Class 9B Tour III (p. 267)

- The biggest source of handwringing about statistical inference boils down to the fact it has become very easy to infer claims that have not been subjected to stringent tests. Sifting through reams of data makes it easy to find impressive-looking associations, even if they are spurious. (Wasserstein and Lazar 2016); hereafter, ASA Guide. Principle 4 of the Guide asserts that:

“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.”

The 2016 continues…

"Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. 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)

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

The Tour begins with an imaginary court case

YOGA EXERCISES FOR RESEARCHERS
PUBLISH or PERISH
DATA
p
© Dario Battisti 2013

Hunting for significance (nominal vs. actual)

> *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)

(Morrison & Henkel's *Significance Test Controversy* 1970!)

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 (a conservative way is 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)

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)

Error control is lost (SIST p. 270)

“[I]f the sampling plan is ignored, the researcher is able to always reject the null hypothesis, even if it is true. This example is sometimes used to argue that any statistical framework should somehow take the sampling plan into account. ..This feeling is, however, contradicted by a mathematical analysis. (E-J Wagenmakers, 2007, 785)

But the “proof” assumes the likelihood principle (LP) by which error probabilities drop out. (Edwards, Lindman, and Savage 1963, Excursion 1 Tour II optional stopping)

A mathematical ‘principle of rationality’ (LP) gets weight over what appears to be common sense

At odds with key way to advance replication:

21 Word Solution

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

- Replication researchers find flexibility with data-dredging and stopping rules major source of failed-replication (the "forking paths", Gelman and Loken 2014))

Statutes are at one in condemning data dredging

Reference Manual on Scientific Evidence for lawyers

How many tests have been done? 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….

Nevertheless, statutes can be changed if their rationale is overturned. The issue is not settled. (Freedman and Kaye 2011)

Many “reforms” offered as alternative to significance tests follow the 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; authors of the *Likelihood Principle*)

- No wonder reformers talk past each other

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

Relinquishing their strongest criticism

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

- The researcher deserving criticism deflects this saying: you can always counter an effect this way (the defense violates our minimal principle of evidence)

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
- Bayesian critics resort to a default Bayesian prior to (a point) null hypothesis

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)

Bem’s response

“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 alternative.

This is known as the Lindley-Jeffreys paradox*: A frequentist [can always] be contradicted by a …Bayesian analysis that concludes that the same data are more likely under the null.” (Bem et al. 2011, 717)

*Bayes-Fisher disagreement (Day #10)

Bayes (Jeffreys)/Fisher disagreement (“spike and smear”)

- 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). $X_i \sim N(\mu, \sigma^2)$

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

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

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

Instead of getting flogged, Bem points to the flexibility of getting a Bayes factor in favor of the null hypothesis.

P-Values Can’t be Trusted Except When Used to Argue That P-values Can’t be Trusted!

Severity can be *improved* by searching (p. 281)

Searching for a DNA match with a criminal’s DNA: The probability is high that we would not obtain a match with person i, if i were not the criminal;

So, finding the match is good evidence that i is the criminal.

(wrong to say a frequentist would have to be penalized for searching here, Dawid)

Quite a lot of background knowledge, numerous assumptions, not knowledge-free or model free

One mistake–spurious effect–is taken care of.

- Deflection effect was a constraint on finding its cause
- Searching for an an animal to manifest (known) birth defects in thalidomide

(Of course, if you're going to test on new data, it's not double-counting, but merely exploratory)

One mistake–spurious effect–is taken care of.

- Deflection effect was a constraint on finding its cause
- Searching for an an animal to manifest (known) birth defects in thalidomide

(Of course, if you're going to test on new data, it's not double-counting, but merely exploratory)

Return to our Court case–the real one

Not only did Hack’s defenders argue,

- everyone data dredges and puts “spin” on their non-significant results
- There’s disagreement (philosophical) about whether to adjust for multiple testing
- there’s controversy about P-values, they aren’t posteriors, aren’t comparative, aren’t measures of effect size, are often misinterpreted

There’s also a selective appeal to best practice guides on selective reporting, such as the 2016 ASA guide (ASA 1)

His defenders argue that “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)

American Statistical Association 2016 (statement on P-values):

- *Principle 4: P*-values and related analyses should not be reported selectively. …Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis.
- 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… such as likelihood ratios or Bayes Factors;(p. 132)

We know some of these 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)

- The same data-dredged hypothesis can occur in a Bayes factor, except now your basis for criticism is lost

Does principle 4 hold for other approaches?

- "The direct grounds to criticize inferences as flouting error statistical control is lost"
- An 11th hour point of controversy: whether to retain "full reporting and transparency" (principle 4) for all methods
- I was a "philosophical observer"
- Or should it apply only to "p-values and related statistics"

Turn to "P-vaues on Trial"