Farewell Keepsake in the Form of: 7 Responses (by severe testers) to critics of statistical significance tests

Sessions 14-15 Phil 6014

Intermingled in today’s statistical controversies are some long-standing, but unresolved, disagreements on the nature and principles of statistical methods and the roles for probability in statistical inference. These have important philosophical dimensions that must be recognized to effectively carry out as well as appraise statistical research in today’s social contexts. To combat the dangers of unthinking, bandwagon effects, practitioners and consumers should be in a position to critically evaluate the ramifications of proposed statistical "reforms," as well as respond to often-rehearsed objections to statistical significance tests. I distill some complex philosophical issues by means of 7 simple responses to key challenges.

In *SIST: How to Get Beyond the Stat Wars Requires Chutzpah* (p. 12):

“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.)”

Not just royalty, now, but ASA!

- I set sail with a very simple tool: If little if anything has been done to probe flaws in a claim, then there’s poor evidence for it

- While I think we can all agree to this much, many reforms flout it

The philosophical issues behind the controversies are complex, hotly debated, typically ignored

I attempt to distill them by means of a simple series of responses to key challenges.

“Several methodologists have pointed out that the high rate of nonreplication (lack of confirmation) 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. ….” (John Ioannidis 2005, 0696)

1. Why use a tool that infers* from a single (arbitrary) P-value that pertains to a *statistical* hypothesis *H_0 to a research claim H*?

1. ***Why use a tool that infers* from a single (arbitrary) P-value that pertains to a *statistical* hypothesis *H_0 to a research claim H*?***

RESPONSE: We don’t.

“[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, p. 14)

Statistical Test Reasoning

- Were H_0 a reasonable description of the process, then with very high probability you would not be able to regularly produce statistically significant results.
- So if you do, it’s evidence H_0 is false in the particular manner probed.
- This is the basis for falsification in science.

Fallacy of Rejection

- Even a genuine statistical effect *H* isn't automatically evidence for a substantive *H**.

- *H** makes claims that haven't been probed by the statistical test; statistical significance isn't substantive significance.

- Moves from experimental interventions to *H** don't get enough attention–beyond statistics into theory and measurement (but your account should block them).

- Neyman-Pearson (N-P) tests explicitly restrict the inference to an alternative statistical claim.

(2) Why use an incompatible hybrid (of Fisher and N-P)?

(2) Why use an incompatible hybrid (of Fisher and N-P)?

RESPONSE: They fall under the umbrella of “tools for appraising and bounding the probabilities (under respective hypotheses) of seriously misleading interpretations of data” (Birnbaum 1970, 1033)–**error probabilities**.

Confidence intervals, N-P and Fisherian tests, resampling, randomization.

- N-P and Fisher showed the error control is nullified by **biasing selection effects**

Get Beyond the Inconsistent Hybrid

- We need to get beyond “inconsistent hybrid”: Fisher–inferential; N-P–long run performance
- It leaves us with caricatures: Fisherians can’t use power
- N-P testers adhere to rigid, fixed error probabilities, and can’t report P-values

“It is good practice to determine … the smallest significance level…at which the hypothesis would be rejected for the given observation. This number, the so-called **P-value gives an idea of how strongly the data contradict the hypothesis. It also enables others to reach a verdict based on the significance level of their choice.”** (Lehmann and Romano 2005, pp. 63-4)

- Drop personality labels and NHST–an illicit animal too often associated with cookbook statistics– “statistical tests” or “error statistical tests” will do.

(3) Why apply a method ***that uses error probabilities*, *the sampling plan*,** researcher "***intentions***"? You should ***condition*** on the data (Likelihood Principle LP).

(3) Why apply a method ***that uses error probabilities***, ***the sampling plan***, researcher "***intentions***"? You should ***condition*** on the data (Likelihood Principle LP).

RESPONSE 1: If I condition on the actual data, this precludes error probabilities.

- What bothers you when cherry pickers selectively report?

- Not a problem with long-runs: You *can't say the case at hand* has done a good job of avoiding the sources of misinterpreting data.

Key Principle in a Skeptical Context

- I haven't been given evidence for a genuine effect if the method makes it very easy to find some impressive-looking effect, even if spurious.

- Else it's utterly lacking in stringency or severity.

A claim passes a *severe test* to the extent it has been subjected to, and passes a test, that probably would have detected flaws in C if present.

- Holds outside of tests, to estimation, prediction, problem solving.

You can agree with those who point out:

“P-values can only be computed once the sampling plan is fully known and specified in advance…few people are keenly aware of their intentions, particularly with respect to what to do when when the data turn out not to be significant,” (Wagenmakers 2007, 784)

“In fact, Bayes factors can be used in the complete absence of a sampling plan… .” (Bayarri, Benjamin, Berger, Sellke 2016, 100)

RESPONSE 2 (polite): We’re in different contexts. I’m in one that led to advise the “21 word solution”:

“We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.” (Simmons, Nelson, and Simonsohn 2012, 4)

- I’m in the setting of a (skeptical) consumer of statistics.
- Have you given yourself lots of extra chances in the “forking paths” (Gelman and Loken 2014) between data and inference?

Yet some accounts of evidence:

> “*Two problems that plague frequentist 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)

- To a severe tester, they have a lot to do with the evidence.

Replication Paradox

- ***Test Critic***: It's too easy to satisfy standard significance thresholds
- ***You***: Why do replicationists find it so hard to achieve significance thresholds (with preregistration)?
- ***Test Critic***: Obviously the initial studies were guilty of P-hacking, cherry-picking, data-dredging (QRPs)
- ***You***: So, the replication researchers want methods that pick up on, adjust, and block these biasing selection effects.
- **Test Critic**: Actually "reforms" recommend methods where the need to alter P-values due to data dredging vanishes

- What’s the value of preregistered reports? *Your appraisal is altered by considering the probability that some hypotheses, stopping point, subgroups, etc. could have led to a false positive* –even if informal

(What’s your justification?)

- True, there are many ways to correct P-values, appropriate for different contexts (Bonferroni, false discovery rates).
- The main thing is to have an alert that the reported P-values are invalid or questionable.

RESPONSE 3 (why care about error probabilities?): I don’t want to relinquish my strongest criticism of findings that are the result of biasing selection effects and fishing expeditions.

- Wanting to promote an account that downplays error probabilities, Bayesian critics turn to other means–give H_0 (no effect) a high prior probability in a Bayesian analysis
- Might work in some cases
- The researcher deserving criticism deflects this saying: you can always counter an effect by giving a high prior to a H_0: no effect

- *Puts the blame in the wrong place.* Data-dependent hypotheses (post-data subgroups) are often believable, that’s what makes them seductive.

- Want to say it’s plausible but this is a poor test of it

(4) Why use methods that exaggerate *evidence* against a null hypothesis?

(4) Why use methods that exaggerate *evidence* against a null hypothesis?

RESPONSE: Whether P-values exaggerate, "depends on one's philosophy of statistics …

..based on directly comparing *P* values against certain quantities (likelihood ratios and Bayes factors) that play a central role as evidence measures in Bayesian analysis … Nonetheless, many other statisticians do not accept these quantities as gold standards,"… (Greenland, Senn, Rothman, Carlin, Poole, Goodman, Altman 2016, 342)

“Bayes/Fisher disagreement”

- The “P-values exaggerate” arguments refer to testing a point null hypothesis, a lump of prior probability given to H_0 (or a tiny region around 0). $X_\mathrm{i} \sim N(\mu, \sigma^2)$

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

- The rest appropriately spread over the alternative, an α significant result can correspond to

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

(Jeffreys-Lindley Paradox)

- To a Bayesian this shows P-values exaggerate evidence against…

- Significance testers object to highly significant results being interpreted as no evidence against the null– or even evidence for it!
  High Type 2 error

- “In fact it is not the case that P-values are too small, but rather that Bayes point null posterior probabilities are much too big!” (Casella and R. Berger 1987b, 344).

- “Concentrating mass on the point null hypothesis is biasing the prior in favor of H_0 as much as possible.” (Casella and R. Berger, 1987a, 111)

- Whether tests should use a lower Type 1 error probability is separate; the problem is supposing there should be agreement between quantities measuring different things.

(5) Why do you use a method that presupposes the underlying statistical model?

(5) Why do you use a method that presupposes the underlying statistical model?

RESPONSE: *Au contraire*. I use (simple) significance tests because I want to test my statistical assumptions and perhaps falsify them.

George Box, a Bayesian eclecticist:

> "Some check is needed on [the fact that] some pattern or other can be seen in almost any set of data or facts. This is the object of diagnostic checks [which] require frequentist theory [of] significance tests… ." (Box 1983, 57)

‘Falsificationist Bayesianism’ (Andrew Gelman)

“What 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, 21)

“with an interpretation of probability that can be seen as frequentist in a wide sense and with an error statistical approach to testing assumptions… .” (Gelman and Hennig 2017, 991)

(Why are you prepared to use Bayesian P-value to check accordance of a model?)

(6) Why use a measure that doesn’t report effect sizes?

(6) Why use a measure that doesn’t report effect sizes?

RESPONSE:

- Who says we only report P-values and stop? Given evidence of a real effect, we estimate its magnitude.

- Could use confidence intervals (inversions of tests): values within the (1 – α) CI are not statistically significant at the α level.

Duality of Tests and CIs

(estimating μ in a Normal Distribution)

$\mu > M_0 - 1.96\sigma/\sqrt{n}$ CI-lower

$\mu < M_0 + 1.96\sigma/\sqrt{n}$ CI-upper

M_0 : the observed sample mean

CI-lower: the value of μ that M_0 is statistically significantly greater than at $P= 0.025$

CI-upper: the value of μ that M_0 is statistically significantly lower than at $P= 0.025$

- You could get a CI by asking for these values, and learn indicated effect sizes with tests

The Severe Tester Prefers to Reformulate Tests

- CIs (as standardly used) inherit problems of N-P tests: dichotomous, treat values in the CI the same, justified in terms of long-run performance.

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

- Ex. If you very probably would have observed a more impressive (smaller) P-value than you did, if $\mu = \mu_1$ $(\mu_1 = \mu_0 + \gamma)$; the data are poor evidence that $\mu > \mu_1$.

We get an inferential rationale absent from CIs

CI Estimator:

CI-lower < μ < CI-upper

Because it came from a procedure with good coverage probability

Severe Tester:

μ > CI-lower because with high probability (.975) we would have observed a smaller M_0 if $\mu \leq$ CI-lower

μ < CI-upper because with high probability (.975) we would have observed a larger M_0 if $\mu \geq$ CI-lower

The reformulation can be developed from either Fisherian or N-P perspectives.

SEV: severity (Mayo 1991,1996; Mayo and Spanos 2006, 2011).

FEV: Frequentist Principle of Evidence (Mayo and Cox 2006/2010).

(7) Why do you use a method that doesn’t provide posterior probabilities?

(7) Why do you use a method that doesn’t provide posterior probabilities?

RESPONSE 1: Which notion of a posterior do you recommend? (Note: posteriors aren’t provided by comparative accounts: Bayes Factors, likelihood ratios or model selections.)

- Most Bayesian accounts are default/non-subjective (with data dominant in some sense).
- There is no agreement on suitable priors.
- Even the ordering of parameters will yield different priors.

- Default priors are not expressing uncertainty or degree of belief;

  "…in most cases, they are *not even proper* probability distributions in that they often do not integrate [to] one." (Bernardo 1997, pp. 159-160)

  If priors are not probabilities, "what interpretation is justified for the posterior?" (Cox 2006, p. 77)

- Coherent updating goes by the board.

RESPONSE 2

- I’m not seeking hypotheses that are highly probable (in any formal sense) but methods that very probably would have unearthed discrepancies and improvements–the probability is on the method.

- Bayesian posteriors require a *catchall factor*: all hypotheses that could explain the data; I just want to split off a piece (or variant of a theory) to test.

- A popular new attempt is based on posterior prevalences from diagnostic screening—we’ve done this

A new/old approach that's caught on: Diagnostic Screening (DS) Model

- If we imagine randomly selecting a hypothesis from an urn of nulls 90% of which are true
- *Consider just 2 possibilities:* H_0: no effect H_1: meaningful effect, all else ignored,
- Take the prevalence of 90% as $\Pr(H_0) = 0.9$, $\Pr(H_1)= 0.1$
- Reject H_0 with a single (just) 0.05 significant result, with cherry-picking, selection effects

Then it can be shown most "findings" are false

- **Pr(H_0|Test T rejects H_0) > 0.5**

really: prevalence of true nulls among those rejected at the 0.05 level . 0.5.

Call this: False Finding rate FFR

- **Pr(Test T rejects H_0 | H_0) = 0.05**

Criticism: N-P Type I error probability ≠ FFR

(Ioannidis 2005, Colquhoun 2014)

Why use an approach where your type I error probability differs from the diagnostic model?

But there are major confusions

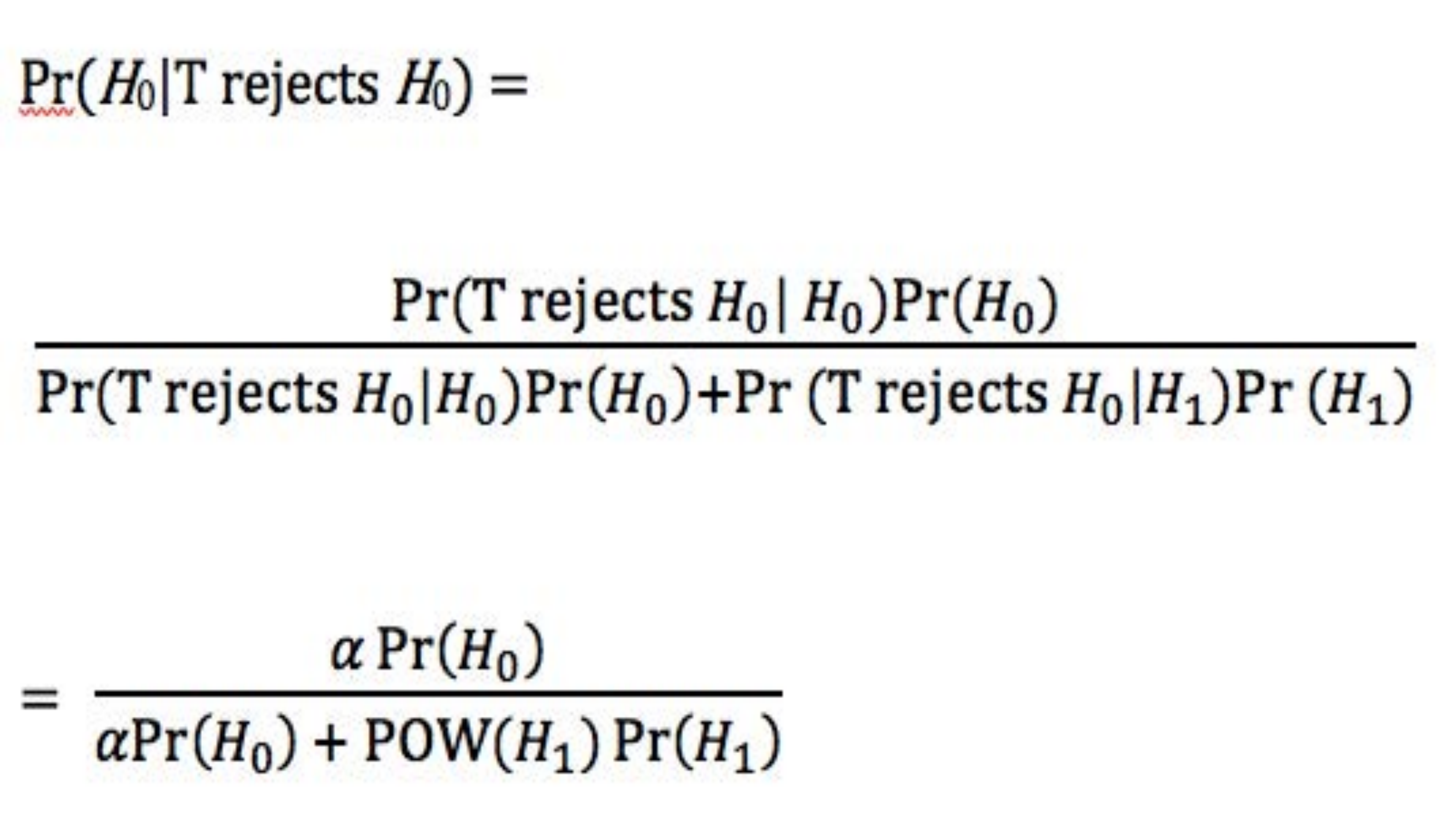
$\Pr(H_0|\text{Test T rejects } H_0)$ is not a Type I error probability.

Transposes conditional-but that's not all

Combines crude performance with a probabilist assignment

OK in certain screening contexts (genomics)

FFR: False Finding Rate



$$\Pr(H_0 | \text{T rejects } H_0) =$$

$$\frac{\Pr(\text{T rejects } H_0 \,|\, H_0)\Pr(H_0)}{\Pr(\text{T rejects } H_0 | H_0)\Pr(H_0) + \Pr\,(\text{T rejects } H_0 | H_1)\Pr\,(H_1)}$$

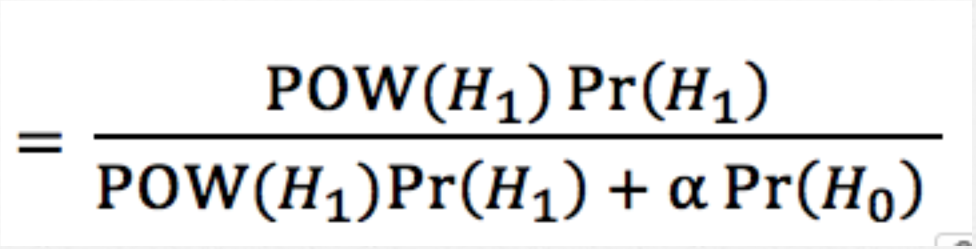
$$= \frac{\alpha \Pr(H_0)}{\alpha \Pr(H_0) + \text{POW}(H_1)\Pr(H_1)}$$

α = 0.05 and (1 – β) = .8, FFR = 0.36, the PPV = .64

PPV

- Complement of FFR is the positive predictive value PPV

$\Pr(H_1 | \text{Test T rejects } H_0)$



$$= \frac{\mathrm{POW}(H_1)\,\Pr(H_1)}{\mathrm{POW}(H_1)\Pr(H_1) + \alpha\Pr(H_0)}$$

What's Pr(H_1) (i.e., *Prev(H_1)*)?

"Proportion of experiments we do over a lifetime in which there is a real effect" (Colquhoun 2014, p. 9)

Proportion of true relationships among those tested in a field. Ioannidis (p. 0696)

Prevalence of GTR hypotheses in 1919

Hypotheses can be individuated in many ways

Probabilistic instantiation fallacy

Even if the prevalence of true effects in the urn is .1 does not follow that a specific hypothesis–say, the GTR deflection effect is 1.75–gets a probability of .1 of being true, for a frequentist

Non-Exhaustive Hypotheses

H_0: 0 effect (μ = 0),

H_1: the alternative vs which the test has power (1 – β).

Non-exhaustive, yet the prior is used up

Is the PPV (complement of the FFR) computation *relevant* to what working scientists want to assess?

***Crud Factor*.** In many fields of social science it’s thought nearly everything is related to everything: “all nulls false”.

These relationships are not, I repeat, Type I errors. They are facts about the world, and with N – 57,000 they are pretty stable. (Meehl, 1990, p. 206).

Will we be better able to replicate results in a field with a high crud factor?

- By contrast: Even in a low prevalence situation, scientists who go beyond the *one* P-value, develop theories, triangulate with other measures have a good warrant for taking the effect as real (stage (i) of GTR).

- *Severe error probing is what's doing the work, not prevalence.*

The Diagnostic Screening model says stay safe

> "Large-scale evidence should be targeted for research questions where the pre-study probability is already considerably high, so that a significant research finding will lead to a post-test probability that would be considered quite definitive" (Ioannidis, 2005, p. 0700).

It was the novelty of the Einstein deflection effect that gave it high corroboration when found.

Upshot: In my problem I need to scrutinize what is warranted to infer – *normative*

Methods must be:

- able to block inferences that violate minimal severity,
- directly altered by biasing selection effects (e.g., post hoc subgroups, outcome-switching etc.),
- able to falsify claims statistically,
- able to test statistical model assumptions.

The Latest ASA Declarations

(Wasserstein et al 2019)

In a purported attempt to avoid abuses of tests:

“’Statistically significant’– don’t say it and don’t use it”

“Don’t conclude anything about scientific or practical importance based on statistical significance (or lack thereof)”

“No p-value can reveal the plausibility, presence, truth, or importance of an association or effect.”

“A declaration of statistical significance is the antithesis of thoughtfulness: … it ignores what previous studies have contributed to our knowledge.”

(1) *Why use a tool that infers* from a single small P-value (that pertains to a *statistical* hypothesis H_0 *) to a research claim H*?*

(2) Why use an incompatible hybrid (of Fisher & N-P)?

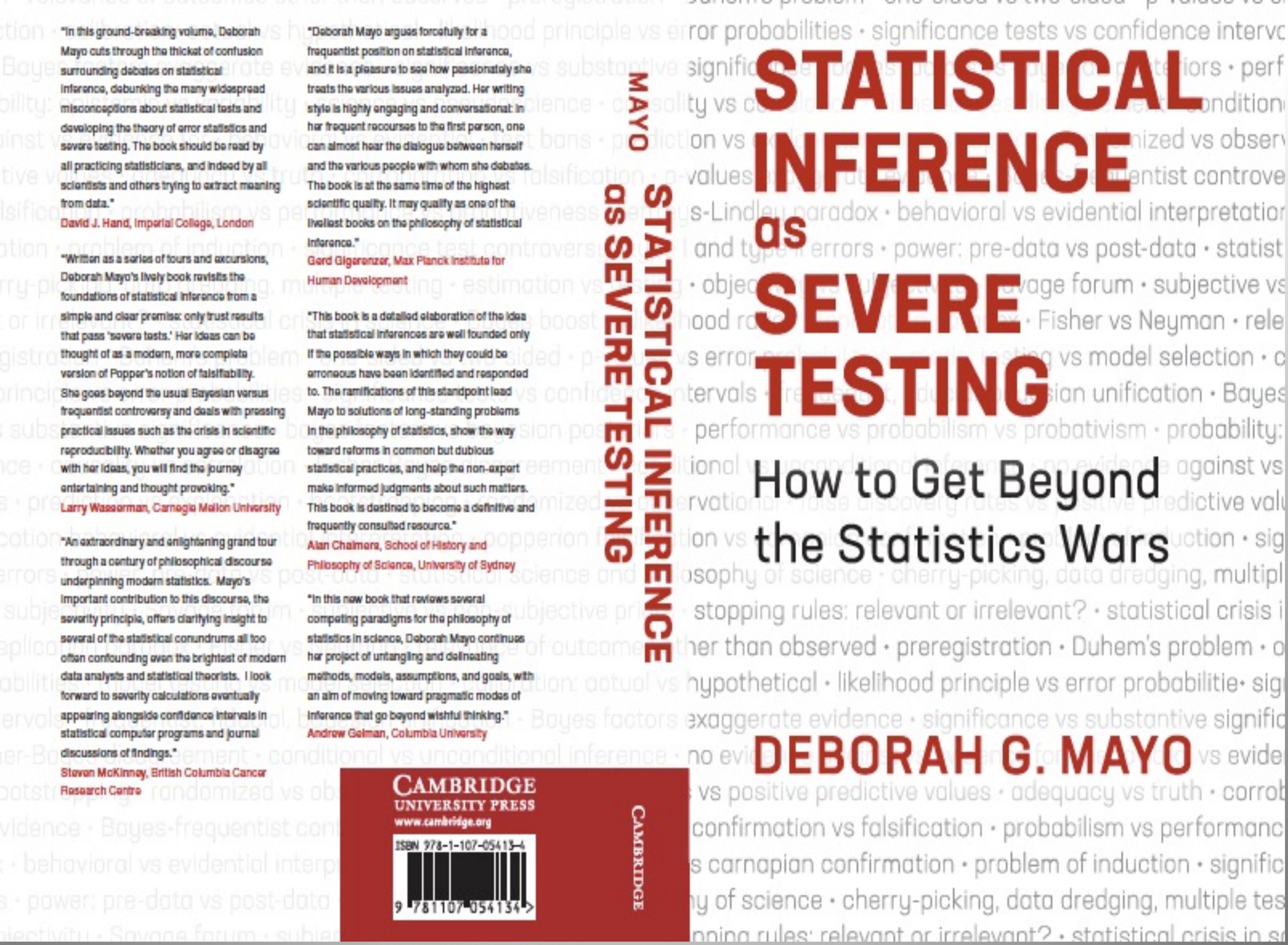
(3) Why apply a method *that uses error probabilities*, *the sampling plan*, researcher "*intentions*"? You should *condition* on the data

(4) Why use methods that exaggerate *evidence* against a null hypothesis?

(5) Why do you use a method that presupposes the underlying statistical model?

(6) Why use a measure that doesn't report effect sizes?

(7) Why do you use a method that doesn't provide posterior probabilities?



"In this ground-breaking volume, Deborah Mayo cuts through the thicket of confusion surrounding debates on statistical inference, debunking the many widespread misconceptions about statistical tests and developing the theory of error statistics and severe testing. The book should be read by all practicing statisticians, and indeed by all scientists and others trying to extract meaning from data."
David J. Hand, Imperial College, London

"Written as a series of tours and excursions, Deborah Mayo's lively book revisits the foundations of statistical inference from a simple and clear premise: only trust results that pass 'severe tests.' Her ideas can be thought of as a modern, more complete version of Popper's notion of falsifiability. She goes beyond the usual Bayesian versus frequentist controversy and deals with pressing practical issues such as the crisis in scientific reproducibility. Whether you agree or disagree with her ideas, you will find the journey entertaining and thought provoking."
Larry Wasserman, Carnegie Mellon University

"An extraordinary and enlightening grand tour through a century of philosophical discourse underpinning modern statistics. Mayo's important contribution to this discourse, the severity principle, offers clarifying insight to several of the statistical conundrums all too often confounding even the brightest of modern data analysts and statistical theorists. I look forward to severity calculations eventually appearing alongside confidence intervals in statistical computer programs and journal discussions of findings."
Steven McKinney, British Columbia Cancer Research Centre

"Deborah Mayo argues forcefully for a frequentist position on statistical inference, and it is a pleasure to see how passionately she treats the various issues analyzed. Her writing style is highly engaging and conversational: in her frequent recourses to the first person, one can almost hear the dialogue between herself and the various people with whom she debates. The book is at the same time of the highest scientific quality. It may qualify as one of the liveliest books on the philosophy of statistical inference."
Gerd Gigerenzer, Max Planck Institute for Human Development

"This book is a detailed elaboration of the idea that statistical inferences are well founded only if the possible ways in which they could be erroneous have been identified and responded to. The ramifications of this standpoint lead Mayo to solutions of long-standing problems in the philosophy of statistics, show the way toward reforms in common but dubious statistical practices, and help the non-expert make informed judgments about such matters. This book is destined to become a definitive and frequently consulted resource."
Alan Chalmers, School of History and Philosophy of Science, University of Sydney

"In this new book that reviews several competing paradigms for the philosophy of statistics in science, Deborah Mayo continues her project of untangling and delineating methods, models, assumptions, and goals, with an aim of moving toward pragmatic modes of inference that go beyond wishful thinking."
Andrew Gelman, Columbia University

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-1-107-05413-4

MAYO STATISTICAL INFERENCE as SEVERE TESTING CAMBRIDGE

STATISTICAL INFERENCE as SEVERE TESTING

How to Get Beyond the Statistics Wars

DEBORAH G. MAYO

Severity for Test T+: SEV(T+, *$d(x_0)$, claim C)*

Normal testing: H_0: $\mu \le \mu_0$ vs. H_1: $\mu > \mu_0$ known σ; discrepancy parameter γ; $\mu_1 = \mu_0 + \gamma$; $d_0 = d(\mathbf{x}_0)$ *(observed value of test statistic)* $\sqrt{n}(M - \mu_0)/\sigma$

***SIR*: (Severity Interpretation with low P-values)**

- (a): (*high*): If there’s a very low probability that so large a d_0 would have resulted, if μ were no greater than μ_1, then d_0 it indicates $\mu > \mu_1$: SEV($\mu > \mu_1$) is high.

- (b): (*low*) If there is a fairly high probability that d_0 would have been larger than it is, even if $\mu = \mu_1$, then d_0 is *not* a good indication $\mu > \mu_1$: SEV($\mu > \mu_1$) is low.

***SIN*: (Severity Interpretation for Negative results. Moderate P-values)**

- (a): *(high)* If there is a very *high* probability that d_0 would have been larger than it is, were $\mu > \mu_1$, then $\mu \leq \mu_1$ passes the test with *high* severity: SEV($\mu \leq \mu_1$) is high.

- (b): (*low*) If there is a *low* probability that d_0 would have been larger than it is, even if $\mu > \mu_1$, then $\mu \leq \mu_1$ passes with *low* severity: SEV($\mu \leq \mu_1$) is low.