Statistical "Reforms": Fixing Science or Threats to Replication and Falsification

Deborah G Mayo

June 7, 2022

The Du Bois-Wells Symposium on Socially Aware Data Science and Ethical Issues in Modeling

Mounting failures of replication give a new urgency to critically appraising proposed statistical reforms.

- While many are welcome
  - preregistration
  - replication
  - avoid cookbook statistics
- Others are quite radical!

Replication Crisis Paradox

Critic of P-values: It's much too easy to get a small P-value

Crisis of replication: It is much too difficult to replicate small P-values (with prespecified hypotheses)

Is it easy or is it hard*?*

- 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 practically guarantee a preferred claim *H* appears supported, even if it's unwarranted

- “We knew many researchers - including ourselves - who readily admitted to [biasing selection effects]..but they thought it was wrong the way it's wrong to jaywalk. …simulations revealed it was wrong the way it's wrong to rob a bank.” (Simmons, Nelson and Simonsohn (2018, 255)

- “21 word solution” (2012)

This underwrites key features of significance tests:

- to bound the probabilities of misleading inferences ***error probabilities***

- to constrain the human tendency to selectively favor views they believe in.

They should not be replaced with methods less able to control erroneous interpretations of data.

Not if we want socially aware data science.

(Simple) Statistical significance tests

Significance tests (R.A. Fisher) are a small part of a rich error statistical methodology:

> "…to test the conformity of the particular data under analysis with H_0 in some respect…."
>
> …the **P-value: the probability of an even larger value of** t_{0bs} merely from background variability or noise (Mayo and Cox 2006, 81)

Testing reasoning, as we see it

- If even larger differences than t_{0bs} occur fairly frequently under H_0 (i.e., P-value is not small), there's scarcely evidence of inconsistency with H_0
- Small P-value indicates H_1 *some* underlying discrepancy from H_0 because **very probably (1–P) you would have seen a smaller** difference than t_{0bs} were H_0 true.
- Even if the small P-value is valid, it isn't evidence of a scientific conclusion H^*

Stat-Sub fallacy $\boldsymbol{H_1 \Rightarrow H^*}$

Neyman-Pearson (N-P) put Fisherian tests on firmer footing (1933):

Introduces alternative hypotheses H_0, H_1

H_0: μ ≤ 0 vs. H_1: μ > 0

- Constrains tests by requiring control of both Type I error (erroneously rejecting) and Type II error (erroneously failing to reject) H_0, and power

(Neyman also developed confidence interval estimation at the same time)

N-P tests tools for optimal performance:

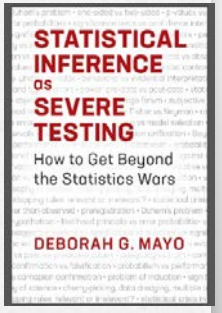
Their success in optimal control of error probabilities gives a new paradigm for statistics

Yet a major criticism: they are only "accept/reject" rules that should be "retired" for anything but quality control decisions, not science (e.g., Amrhein et al 2021)

- Our view is that reliable error control is crucial for warranting specific inferences (not mere long-run quality control)
- What bothers you with selective reporting, cherry picking, stopping when the data look good, P-hacking?
- Not a problem about long-run performance —

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

Basis for the severe testing philosophy of evidence



- Use error probabilities to assess capabilities of tools to probe various flaws ("*probativism*")
- Data supply good evidence for a claim only if it has been subjected to and passes a test that probably would have found it flawed or specifiably false (if it is).
- Altering this capability changes the evidence (on our view)

On a rival view of evidence…

"Two problems that plague frequentist [error statistical] inference: multiple comparisons and multiple looks, or, …data dredging and peeking at the data. The frequentist solution to both problems involves adjusting the P-value…

But adjusting the measure of evidence because of considerations that have nothing to do with the data defies scientific sense" (Goodman 1999, 1010)

(Meta-Research Innovation Center at Stanford)

This view of evidence (probabilist) is based on the Likelihood Principle (LP)

All the evidence is contained in the ratio of likelihoods:

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

x support H_0 less well than H_1 if H_0 is less likely than H_1 in this technical sense

Likelihood Principle (LP) vs error statistics

- Any hypothesis that perfectly fits the data is maximally likely
- Pr(H_0 is less well supported than H_1; H_0) is high for some H_1 or other
- So even with strong support, the error probability associated with the inference is high

All error probabilities violate the Likelihood Principle (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)

Many "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

Bayesian sequential (adaptive) analysts say:

"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)

Our view: Why buy a philosophy that relinquishes error probability control?

This is of relevance to AI/ML

If the field follows critics of explainable AI/ML:

“regulators should place more emphasis on well-designed clinical trials, at least for some higher-risk devices, and less on whether the AI/ML system can be explained”. (Babic et al. 2021)

“Beware Explanations From AI in Health Care”

Bayesians may (indirectly) block implausible inferences

- With a low prior degree of belief on *H* (e.g., *real effect*), the Bayesian can block inferring *H*

Concerns:

- Doesn’t show what has gone wrong—it’s the multiplicity
- The believability of post hoc hypotheses is what makes them so seductive
- Claims can be highly probable (or even known) while poorly probed.

How to obtain and interpret Bayesian priors?

- Most (?) use nonsubjective or default priors, to prevent prior beliefs from influencing posteriors—data dominant
- There is no agreement on which of rival systems to use. (They may not even be probabilities.)

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

There may be ways to combine Bayesian and error statistical accounts

(Gelman: Falsificationist Bayesian; Shalizi: error statistician)

> *"[C]rucial parts of Bayesian data analysis,* such as model checking, *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, 10, 20).

- Can't also champion "abandoning statistical significance", as in a recent "reform"

A recent recommended "reform": Don't say 'significance', don't use P-value thresholds

- In 2019, executive director of the American Statistical Association (ASA) (and 2 co-authors*) announce: "declarations of 'statistical significance' be abandoned"
- "It is time to stop using the term 'statistically significant' entirely."

*Wasserstein, Schirm & Lazar

- We agree the actual P-value should be reported (as all the founders of tests recommended)

- But the 2019 Editorial says prespecified P-value thresholds should not be used at all in interpreting results.

- Many who signed on to the “no threshold view” think by removing P-value thresholds, researchers lose an incentive to data dredge and multiple test and otherwise exploit researcher flexibility

- D. Hand and I (2022) argue: banning the use of P-value thresholds in interpreting data does not diminish but rather exacerbates data-dredging

- In a world without predesignated thresholds, it would be hard to hold the data dredgers accountable for reporting a *nominally* small P-value through ransacking, data dredging, trying and trying again

- What distinguishes data dredged P-values from valid ones is that they fail to meet a prespecified error probability

"Statistical Significance and its Critics: Practicing damaging science, or damaging scientific practice?" (Mayo and Hand 2022)

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, then you do not have a test of C
- Why insist on replications if at no point can you say, the effect has failed to replicate?
- Nor can you test the assumptions of statistical models and likelihood functions

ASA (President's) Task Force on Statistical Significance and Replicability

- In 2019 ASA President appointed a task force of 14 statisticians put in the odd position of needing:

  "to address concerns that [the ASA executive director's "no significance" editorial] might be mistakenly interpreted as official ASA policy"

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

"P-values and significance tests are among the most studied and best understood statistical procedures in the statistics literature".

As an aside to this, Stephen Stigler asks:

"…*Which of the exciting new methods on modern data science machine learning can the same be said?"*

Our view: Reformulate Tests

- Instead of a binary cut-off (significant or not) the particular outcome is used to infer discrepancies that are or are not well-warranted

- Avoids fallacies of significance and nonsignificance, and improves on confidence interval estimation

Severity Reformulation

Severity function: 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

Akin to confidence distributions

To avoid Fallacies of Rejection
(e.g., magnitude error)

Testing the mean of a Normal distribution: H_0: μ ≤ 0 vs. H_1: μ > 0, consider

$$H_0: \mu \leq \mu_1 \text{ vs. } H_1: \mu > \mu_1$$

for $\mu_1 = \mu_0 + \gamma$. If you very probably would have observed a more impressive (smaller) P-value if μ = μ_1 then the data are poor evidence that μ > μ_1.

SEV(μ > μ_1) is low

Power vs Severity for $\mu > \mu_1$

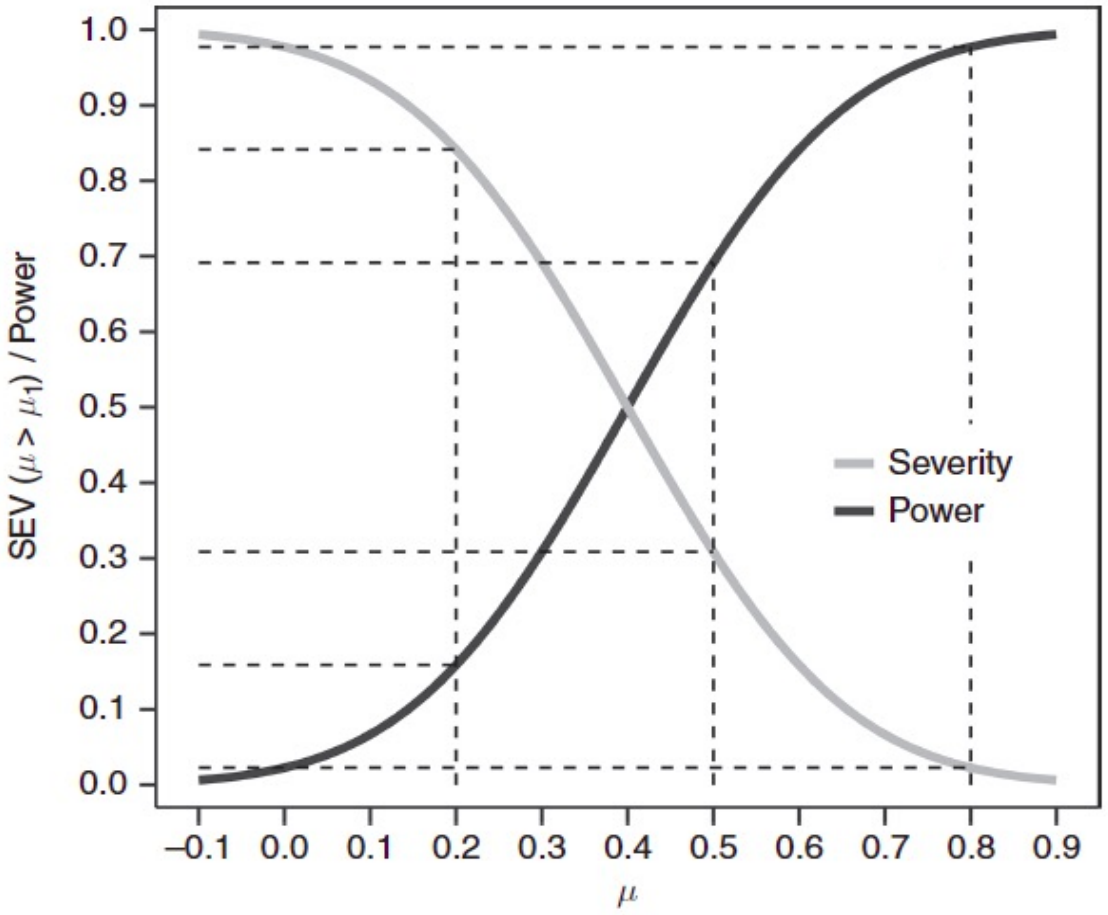


Figure 5.4 Severity for $(\mu > \mu_1)$ vs power (μ_1).

To give an informal construal, consider how severity avoids the "large *n* problem"

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

- Get to a point where ***x*** is closer to the null than various alternatives

Severity tells us:

- an observed difference just statistically 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

[assumptions are presumed to hold]

What About Fallacies of Non-Significant Results?

- They don't warrant 0 discrepancy
- *Using severity reasoning*: rule out discrepancies that very probably would have resulted in larger differences than observed — set upper bounds
- If you very probably would have observed a larger value of test statistic (smaller P-value), were $\mu = \mu_1$ then the data indicate that $\mu < \mu_1$

  $SEV(\mu < \mu_1)$ is high

Brief overview

- The sources of irreplication are not mysterious: in many fields, latitude in collecting and interpreting data makes it too easy to dredge up impressive looking findings even when spurious.

- Some of the reforms intended to fix science enable rather than reveal illicit inferences due to multiple testing, and data-dredging. (either they obey the LP or block thresholds)

- Banning the use of P-value thresholds in interpreting data does not diminish but rather exacerbates data-dredging and biasing selection effects.
- If an account cannot specify outcomes that will not be allowed to count as evidence for a claim—if all thresholds are abandoned—then there is no test of that claim
- We should instead reformulate tests so as to avoid fallacies and report the extent of discrepancies that are and are not indicated with severity.

Mayo (1996, 2018); Mayo and Cox (2006): Frequentist Principle of Evidence (FEV); SEV: Mayo and Spanos (2006), Mayo and Hand (2022)

FEV/SEV *significant result* : A small P-value is evidence of discrepancy γ from H_0, if and only if, there is a high probability the test would have $\boldsymbol{d(X) < d(x_0)}$ were a discrepancy as large as γ absent

FEV/SEV: *insignificant result*: A moderate *P*-value is evidence of the absence of a discrepancy γ from H_0, only if there is a high probability the test would have given a worse fit with H_0 (i.e., $\boldsymbol{d(X) > d(x_0)}$) were a discrepancy γ to exist

References

- Amrhein, V., Greenland, S., & McShane, B. (2019). Comment: Scientists rise up against statistical significance. *Nature*, 567, 305–307. https://doi.org/10.1038/d41586-019-00857-9
- Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from ai in health care. Science, 373(6552), 284–286. https://doi.org/10.1126/science.abg1834
- 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.
- 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.* (Online June 20, 2021.)
- Berger, J. O. and Wolpert, R. (1988). *The Likelihood Principle*, 2nd ed. Vol. 6 Lecture Notes-Monograph Series. Hayward, CA: Institute of Mathematical Statistics.
- Fisher, R. A. 1955. "Statistical Methods and Scientific Induction." *Journal of the Royal Statistical Society, Series B (Methodological)* 17 (1) (January 1): 69–78.
- 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.
- Goodman SN. (1999). "Toward Evidence-based Medical Statistics. 2: The Bayes factor." *Annals of Internal Medicine* 1999; 130:1005 –1013.
- Lindley, D. V. (1971). "The Estimation of Many Parameters." In *Foundations of Statistical Inference*, edited by V. P. Godambe and D. A. Sprott, 435–455. Toronto: Holt, Rinehart and Winston.
- Mayo, D. (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.
- Mayo, D. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars,* Cambridge: Cambridge University Press.

References cont.

- 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, Lecture Notes-Monograph Series, Volume 49, Institute of Mathematical Statistics, pp. 247-275.
- Mayo, D.G., Hand, D. Statistical significance and its critics: practicing damaging science, or damaging scientific practice? *Synthese* **200,** 220 (2022). 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.
- Neyman, J. and Pearson, E.S. (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.
- Ryan, 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 Methodology*, 20(1), 1-9.
- Simmons, J. Nelson, L. and Simonsohn, U. (2012). "A 21 Word Solution." *Dialogue: The Official Newsletter of the Society for Personality and Social Psychology*, 26(2), 4–7.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-positive citations. *Perspectives on Psychological Science, 13*(2), 255 259. https://doi.org/10.1177/1745691617698146
- Stigler, S. (2022). Comment at the NISS Awards Ceremony & Affiliate Luncheon Program (Discussion on discussions leading to Task Force's final report) on August 2, 2021 5 pm ET. https://www.niss.org/events/niss-awards-ceremony-affiliate-luncheon-program
- 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