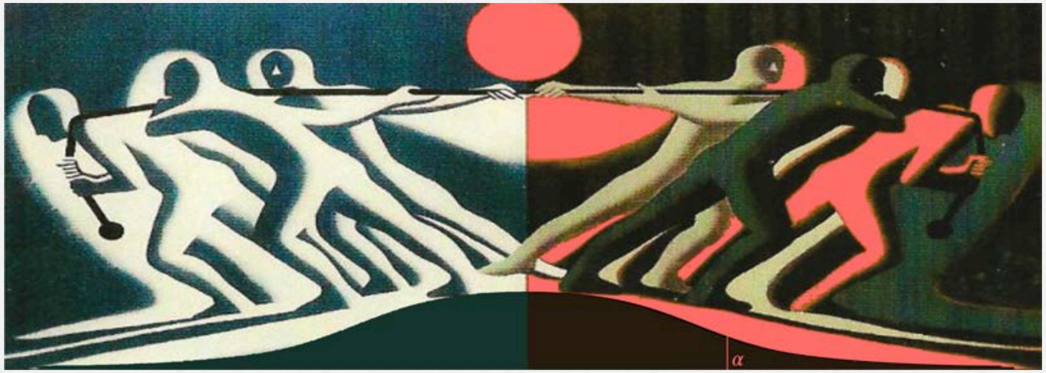
LSE Research Seminar PH500

Current Controversies in Phil Stat or The Statistics Wars (and their casualties)



Deborah G. Mayo

Virginia Tech

The Statistics Wars

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

Role of Probability: performance or probabilism? (Frequentist vs. Bayesian)

- End of foundations? (Unifications and Eclecticism)

- Long-standing battles still simmer below the surface (agreement on numbers)

Let’s brush the dust off the pivotal debates, walk into the museums to hear the founders: Fisher, Neyman, Pearson, Savage and many others in relation to today’s statistical crisis in science (xi)

Stat Sci/Phil Sci
MUSEUM
How-to-draw-funny-cartoons.com
How-to-draw-funny-cartoons.com

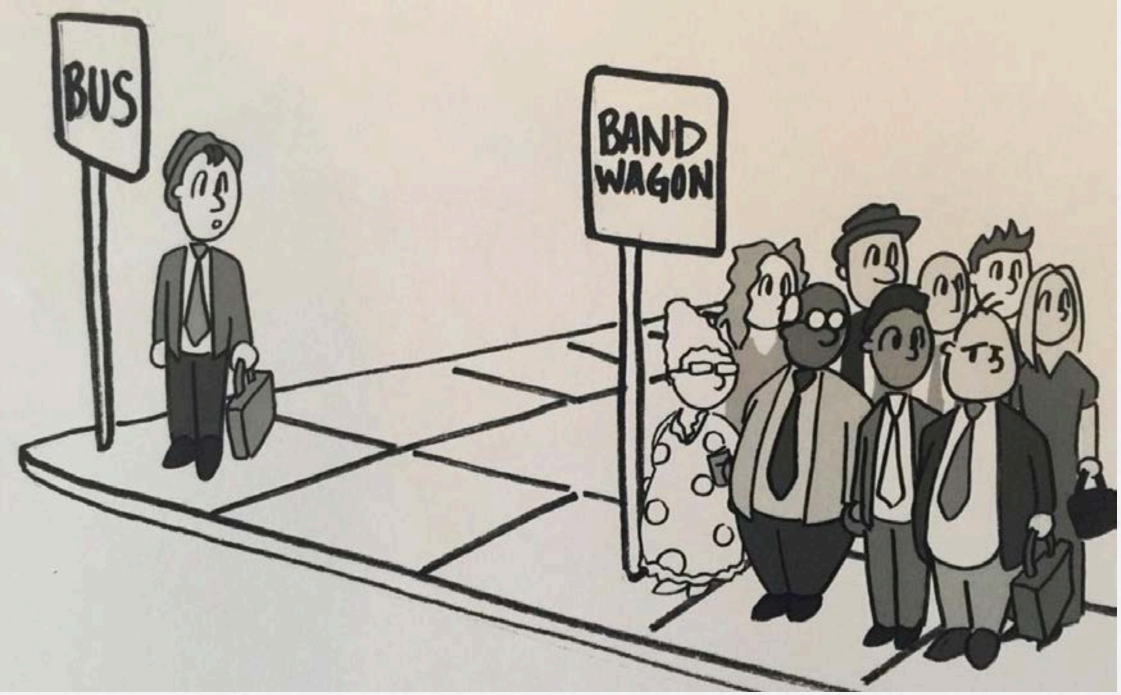
Statistical inference as severe testing

- Main source of the statistical crisis in science?
- We set sail with a simple tool: if little or nothing has been done to rule out flaws in inferring a claim, it has not passed a *severe test*
- Sufficiently general to apply to any methods now in use
- Excavation tool: You needn't accept this philosophy to use it to get beyond today's statistical wars and appraise reforms

Statistical reforms

- Several are welcome: preregistration, avoidance of cookbook statistics, calls for more replication research
- Others are quite radical, and even violate our minimal principle of evidence

The statistics wars have had serious casualties: self-defeating “reforms” and unthinking bandwagon effects



Chutzpah, No Proselytizing

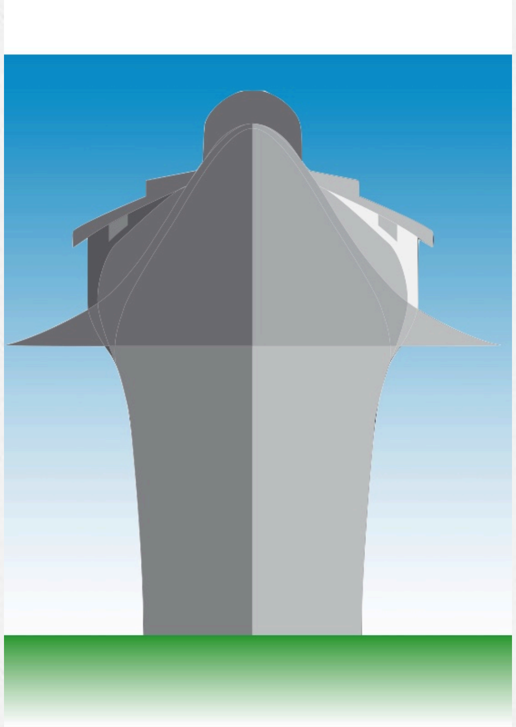
"You will need to critically evaluate …brilliant leaders, high priests, maybe even royalty. Are they asking the most unbiased questions in examining methods, or are they like admen touting their brand, dragging out howlers to make their favorite method look good? (I am not sparing any of the statistical tribes here.)" (p. 12)

- Statistics wars as proxy wars

A philosophical excursion

“I will not be proselytizing for a given statistical school…they all have shortcomings, insofar as one can even glean a clear statement of a given ‘school’” (p.12)

Taking the severity principle, and the aim that we desire to find things out… let’s set sail on a philosophical excursion to illuminate statistical inference.”



Rejected cruise ship

Excursion 1 How to Tell What's True About Statistical inference

Tour I: Beyond Probabilism and Performance

Most often used tools are most criticized

“Several methodologists have pointed out that the high 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)

Do researchers do that?

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)

p-value. …to test the conformity of the particular data under analysis with H_0 in some respect:

…we find a function $d(\boldsymbol{X})$ of the data, the **test statistic**, such that

- the larger the value of $d(\boldsymbol{X})$ the more inconsistent are the data with H_0;
- $d(\boldsymbol{X})$ has a known probability distribution when H_0 is true.

…the p-value corresponding to any $d(\boldsymbol{x})$ (or $d_{0bs})$

$$p = p(t) = Pr(d(\boldsymbol{X}) \geq d(\boldsymbol{x}); H_0)$$

(Mayo and Cox 2006, 81; d for t, x for y)

Testing Reasoning

- If even larger differences than d_{0bs} occur fairly frequently under H_0 (i.e., P-value is not small), there's scarcely evidence of incompatibility with H_0
- Small P-value indicates *some* underlying discrepancy from H_0 because **very probably you would have seen a less impressive** difference than d_{0bs} were H_0 true.
- This still isn't evidence of a genuine statistical effect H_1, let alone a scientific conclusion *H**

Stat-Sub fallacy ***H => H****

Don’t let the tail wag the dog

“It would be a mistake to allow the tail to wag the dog by being overly influenced by flawed statistical inferences” (Cook et al. 2019)

In response to suggesting the concept of statistical significance be dropped (Amhrein et al. 2019)

Fallacy of rejection

- *H** makes claims that haven’t been probed by the statistical test

- The moves from experimental interventions to *H** don’t get enough attention–but your statistical account should block them

Neyman-Pearson (N-P) tests:

A null and alternative hypotheses H_0, H_1 that are exhaustive*

H_0: μ ≤ 0 vs. H_1: μ > 0

“no effect” vs. “some positive effect”

- So this fallacy of rejection ***H_1 ➔ H**** is blocked
- Rejecting H_0 only indicates statistical alternatives H_1 (how discrepant from null)

*(introduces Type 2 error)

Casualties from (Pathological) Fisher-N-P Battles

- Neyman Pearson (N-P) give a rationale for Fisherian tests
- All is hunky dory until Fisher and Neyman began fighting (from 1935) largely due to professional and personality disputes
- A huge philosophical difference is read into their in-fighting

Get beyond the inconsistent hybrid charge

- Major casualties from the “inconsistent hybrid”: Fisher–inferential; N-P–long run performance

- Fisherians can’t use power

- N-P testers must adhere to fixed error probabilities (P <); can’t report P-values (P =)

Erich Lehmann (Neyman’s student)

“It is good practice to determine … the smallest significance level…at which the hypothesis would be rejected for the given observation. ..**the P-value gives an idea of *how strongly the data contradict the hypothesis [and] enables others to reach a verdict based on the significance level of their* choice.**”

(Lehmann and Romano 2005, 63-4)

Error Statistics

- Fisher and N-P both fall under tools for "appraising and bounding the probabilities (under respective hypotheses) of seriously misleading interpretations of data" (Birnbaum 1970, 1033)–*error probabilities*
- I place all under the rubric of *error statistics*
- 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, significance seeking, multiple testing, post-data subgroups, trying and trying again—may practically guarantee a preferred claim *H* gets support, even if it’s unwarranted by evidence
- Violates severity

Severity Requirement:

If the test had little or no capability of finding flaws with *H* (even if *H is* incorrect), then agreement between data $\boldsymbol{x}_0$ and *H* provides poor (or no) evidence for *H*

- Such a test fails a *minimal requirement* for a stringent or severe test
- N-P and Fisher did not put it in these terms but our severe tester does

Need to Reformulate Tests

Severity function: SEV(Test T, data ***x***, claim *C*)

- Tests are reformulated in terms of a discrepancy γ from H_0

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

Requires a third role for probability

Probabilism. To assign a degree of probability, confirmation, support or belief in a hypothesis, given data $\boldsymbol{x}_0$ (absolute or comparative)

(e.g., Bayesian, likelihoodist, Fisher (at times))

Performance. Ensure long-run reliability of methods, coverage probabilities (frequentist, behavioristic Neyman-Pearson, Fisher (at times))

Only probabilism is thought to be inferential or evidential

What happened to using probability to assess error probing capacity?

- Neither “probabilism” nor “performance” directly captures it
- Good long-run performance is a necessary, not a sufficient, condition for severity

Key to solving a key problem for frequentists

- Why is 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 problems about long-runs—

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

© Dario Battisti 2013

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 with low long-run error

- ***Probativism (severe testing)*** unless something (a fair amount) has been done to probe ways we can be wrong about *C*

Rogues Gallery of informal Examples of BENT

Bad Evidence No Test (BENT): Texas marksman

(Having drawn a bull’s eye around tightly clustered shots, claims marksman ability) (p. 19)

A severe test: My weight

Informal example: To test if I've gained weight between the time I left for London and my return, I use a series of well-calibrated and stable scales, both before leaving and upon my return.

All show an over 4 lb gain, none shows a difference in weighing EGEK, I infer:

H: I've gained at least 4 pounds

- Properties of the scales are akin to the properties of statistical tests (performance).

- No one claims the justification is merely long run, and can say nothing about my weight.

- We infer something about the source of the readings from the high capability to reveal if any scales were wrong

- “lift-off”: an overall inference is more reliable that the premises

The severe tester assumed to be in a context of wanting to find things out

- I could insist all the scales are wrong—they work fine with weighing known objects—but this would prevent correctly finding out about weight…..

- What sort of *extraordinary circumstance* could cause them all to go astray just when we don't know the weight of the test object?

- So we have ***weak*** and ***strong*** severity (p. 23)

Jump to Tour II: Error Probing Tools vs. Logics of Evidence (p. 30)

To understand the stat wars, start with the holy grail–a purely formal (syntactical) logic of evidence

It should be like deductive logic but with probabilities

Engine behind probabilisms (e.g., Carnapian confirmation theories Likelihood accounts, Bayesian posteriors)

Battles about roles of probability trace to philosophies of inference

According to modern logical empiricist orthodoxy, in deciding whether hypothesis h is confirmed by evidence e, . . . we must consider only the statements h and e, and *the logical relations [C(h,e)] between them.* It is quite irrelevant whether e was known first and h proposed to explain it, or whether e resulted from testing predictions drawn from h”.
(Popperian, Alan Musgrave 1974, p. 2)

Likelihood Principle (LP)

In logics of induction, like probabilist accounts (as I’m using the term) the import of the data is via the ratios of *likelihoods* of hypotheses

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

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

A pivotal disagreement in the philosophy of statistics battles

Comparative logic of support

- **Ian Hacking (1965)** "Law of Likelihood":

  x support hypothesis H_0 less well than H_1 if,

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

(rejects in 1980)

- 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).

Royall’s trick deck

”According to the [LL], the hypothesis that the deck consists of 52 aces of diamonds is better supported than the hypothesis that the deck is normal [by the factor 52]. (p. 38)

“even if the deck is normal we will always claim to have found strong evidence that it is not”.

Error Probability:

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

"In order to fix a limit between 'small' and 'large' values of [the likelihood ratio] *we must know how often such values appear when we deal with a true hypothesis*." (Pearson and Neyman 1967, 106)

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)

Optional Stopping:

Error probing capacities are altered not just by data dredging, but also via data dependent stopping rules:

H_0: no effect vs. H_1: some effect

$X_i \sim N(\mu, \sigma^2)$, 2-sided H_0: $\mu = 0$ vs. H_1: $\mu \neq 0$.

Instead of fixing the sample size *n* in advance, in some tests, *n* is determined by a *stopping rule*:

- Keep sampling until H_0 is rejected at ("nominal") 0.05 level

Keep sampling until sample mean M $\geq$ 1.96 SE

- *Trying and trying again*: Having failed to rack up a 1.96SE difference after 10 trials, go to 20, 30 and so on until obtaining a 1.96 SE difference

SE = $\sigma/\sqrt{n}$

Nominal vs. *Actual* significance levels:

- With *n* fixed the Type 1 error probability is 0.05

- With this stopping rule the actual significance level differs from, and will be greater than 0.05

  (proper stopping rule)

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)

- Understandably, they observe, the significance tester frowns on this, or at least requires adjustment of the P-values

- “Imagine instead if an account advertised itself as ignoring stopping rules” (43)
- “[the] irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design that had been lost by classical emphasis on significance levels (in the sense of Neyman and Pearson).” (Edwards, Lindman, and Savage 1963, 239)
- “…these same authors who warn that to ignore stopping rules is to guarantee rejecting the null hypothesis even if it’s true” (43) declare it irrelevant.

What counts as cheating depends on statistical philosophy

- Are they contradicting themselves?

- “No. It is just that what looks to be, and indeed is, cheating from the significance testing perspective is not cheating from [*their*] Bayesian perspective.” (43)

21 Word Solution

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

- Statistical critics, Simmons, Nelson, and Simonsohn (2011) place at the top of their list:

“Authors must decide the rule for terminating data collection before data collection begins and report this rule in the articles” (Simmons, Nelson, and Simonsohn 2011, 1362).

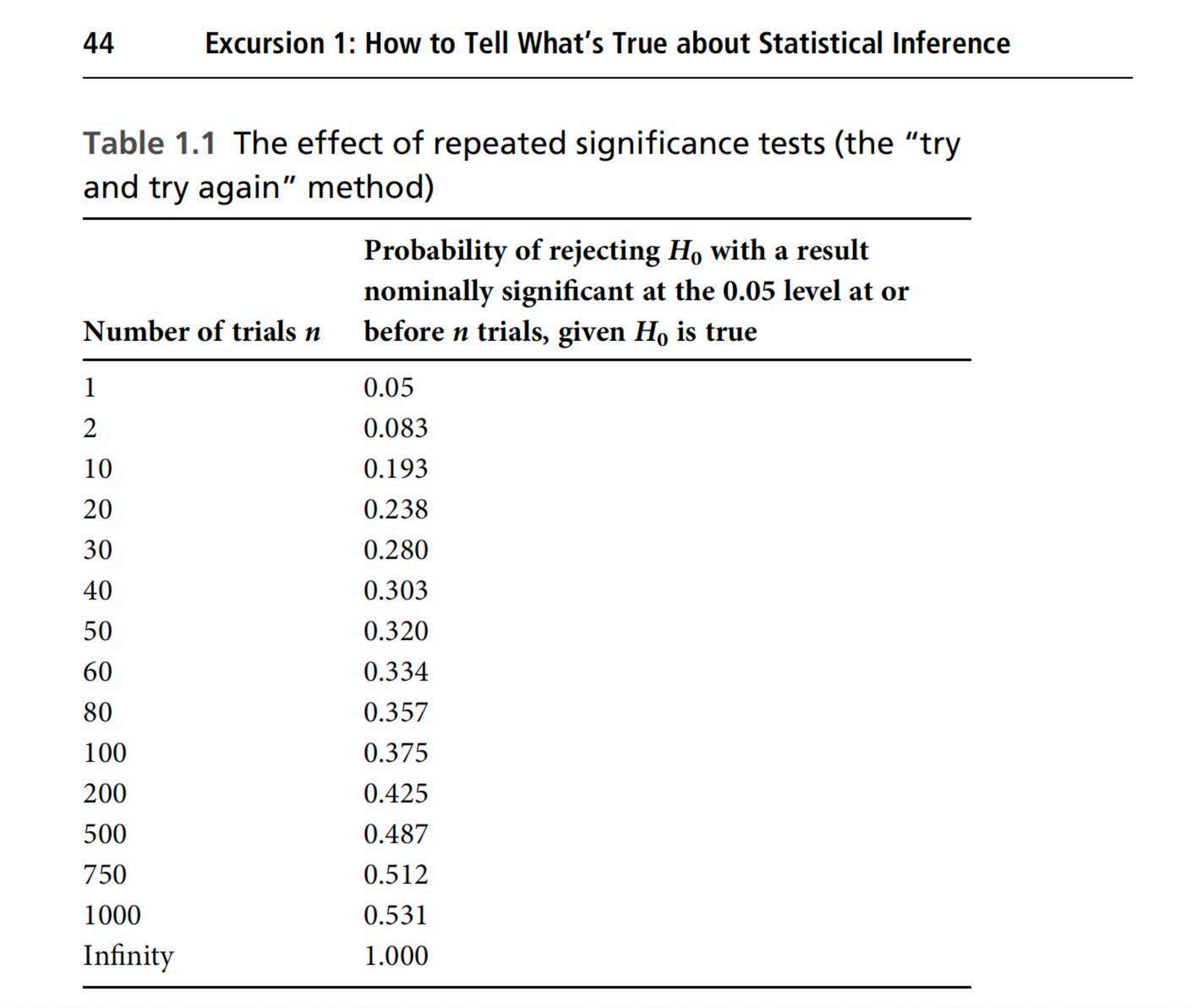


Table 1.1 The effect of repeated significance tests (the "try and try again" method)

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

Berger and Wolpert: LP

- 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)

- “A significance test inference, therefore, depends not only on the outcome that a trial produced, but also on the outcomes that it could have produced but did not” (Howson and Urbach 1993, 212 , SIST p. 49))

- “intentions” should be irrelevant
- “How Stopping Rules Drop Out” (ibid., 45) is derived

Probabilists can still block intuitively unwarranted inferences
(without error probabilities)?

- Supplement with subjective beliefs: What do I believe? As opposed to What is the evidence? (Royall)

- Likelihoodists + prior probabilities

Richard Royall

1. What do I believe, given x
2. What should I do, given x
3. How should I interpret this observation x as evidence? (comparing 2 hypotheses)

For #1–degrees of belief, Bayesian posteriors

For #2–frequentist performance

For #3–LL (p. 33)

Don’t confuse evidence and belief (in the case of the trick deck), p. 38

Bayes Rule

$$\Pr(H|\boldsymbol{x}) = \frac{\Pr(\boldsymbol{x}|H)\Pr(H)}{\Pr(\boldsymbol{x}|H)\Pr(H) + \Pr(\boldsymbol{x}|{\sim}H)\Pr({\sim}H)}$$

(p. 24)

- Requires an exhaustive set of hypotheses

Problems with appealing to priors to block inferences based on selection effects

- Could work in some cases, but still wouldn’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 and biasing selection effects

No help with the severe tester's key problem

- How to distinguish the warrant for a single hypothesis *H* with different methods

  (e.g., one has biasing selection effects, another, pre-registered results and precautions)?

- Since there's a single *H,* its prior would be the same

Casualties: Criticisms of data-dredgers lose force

If a Bayesian critic turns to giving H_0 (no effect) a high prior probability to mount a criticism, the data-dredger can deflect this saying “you can always counter any effect this way”

Current State of Play in Bayesian-Frequentist Wars

1.3 View from a Hot-Air Balloon (p. 23)

How can a discipline, central to science and to critical thinking, have two methodologies, two logics, two approaches that frequently give substantively different answers to the same problems? Is complacency in the face of contradiction acceptable for a central discipline of science? (Donald Fraser 2011, p. 329)

Bayesian-Frequentist Wars

The Goal of Frequentist Inference: Construct procedures with frequentist guarantees

good long-run performance

The Goal of Bayesian Inference: Quantify and manipulate your degrees of beliefs

belief probabilism

Larry Wasserman (p. 24)

But now we have marriages and reconciliations (pp. 25-8)

Most Bayesians (last decade) use “default” priors: unification (pp 25-8)

- ‘Eliciting’ subjective priors too difficult, scientists reluctant for subjective beliefs to overshadow data

> “[V]irtually never would different experts give prior distributions that even overlapped” (J. Berger 2006, 392)

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

Marriages of Convenience

Why?

For subjective Bayesians, to be less subjective

For frequentists, to have an inferential or epistemic interpretation of error probabilities

How should we interpret them?

- “The priors are not to be considered expressions of uncertainty, ignorance, or degree of belief. Conventional priors may not even be probabilities…” (Cox and Mayo 2010, 299)
- No agreement on rival systems for default/non-subjective priors

Unificationist Bayesians

Contemporary nonsubjective Bayesians concede they "have to live with some violations of the likelihood and stopping rule principles" (Ghosh et al.,2006, 148), since their prior probability distributions are influenced by the sampling distribution.

Is it because ignoring stopping rules can wreak havoc with the well-testedness of inferences?

If that is their aim too, then that is very welcome. Stay tuned (SIST p. 51)

Some Bayesians reject probabilism (Falsificationist Bayesians)

- *“[C]rucial parts of Bayesian data analysis, such as model checking, can be understood as ‘error probes’ in Mayo’s sense” which might be seen as using modern statistics to implement the Popperian criteria of severe tests.*

  (Andrew Gelman and Cosma Shalizi 2013, 10).

Decoupling

Break off stat methods from their traditional philosophies

Can Bayesian methods find a new foundation in error statistical ideas? (p. 27)

Excursion 6: (probabilist) foundations lost; (probative) foundations found

Sum-up

To get beyond today’s statistics wars, we need to understand the jumble of philosophical, statistical, historical and other debates

These are largely hidden in today’s debates & the reforms put forward for restoring integrity

There are ago-old: significance test controversy, Bayes-Frequentist battles

Newer debates: reconciliations (default vs. subjective vs. empirical) Bayesians

To evaluate the consequences of reforms need to excavate the jungle

- As impossible as it seems, I set out to do this
- I begin with a simple tool, underwritten by today's handwringing: the minimal requirement for evidence

(evidence for *C* only if it has been subjected to and passes a test it probably would have failed if false)

- Biasing selection effects make it easy to find impressive-looking effects erroneously
- They alter a method's error probing capacities
- They may not alter evidence (in *traditional* probabilisms): Likelihood Principle
- Members of different tribes talk past each other
- To the LP holder: worry about what could have happened but didn't is to consider "imaginary data" and "intentions"

- To the severe tester, probabilists are robbed from a main way to block spurious results
- Constructive role of replication crisis:

  Biasing selection effects *impinge on* error probabilities

  Error probabilities *impinge on* well-testedness

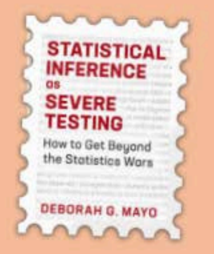
- It directs the reinterpretation of significance tests and other methods
- Probabilists may block inferences without appeal to error probabilities: high prior to H_0 (no effect) can result in a high posterior probability to H_0

- Gives a life-raft to the P-hacker and cherry picker; puts blame in the wrong place
- Severe probing (formal or informal) must take place at every level: from data to statistical hypothesis; from there to substantive claims
- A silver lining to distinguishing highly probable and highly probed–can use different methods for different contexts
- Some Bayesian tribes may find their foundations in error statistics
- Last excursion: (probabilist) foundations lost; (probative) foundations found

SOUVENIR A

HANDY SLOGANS ON HOW TO AVOID LYING WITH STATISTICS:

- Association is not causation
- Statistical significance is not substantive significance
- No evidence of risk is not evidence of no risk
- If you torture data enough, they will confess



STATISTICAL INFERENCE as SEVERE TESTING

How to Get Beyond the Statistics Wars

DEBORAH G. MAYO