955)

Also, “no scientific worker has a fixed level of significance” used habitually (Fisher 1956)

Lehmann (1993):“The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?



“Unlike Fisher, Neyman and Pearson ...did not recommend a standard [significance] level but suggested that ‘how the balance [between the two kinds of error] should be struck must be left to the investigator.’” (1244)…

“Fisher, although originally recommending the use of such levels, later strongly attacked any standard choice”. (1248)

Earlier phrases are changed

Lehmann’s view might surprise: “you can have your cake and eat it too”

“Should [the report consist] merely of a statement of significance or nonsignificance at a given level, or should a p value be reported? ...One should routinely report the p value and, where desired, combine this with a statement on significance at any stated level.” (1247)

“it gives an idea of how strongly the data contradict the hypothesis” (Lehmann and Romano 2005, 63-4)

p-value as post data error probability

“p_{obs} [the observed p-value] is the probability that we would mistakenly declare there to be evidence against H_0, were we to regard the data under analysis as just decisive against H_0.” (Cox and Hinkley 1974, p. 66)

error probabilities are counterfactual

evidence (p’s) vs error (α’s)?

Neymanian Lehmann is viewing the p-value as evidence

Fisherian Cox, as, performance (or calibration)

Critics say this is to confuse evidence (p’s) and error (α’s)

Severity says it is not.

Neyman’s applied work is more evidential: decision or conclusion

…

“A study of any serious substantive problem involves a sequence of incidents at which one is forced to pause and consider what to do next. In an effort to reduce the frequency of misdirected activities one uses statistical tests” (Neyman 1976).

Since he regards the only logic to be deductive, he uses “inductive behavior”.

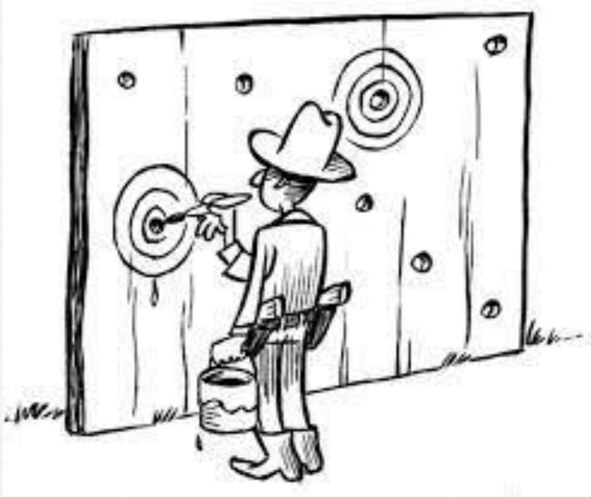
The “act” can be to declare evidence against H_0

Severity gives an evidential twist

- What is the epistemic value of good performance relevant for inference in the case at hand?
- What bothers you with selective reporting, cherry picking, stopping when the data look good, P-hacking
- Not just long-run quality control

We cannot say the test has done its job in the case at hand in avoiding sources of misinterpreting data

The successes are due to the biasing selection effects, not *C*’s truth (e.g., marksmanship)



A claim *C* is not warranted _______

- ***Probabilism****:* unless *C* is true or probable (gets a probability boost, made comparatively firmer)
- ***Performance***: unless it stems from a method that would rarely err
  - ***Probativism (severe testing)*** unless *C* passes a test highly capable of uncovering mistakes

The view of evidence underlying ‘probabilism’ is based on the Likelihood Principle (LP)

All the evidence is contained in the ratio of likelihoods:

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

$\boldsymbol{x}$ supports H_0 less well than H_1 if H_0 is less likely than H_1 in this technical sense

Ian Hacking (1965) “Law of Likelihood”

Comparative logic of support

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

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

(comes from the *sampling distribution* of T)

Hacking: “There is no such thing as a logic of statistical inference”

“I now believe that Neyman, Peirce, and Braithwaite were on the right lines to follow in the analysis of inductive arguments” (1980, 141)

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)

Data Dredging and multiplicity

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

“Repeated testing complicates the interpretation of significance levels. If enough comparisons are made, random error almost guarantees that some will yield ‘significant’ findings, even when there is no real effect . . .” (Kaye and Freedman 2011, 127)

Optional Stopping

- “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, 239)

- “the import of the...data actually observed will be exactly the same as it would had you planned to take *n* observations.” (ibid)

A question

If a method is insensitive to error probabilities, does it escape inferential consequences of gambits that inflate error rates?

It depends on a critical understanding of “escape inferential consequences”

It would not escape consequences for a severe tester:

- What alters error probabilities alters the method’s error probing capability

- Alters the severity of what’s inferred

- Bayesians and frequentists use sequential trials—difference is whether/how to adjust

- Post-data selective inference is a major research area of its own

- With Big Data, new methods to deal with multiplicity are needed, and Juliet Shaffer (1995), a faculty member here, is among the first

Probabilists may block intuitively unwarranted inferences

(without error probabilities)

- Give high prior probability to “no effect” (spike prior)

Can work in some cases, but :

- The believability of data-dredged hypotheses is what makes them so seductive

- Additional source of flexibility, can also be data-dependent

“Why should one's knowledge, or ignorance, of a quantity depend on the experiment being used to determine it” (Lindley 1972)

Data-dredging need not be pejorative

- It’s a biasing selection effect only when it injures severity

- It can even increase severity.

- There’s a difference between ruling out chance variability and explaining a known effect (e.g., DNA testing)

Neyman initially sought a Bayesian account

- rare availability of frequentist priors led him (and Pearson) to develop tools whose validity did not depend on them

- opposes rules for the ideal rational mind (fiducial or Bayesian)--dogmatic, not testable

[Aside: the end of Lehmann’s paper describes where Lehmann thinks Fisher and Neyman do disagree]

Turning to more recent “statistics wars”

Statistics has kept its reputation as being a subject of philosophical debate marked by unusual heights of passion and controversy

American Statistical Association (ASA):

2015: Concerned about the “statistical crisis of replication,” it assembles ~2 dozen researchers

Phillip Stark was one

I was a “philosophical observer”

ASA: 2016 Statement on P-values and Statistical significance [1]

- Warns: p-values aren’t effect-size measures or posteriors
- Data dredging “renders the reported *p*-values essentially uninterpretable”

[1] Wasserstein & Lazar 2016

ASA Executive Director’s Editorial (2019) Abandon ‘statistical significance’: No threshold

- “The 2016 Statement “stopped just short of recommending that declarations of ‘statistical significance’ be abandoned”

- “We take that step here ...Whether a *p-value* passes any arbitrary threshold should not be considered at all when deciding which results to present or highlight”

- Not ASA policy

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

- It is also much harder to hold data-dredgers accountable

- No thresholds, no tests, no falsification

2019 The ASA (President’s) Task Force on Statistical Significance and Replicability (2019-2021) :

“The use of 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 (Karen Kafadar) 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

Phil Stat underlying criticisms

There are plenty of misinterpretations, but there is also a deep difference in philosophies of evidence:

P-values

- exaggerate evidence
- are not evidence,
- must be misinterpreted to provide evidence

P-values

- exaggerate evidence, (if equated with a posterior probability of H_0)
- are not evidence (if the evidence requires the LP)
- must be misinterpreted to provide evidence (if evidence must be probabilist)

- Highly probable differs from being highly well-probed (in the error statistical sense)
- The goals are sufficiently different that we may accommodate both
- Trying to reconcile creates confusion—members of rival school talk past each other

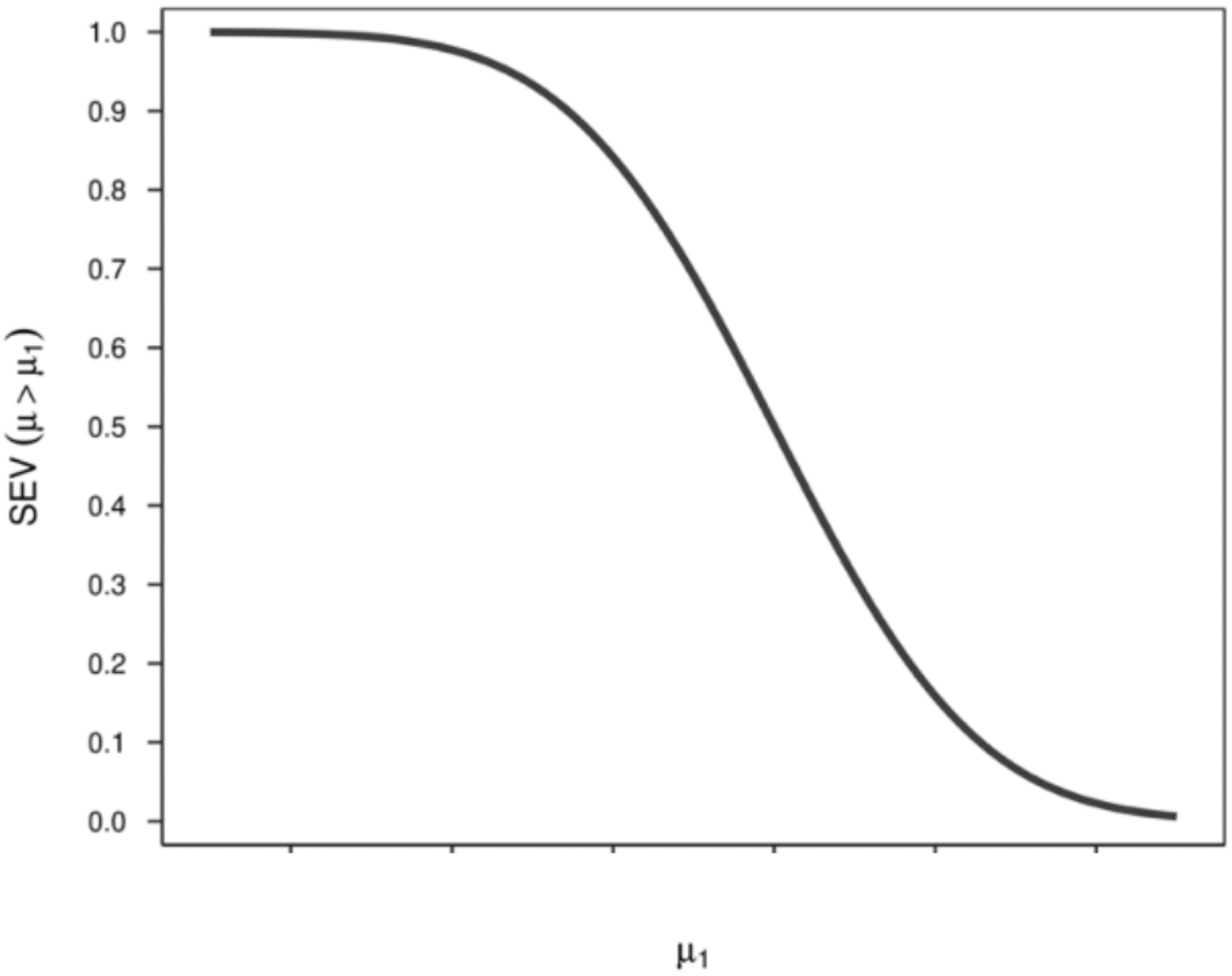
Severity Reformulates tests

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

Mayo 1991-; Mayo and Spanos (2006, 2011); Mayo and Cox (2006); Mayo and Hand (2022)


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
SEV ($\mu > \mu_1$)
μ_1

Severity requires auditing

Check for

1. biasing selection effects,
2. violations of the statistical model assumptions

- generally rely on simple significance tests (e.g., IID)

- “diagnostic checks and tests of fit which, I will argue, require frequentist theory significance tests for their formal justification” (Box 1983, p. 57)

Auditing

3. unwarranted substantive interpretations

The most serious problem, I concur with Philip Stark (2022) is “testing statistical hypotheses that have no connection to the scientific hypotheses”

Falsifying inquiries

We should stringently test (and perhaps falsify) some of the canonical types of inquiry

An inquiry is falsified by showing its inability to severely probe the question of interest.

e.g., cleanliness and morality: *Does unscrambling soap words make you less judgmental on moral dilemmas?*

Falsifying inquiries

We should stringently test (and perhaps falsify) some of the canonical types of inquiry

An inquiry is falsified by showing its inability to severely probe the question of interest.

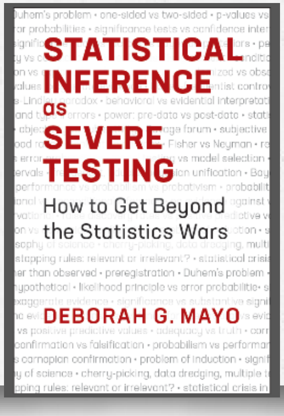
e.g., cleanliness and morality: *Does unscrambling soap words make you less judgmental on moral dilemmas?*

Concluding remark

- I begin with a simple tool: the minimal requirement for evidence
- We have evidence for *C* only to the extent *C* has been subjected to and passes a test it probably would have failed if false
- In formal tests, (relevant) error probabilities can be used to assess this capacity

In the severe testing context, statistical 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



STATISTICAL INFERENCE as SEVERE TESTING

How to Get Beyond the Statistics Wars

DEBORAH G. MAYO

- The goals of the probabilist and the probativist differ
- There are also differences in the goals of formal epistemology and philosophy of statistics

(I’m keen to learn what Snow Zhang thinks)

Statistical inference is broadened

- **Statistical inference:** using data to solve problems of error prone inquiry, qualified with an assessment of the capability of the method to have avoided erroneous interpretations of data
- (Usually) using data to learn about the data generating mechanism qualified...

“inductive learning [in] the frequentist philosophy is that inductive inference requires building up incisive arguments and inferences by putting together several different piece-meal results; ...Although the complexity ...makes it more difficult to set out neatly, ...the payoff is an account that approaches the kind of arguments that scientists build up in order to obtain reliable knowledge and understanding of a field” (Mayo and Cox 2006).

“Frequentist Statistics as a Theory of Inductive Inference” in the proceedings of the second Lehmann Symposium

I see connections with all the panelists:

- With Snow Zhang, the importance of clarifying concepts of knowledge philosophically

- I find the severity requirement in sync with a number of papers by Philip Stark

Bin Yu's veridical data science, with its goal ensuring that every step of the data-to-inference pipeline is robust and trustworthy

Ben Recht's emphasis on probing algorithms for potential weaknesses

How far can data science, AI/ML, explainable AI and black box modeling have error control, at least qualitatively?

Thank you! I’ll be glad to have questions and learn from the panelists!

Jerzy Neyman (1894 - 1981)

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.

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. and Wolpert, R. (1988). *The Likelihood Principle*, 2nd ed. Vol. 6 Lecture Notes-Monograph Series. Hayward, CA: Institute of Mathematical Statistics.

Birnbaum, A. (1969). ‘Concepts of Statistical Evidence’, in Morgenbesser, S., Suppes, P., and White, M. (eds.), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, New York: St. Martin’s Press, pp. 112–43.

Box, G. (1983). ‘An Apology for Ecumenism in Statistics’, in Box, G., Leonard, T., and Wu, D. (eds.), Scientific Inference, Data Analysis, and Robustness, New York: Academic Press, 51–84.

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* pp. 276–304. Cambridge: CUP.

Cox, D. and Mayo, D. (2011). ‘A Statistical Scientist Meets a Philosopher of Science: A Conversation between Sir David Cox and Deborah Mayo’, in Rationality, Markets and Morals (RMM) 2, 103–14.

Edwards, W., Lindman, H., and Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193-242.

Fisher, R. A., (1955), Statistical Methods and Scientific Induction, *J R Stat Soc* (B) 17: 69-78.

Fisher, R. A. (1958). *Scientific Methods for Research Workers* 13th ed., New York: Hafner.

Good, I. J. (1971). ‘The Probabilistic Explication of Information, Evidence, Surprise, Causality, Explanation, and Utility’ and ‘Reply’, in Godambe, V. and Sprott, D. (eds.), pp. 108–22, 131–41. *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.

Kafadar, K. (2019). The year in review…And more to come. President’s corner. Amstatnews, 510, 3–4.

Kaye, D. and Freedman, D. (2011). ‘Reference Guide on Statistics’, in Reference Manual on Scientific Evidence, 3rd edn. pp. 83–178.

Kempthorne, O. (1976). ‘Statistics and the Philosophers’, in Harper, W. and Hooker, C. (eds.), *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science, Volume II,* pp. 273–314. Boston, MA: D. Reidel.

Kuhn, Thomas S. (1970). *The Structure of Scientific Revolutions*. Enlarged (2nd ed.). University of Chicago Press.

Lakatos, I. (1978). *The methodology of scientific research programmes*. Cambridge University Press.

Lehmann, E. (1993). ‘The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?’, Journal of the American Statistical Association 88 (424), 1242–9.

Lehmann, E. and Romano, J. (2005). Testing Statistical Hypotheses, 3rd edn. New York: Springer.

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.

Lindley, D. V. (1972). *Bayesian statistics, a review*. Society for Industrial and Applied Mathematics.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Science and Its Conceptual Foundation. Chicago: University of Chicago Press.

Mayo, D. (2014). On the Birnbaum Argument for the Strong Likelihood Principle (with discussion). *Statistical Science* 29(2), 227–39; 261–6.

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. (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 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. (1934). 'On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection', The Journal of the Royal Statistical Society 97(4), 558–625. Reprinted 1967 Early Statistical Papers of J. Neyman, 98–141.

Neyman, J. (1937). 'Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability', Philosophical Transactions of the Royal Society of London Series A 236(767), 333–80. Reprinted 1967 in Early Statistical Papers of J. Neyman, 250–90.

Neyman, J. (1941). ‘Fiducial Argument and the Theory of Confidence Intervals’, Biometrika 32(2), 128–50. Reprinted 1967 in Early Statistical Papers of J. Neyman: 375–94.

Neyman , J. (1956). Note on an Article by Sir Ronald Fisher, *J R Stat Soc* (B) 18: 288-294.

Neyman, J. (1957). ‘“Inductive Behavior” as a Basic Concept of Philosophy of Science’, Revue de l‘Institut International de Statistique/Review of the International Statistical Institute 25(1/3), 7–22.

Neyman, J. (1976). ‘Tests of Statistical Hypotheses and Their Use in Studies of Natural Phenomena’, Communications in Statistics: Theory and Methods 5(8), 737–51.

Neyman, J. and Pearson, E. (1928). ‘On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I’, Biometrika 20A(1/2), 175–240. Reprinted in Joint Statistical Papers, 1–66.

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.

Pearson, E., (1955). Statistical Concepts in Their Relation to Reality, *J R Stat Soc* (B) 17: 204-207.

Pearson, E., (1970). The Neyman Pearson Story: 1926–34. In Pearson, E. and Kendall, M. (eds.), *Studies in History of Statistics and Probability*, *I*. London: Charles Griffin & Co., 455–77.

Popper, K. (1962). Conjectures and Refutations: The Growth of Scientific Knowledge. New York: Basic Books.

Recht, B. (current). Blog: Arg min (https://www.argmin.net/).

Reid, C. (1998). *Neyman*. New York: Springer Science & Business Media.

Savage, L. J. (1962). *The Foundations of Statistical Inference: A Discussion*. London: Methuen.

Shaffer, J. (1995). Multiple hypothesis testing. *Ann. Rev. Psychol. 46*: 561-84.

Simmons, J. Nelson, L. and Simonsohn, U. (2018). False-positive citations. *Perspectives on Psychological Science 13*(2) 255–259.

Stark, P. B. (2022). Reproducibility, p-values, and type III errors: response to Mayo (2022). *Conservation Biology*, *36*(5), e13986-n/a. https://doi.org/10.1111/cobi.13986

Stark, P.B. (2022). Pay No Attention to the Model Behind the Curtain. *Pure Appl. Geophys. 179,* 4121–4145. https://doi.org/10.1007/s00024-022-03137-2

Suppes, P. (1969). Models of Data, in *Studies in the Methodology and Foundations of Science*, Dordrecht, The Netherlands: D. Reidel, pp. 24–35.

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

Yu, B. and Barter, R. (2024).*Veridical Data Science: The Practice of Responsible Data Analysis and Decision Making* Adaptive Computation and Machine Learning series, The MIT Press. (https://vdsbook.com/)