Excursion 3 Tour III: Biasing Selection Effects: Biggest source of handwringing

.

It is easy to data dredge impressive-looking effects that are spurious

(minimal) Severity Requirement:

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

("too cheap to be worth having" Popper 1983)

Severe Testing Account of Evidence

- You have evidence for a claim if it's subjected to and passes a test that probably would have found it false, if it is: this probability is the severity with which it has passed

With this construal of probability in inference, you may be able to have your cake and eat it too

In approaching an account that uses probability to measure degrees of belief or support, a severe tester still wants to know if it's arriving at strong belief for false claims with high probability

In statistical settings, error probabilities may give a direct way to measure *biasing selection effects* (P-hacking, data-dredging, cherry picking, etc.)

Data Dredging (Torturing): Hunting for Subgroups in RCTs

Tour starts with an imaginary case: the Drug CEO:

- No statistically significant benefit on the primary endpoint (improved lung function)
- Nor on any 10 secondary endpoints
- Ransacks the unblinded data for a subgroup where those on the drug did better.
- Reports it as a statistically significant result from a double-blind study

The method has a high probability of reporting drug benefit (in some subgroup or other), even if none exists—illicit P-value.

But some cases of multiplicity & data dredging satisfy severity

Searching a full database for a DNA match with a criminal's DNA:

- The probability is high of a mismatch with person i, if i were not the criminal;
- So, the match is good evidence that i is the criminal.
- A non-match virtually excludes the person thereby strengthening the inference.

How to distinguish? It's the severity or lack of it that distinguishes if a data dredged claim is warranted

The data dredging at issue involves some element of double-counting

When data-driven discoveries are tested on new data, it's not data dredging

- The FDA gave the drug CEO funds to test his 'exploratory' hypothesis

- When the new study was stopped for futility (2009), FDA said: you're going to jail (for the misleading press report)!

(reached Supreme Court, 2013, Mayo 2020)

- Even if the follow-up had succeeded, the initial data poorly tested the dredged claim

Ruling out chance vs explaining a known effect

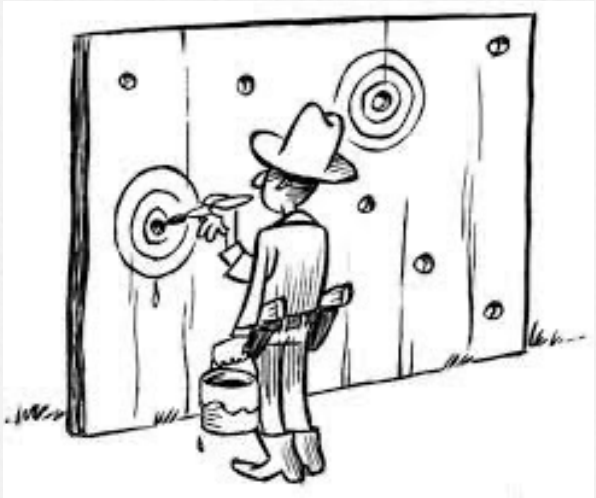
The dredged hypotheses need not be prespecified to be kosher

- The same data were used to arrive at and test the source of a set of blurred 1919 eclipse data (mirror distortion by the sun's heat)
- Nor is it a problem that the same data are used to test multiple claims using statistical significance tests (Fisher recommended)

The problem is selection

The problem is when the results (or hypotheses) are related in such a way that the tester ensures the only ones to emerge or be reported are in sync with the claim *C* at issue, (even if false).

The successes are due to the biasing selection effects, not C's truth



The Severity Requirement with Data Dredging and Multiplicity

- has to be assessed according to the type of error that *C* claims is well ruled out by the data ***x.***

Biasing Selection Effects:

When data or hypotheses are selected, generated or interpreted in such a way as to fail the severity requirement

(includes inability to assess severity even approximately)

It's easy to lie with biasing selection effects

"We're more fooled by noise than ever before, and it's because of a nasty phenomenon called 'big data'. With big data, researchers have brought cherry-picking to an industrial level" (Taleb 2013).

Selection effects alter a method's error probabilities and yet a fundamental battle in the statistics wars revolves around their relevance

2016 ASA Guide: Principle 4

"Proper inference requires full reporting and transparency. P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain *p*-values (typically those passing a significance threshold) renders the reported *p*-values essentially uninterpretable."
(Wasserstein and Lazar 2016)

Capitalizing on Chance (nominal vs. actual P-values)

*Suppose that twenty sets of differences have been examined, that one difference seems large enough to test and that this difference turns out to be 'significant at the 5 percent level.' ….**The actual level of significance is not 5 percent, but 64 percent!*** (Selvin 1970, 104)

Pr(no successes in 20 ind trials) = $(.95)^{20}$

Bonferroni adjustment: multiply P-value by n

SIST p. 274

Spurious P-Value

The hunter reports: Such results *would be difficult to achieve* under the assumption of H_0

When in fact such results are easy to get under the assumption of H_0

- There are many more ways to be wrong with hunting (different sample space)
- Need to adjust P-values
- or at least report the multiple testing (e.g., the Bonferroni adjustment, multiply the P-value by N, the number of tests)

Some accounts of evidence object:

"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)

(Co-director, with Ioannidis, the Meta-Research Innovation Center at Stanford)

So far

I) Multiplicity and data dredging can alter error probabilities

a. Agreed, but appropriate data-dredging can satisfy relevant error probabilities

Next part:

b. Agreed, but altering error probabilities don't matter for evidence

Likelihood Principle (LP)

If the statistical model is correct, then all the information from the data (for inference about a parameter in that model) comes through the likelihood ratio.

Held by Bayesians and Likelihoodists (qualifications to arise)

First 2 rounds of the case….

Dr. Paul Hack, CEO of Best Drug Co., is accused of issuing a report on the benefits of drug X that exploits a smattering of questionable research practices (QRPs): It ignores multiple testing, uses data-dependent hypotheses, and is oblivious to a variety of selection effects. What happens shines a bright spotlight on a mix of statistical philosophy and evidence. The case is fictional; any resemblance to an actual case is coincidental.

Round 1 The prosecution marshals their case, calling on a leader of an error statistical tribe who is also a scientist at Best: *Confronted with a lack of statistically significant results on any of 10 different prespecified endpoints (in randomized trials on Drug X), Dr. Hack proceeded to engage in a post-data dredging expedition until he unearthed a subgroup wherein a nominally statistically significant benefit [B] was found. That alone was the basis for a report to share-holders and doctors that Drug X shows impressive benefit on factor B.* Colleagues called up by the prosecution revealed further details: *Dr. Hack had ordered his chief data analyst to "shred and dice the data into tiny julienne slices until he got some positive results" sounding like the adman for*

Round 2 Next to be heard from are defenders of Dr. Hack: *There's no need to adjust for post-hoc data dredging, the fact that significance tests require such adjustments is actually one of their big problems. What difference does it make if Dr. Hack intended to keep trying and trying again until he found something? Intentions are irrelevant to the import of data.* Others insist that: *the position on cherry picking is open to debate, depending on one's philosophy of evidence. For the courts to take sides would set an ominous precedent.* They cite upstanding statisticians who can attest to the irrelevance of such considerations.

Round 3 A second wave of Hack's defenders (which could be the same as in Round 2) pile on, with a list of reasons for *P*-phobia: *Significance levels exaggerate evidence, force us into dichotomous thinking, are sensitive to sample size, aren't measures of evidence because they aren't comparative reports, and violate the likelihood principle.* Even expert prosecutors, they claim, construe a *P*-value as the probability the results are due to chance, which is to commit the prosecutor's fallacy (misinterpreting *P*-values as posterior probabilities), so they are themselves confused.

Dr. Hack's lawyer jumps at the opening before him: *You see there is disagreement among scientists, at a basic philosophical level. To hold my client accountable would be no different than banning free and open discussion of rival interpretations of data amongst scientists.*

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

SIST 283-4

Wagenmakers looks askance at adjusting for selection effect:

> "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)

Instead of saying they ought to adjust, Wagenmakers dismisses a concern with imaginary data (SIST 284)

Exhibit (x): Bem's "Feeling the future" 2011: ESP?

- Daryl Bem (2011): subjects do better than chance at predicting the (erotic) picture shown in the future
- Some locate the start of the replication crisis with Bem
- Bem admits data dredging
- Bayesians say this shows the need to replace P-values with a default Bayesian prior to (a point) null hypothesis
- (& consider a large effect for the alternative)

It Relinquishes their strongest criticism

- Bayesians can block by giving H_0 (no effect) a high prior probability in a Bayesian analysis

- The researcher deserving criticism deflects this saying: you can always counter an effect this way

Bem's response: Jeffreys-Lindley paradox

"Whenever the null hypothesis is sharply defined but the prior distribution on the alternative hypothesis is diffused over a wide range of values, as it is [here] it boosts the probability that any observed data will be higher under the null hypothesis than under the ǎlternative."

Bayes-Fisher disagreement (meeting 11)

Bayes/Fisher Disagreement: Spike and Smear

- A point null hypothesis, a lump of prior probability on H_0 or a tiny area around it $[X_i \sim N(\mu, \sigma^2)]$

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

- Depending on how you spike and how you smear, an α significant result can even correspond to

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

- But $\Pr(H_0|\boldsymbol{x})$ can also agree with the small α

A recent paper by Bickel

•Bickel, D. R. (2021): If P-values disagree with the posterior on H_0, your prior fails a model check.

"Null hypothesis significance testing defended and calibrated by Bayesian model checking." *The American Statistician*, *75*(3), 249–255.

Can be right for the wrong reason:

"Bayesians can easily discount my statistically significant result this way".

•Even if it's correct to reject the data dredged claim—as in this case--, it can be right for the wrong reason:

•Put the blame where it belongs.

Return to our Court case–the real one: “P-values on Trial”*

In 2009, Scott Harkonen (CEO of InterMune) found guilty of wire fraud for a misleading press report on results of a drug Actimmune in 2002.

In SIST I don’t name names, but it came up again soon after it was published

Many rounds of appeals failed

2013 amicus brief in support of Harkonen that "Multiple Testing Does Not Undermine the Meaning of P-Values" (Rothman et al., p. 19)--in tension with Principle 4 of the ASA Guide.

Nevertheless, in 2018, Harkonen takes the 2016 ASA Guide to show his "actual innocence"

- In 2018, Harkonen & defenders argued the 2016 ASA Statement provides “compelling new evidence that the scientific theory upon which petitioner’s conviction was based [that of statistical significance testing] is demonstrably false” (Goodman Amicus 2018, p. 3).
- They claim “the conclusions from the ASA Principles are the opposite of the “ FDA’s conclusion that his construal of the data was misleading (goes to SCOTUS)

*Mayo D. “P-Values on Trial: (2020).

The 2016 ASA Guide doesn't show statistical significance testing "is demonstrably false,"

but it might be seen to communicate a message that is in tension with itself on one of the most important issues of statistical inference."

Despite principle 4

The ASA 2016 Guide's Six Principles

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Principle 4 goes on longer than all the others …

"Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all *p*-values computed. Valid scientific conclusions based on *p*-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including *p*-values) were selected for reporting." (ASA I pp. 131-132)

Why do I say

“it might be seen to communicate a message that is in tension with itself on one of the most important issues of statistical inference."

Immediately after the principles in the Guide: "*Other Approaches*".

"In view of the prevalent misuses of and misconceptions concerning *p*-values, some statisticians prefer to supplement or even replace *p*-values with other approaches [including] estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors". (p. 132)

We know some of these alternatives follow the LP

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

However…

- The same data-dredged hypothesis can occur in a Bayes factor
- But your grounds for criticism is gone

The data-dredged hypothesis

"H_1^{PD}: Actimmune increases survival in IPF patients in the post data subgroup:

$$\Pr(\boldsymbol{x}|H_1^{PD})/\Pr(\boldsymbol{x}|H_0^{PD}).$$

The alternative H_1^{PD} would be comparatively better supported (for the likelihoodist) or more probable (for the Bayes factor theorist).

It has been deliberately selected for this purpose

Harkonen's defenders claim:
his conviction is "premised on the fundamentally flawed view that a non-significant *p*-value, by itself, falsifies a claim that a relationship exists" (Goodman, 2018, p. 10).

Round 3: P-values are flawed

The government's position against Harkonen merely denies that H_1^{PD} was well-tested by the same data that was used to find the subgroup--very different from denying the post-data claim altogether.
The latter would be to assert H_1^{PD} is false and H_0^{PD} true

Most criticism is just logic .

Does principle 4 hold for other approaches?

- An 11th hour point of controversy: whether to retain “full reporting and transparency” (principle 4) for all methods
- Or should it apply only to “p-values and related statistics”

“Either the other approaches require the same treatment of multiple testing and post-data subgroups as statistical significance tests or they do not.

If they do, then the basis for criticizing Harkonen remains.

If they do not, then the message from the Guide may well be seen to absolve Harkonen of blame for his interpretation".

- Granted, the fact that someone uses their construal of data when serving as an expert witness doesn’t by itself show anything wrong (ad hominem), but you might want to look more closely.

- “P-values on Trial” refers to the 2016 ASA Statement on P-values
- A more drastic editorial (March 2019) claimed the 2016 report “stopped just short of recommending that declarations of ‘statistical significance’ be abandoned….We take that step here.”
- It was never an ASA document, so most thought it was since Wasserstein is ASA Executive Director

- “If the 2016 guide opens the door to giving data dredgers a free pass [the Wasserstein 2019 editorial] swings the door wide open” (Mayo 2020)

- Most thought it was a continuation since Wasserstein is ASA Executive Director

- While “P-values on trial” was in press, the president of the ASA said no, it never was

The 2019: Don’t say ‘significance’, don’t use P-value thresholds

- Editors of the March 2019 issue TAS "A World Beyond p < 0.05"—Wasserstein, Schirm, Lazar—aver that "declarations of ‘statistical significance’ be abandoned" (p. 2).
- On their view: Prespecified P-value thresholds should never be used in interpreting results.
- it is not just a word ban but a gatekeeper ban

No tests, no falsification

- The “no thresholds” view also blocks common uses of confidence intervals and Bayes factor standards
- If you cannot say about any results, ahead of time, they will not be allowed to count in favor of a claim, then you do not have a test of it
- Don’t confuse having a threshold for a terrible test with using a fixed P-value across all studies in an unthinking manner
- We should reject the latter

“Retiring statistical significance would give bias a free pass".

John Ioannidis (2019)

> "...potential for falsification is a prerequisite for science. Fields that obstinately resist refutation can hide behind the abolition of statistical significance but risk becoming self-ostracized from the remit of science”.

I agree, “P-value Thresholds: Forfeit at Your Peril” (2019)

- To be fair: some claim that by removing P-value thresholds, researchers lose an incentive to data dredge, and otherwise exploit researcher flexibility
- I sy it’s the opposite

- Even without the word significance, eager researchers still can’t take the large (non-significant) P-value to indicate a genuine effect
- *It would be to say*: Even though larger differences would frequently occur by chance variability alone, my data provide evidence they are not due to chance variability
- In short, he would still need to report a reasonably small P-value
- The eager investigator will need to "spin" his results, ransack, data dredge

- In a world without predesignated thresholds, it would be hard to hold him accountable for reporting a nominally small P-value:
- “whether a *p*-value passes any arbitrary threshold should not be considered at all" in interpreting data (Wasserstein et al. 2019, 2)

My view: Reformulate Tests

- I agree with ousting strict binary uses of tests, and recipe-like behavioristic interpretations (NHST—a fallacious animal)
- Instead of a binary cut-off (significant or not) the particular outcome is used to infer discrepancies that are or are not warranted
- Avoids fallacies of significance and nonsignificance, and improves on confidence interval estimation

2022 disclaimer: 2019 editorial not an ASA policy

- Wasserstein et al., (2019) claim the 2016 Guide “stopped just short of recommending that declarations of ‘statistical significance’ be abandoned….We take that step here.”
- It was never an ASA document (though Wasserstein is ASA Executive Director)

New ASA Task Force on Significance Tests and Replication

- to “prepare a …piece reflecting “good statistical practice,” without leaving the impression that *p*-values and hypothesis tests…have no role.” (Karen Kafadar 2019)
- This was December, soon after came the pandemic

"The Statistics Wars and Intellectual Conflicts of Interest"*

Deborah G. Mayo
Virginia Tech
Phil Stat forum
11 January 2022
"Statistical Significance Test Anxiety"

Acknowledgments: Mark Burgman, Jean Miller, David Hand, Nathan Schachtman

Souvenir T: Even Big Data Calls for Theory and Falsification

Historically, epidemiology has focused on minimizing Type II error (missing a relationship in the data), often ignoring multiple testing considerations, while traditional statistical study has focused on minimizing Type I error …When traditional epidemiology met the field of GWAS, a flurry of papers reported findings which eventually became viewed as nonreplicable.

(Lambert and Black 2012, p. 199)

Assignment 3 (from Excursion 4 Tour I, objectivity)

(There will also be a question on power)

1. What is the argument against objectivity based on "dirty hands"? (explain as fully as you can). Should we reject or accept it? Or retain it in part? (222-5)
2. Compare: "how well have you probed" and "how strongly do/should you believe it? In explaining these, bring out some central linguistic ambiguities. 226)
3. How might you respond to the argument to "embrace your subjectivity"? Explain the argument. Do you agree with the position in this section of SIST? (228)
4. What are "objective" (default, non-subjective) Bayesians (230-1 and elsewhere in SIST)? Why are there no "uninformative" priors? Why does J. Berger argue for O-Bayesianism? Why does Kadane argue against it? (230-231)

Excursion 4 Tour I Objectivity:

6. Evaluate the argument

(i) that prior probabilities let us be explicit about bias (232-3)

(ii) that prior probabilities allow combining background information

(See also “grace and amen Bayesians” (413-415)

7. Objectivity in epistemology; 235-6. Evaluate the links between objectivity and

(i) Externalism

(ii) Diversity of knowers

8. In the Farewell Keepsake (436-) points 1-8 are often taken as central criticisms of statistical significance tests: First explain, and then critically appraise, two of them.