Replication Crises and Hidden Controversies in Phil Stat

OVERVIEW

APRIL 3, 2019 6334/6614

- High-profile failures of replication have brought forth reams of reforms
- How should the integrity of science be restored?
- Experts do not agree.
- We need to pull back the curtain on why.

Hidden Controversies

1. **The Statistics Wars**: Age-old abuses, fallacies of statistics (Bayesian-frequentist, Fisher-N-P debates) simmer below todays debates; reformulations of frequentist methods are ignored
2. **Replication Paradoxes**: While a major source of handwringing stem from biasing selection effects—cherry picking, data dredging, trying and trying again, some reforms and preferred alternatives conflict with the needed error control.
3. **Underlying assumptions** and their violations

- A) statistical model assumptions
- B) links from experiments, measurements & statistical inferences to substantive questions

Significance Tests

- The most used methods are most criticized
- Statistical significance tests are a small part of a rich set of:

  > "techniques for systematically appraising and bounding the probabilities … of seriously misleading interpretations of data" (Birnbaum 1970, 1033)

- These I call *error statistical methods* (or sampling theory).

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

…we find a function $T = t(\mathbf{y})$ of the data, to be called the **test statistic**, such that

- the larger the value of T the more inconsistent are the data with H_0;
- The random variable $T = t(\mathbf{Y})$ has a (numerically) known probability distribution when H_0 is true.

…the p-value corresponding to any t_{0bs} as

$$p = p(t) = Pr(T \geq t_{0bs}; H_0)"$$

(Mayo and Cox 2006, 81)

Testing Reasoning

- 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 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 t_{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****

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

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

As opposed to NHST

Get beyond N-P/ Fisher incompatibilist caricature

- Beyond “inconsistent hybrid” (Gigerenzer 2004, 590): Fisher–inferential; N-P–long run performance
- They both use a method’s sampling distribution to assess and control error probabilities
- Results in P-value users robbed of features from N-P tests they need (power)
- Wrongly supposes N-P testers have fixed error probabilities (no balance), & don’t report P-value

Replication Paradox
(for Significance Test Critics)

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

You: Why do they find it so difficult to replicate the small P-values others found?

Is it easy or is it hard*?*

Both Fisher and N-P: it's easy to lie with statistics by selective reporting

- Sufficient finagling—**cherry-picking, P-hacking, significance seeking, multiple testing, look elsewhere**—may practically guarantee a preferred claim *H* gets support, even if it's unwarranted by evidence

Severity Requirement:

Everyone agrees: If the test had little or no capability of finding flaws with *H* (even if *H is* incorrect), then agreement between data x_0 and *H* provides poor (or no) evidence for *H*

("too cheap to be worth having" Popper 1983, 30)

- Such a test fails a *minimal requirement* for a stringent or severe test
- My account: severe testing based on error statistics (requires reinterpreting tests)

This alters the role of probability: typically just 2

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

(e.g., Bayesian, likelihoodist)—with regard for inner coherency

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

- Problems with selective reporting, cherry picking, stopping when the data look good, P-hacking, are not problems about long-runs—
- It's that *we cannot say the case at hand* has done a good job of avoiding the sources of misinterpreting data

Key to revising the role of error probabilities

A claim *C* is not warranted _____

- ***Probabilism****:* unless *C* is true or probable (gets a probability boost, is 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*

Biasing selection effects:

One function of severity is to identify problematic selection effects (not all are)

- ***Biasing selection effects***: when data or hypotheses are selected or generated (or a test criterion is specified), in such a way that **the minimal severity requirement is violated, seriously altered or incapable of being assessed**
- Compare *to explaining a known effect* (SIST p. 281), as in DNA matching

Nominal vs. actual Significance levels

SIST p. 275

*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)

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

Spurious P-Value

You report: Such results *would be difficult to achieve* under the assumption of H_0

When in fact such results are common under the assumption of H_0

There are many more ways you can be wrong with hunting (different sample space)

(Formally):

- You say Pr(P-value ≤ P_{obs}; H_0) ~ P_{obs} = small
- But in fact Pr(P-value ≤ P_{obs}; H_0) = high

Scapegoating

- Nowadays, we're likely to see the tests blamed
- My view: Tests don't kill inferences, people do
- Even worse are those statistical accounts where the abuse vanishes!

Some say taking account of biasing selection effects "defies scientific sense" SIST P. 269

Two problems that plague frequentist inference: multiple comparisons and multiple looks, or, as they are more commonly called*, 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 his credit, he's open about this; heads the Meta-Research Innovation Center at Stanford)

Likelihood Principle (LP)

The vanishing act links to a pivotal disagreement in the philosophy of statistics battles

In probabilisms, 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

Jimmy Savage on the LP:

"According to Bayes' theorem,…. if ***y*** is the datum of some other experiment, and ***if it happens that P(x|μ) and P(y|μ) are proportional functions of μ (that is, constant multiples of each other), then each of the two data x and y have exactly the same thing to say about the values of μ…"*** (Savage 1962, p. 17)

All error probabilities violate the LP
(even without selection effects):

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)

The LP implies…the irrelevance of predesignation, of whether a hypothesis was thought of before hand or was introduced to explain known effects
(Rosenkrantz 1977, 122) SIST P. 269

Optional Stopping:

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

$X_i \sim N(\mu, \sigma^2)$, 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 0.05 level

i.e., keep sampling until $M \geq 1.96\ \sigma/\sqrt{n}$

- *Trying and trying again*: Having failed to rack up a $1.96\ \sigma/\sqrt{n}$ difference after 10 trials, go to 20, 30 and so on until obtaining a $1.96\ \sigma$ difference

Nominal vs. *Actual* significance levels again:

- 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)

Probabilist side:

Jmmie Savage (1961, 583) declared:

“optional stopping is no sin”

so the problem must be with significance levels (because they pick up on it).

Performance *(and probative testing)* ***side:***

Peter Armitage:

“thou shalt be misled”

if thou dost not know the person tried and tried again. (1962, 72)

- Equivalently, you can ensure that 0 is excluded from a confidence interval (or credibility interval) even if true (Berger and Wolpert 1988)
- Likewise the same p-hacked hypothesis can occur in Bayes factors, credibility intervals, likelihood ratios

With One Big Difference:

- The direct grounds to criticize inferences as flouting error statistical control is lost
- They condition on the actual data,
- Error probabilities take into account other outcomes that could have occurred but did not (sampling distribution)

What Counts as Cheating? 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. Some people feel that 'optional stopping' amounts to cheating…. This feeling is, however, contradicted by a mathematical analysis. (Eric-Jan Wagenmakers, 2007, 785)

But the "proof" assumes the likelihood principle (LP) by which error probabilities drop out. (Edwards, Lindman, and Savage 1963)

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

"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).

- Or report how their stopping plan alters relevant error probabilities.

- ***Critic***: It's too easy to satisfy standard significance thresholds
- ***You***: Why do replicationists find it so hard to achieve significance thresholds (with preregistration)?
- ***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.
- **Critic**: Actually "reforms" recommend methods where the need to alter P-values due to data dredging vanishes

Error Probabilities Violate the LP

> [I]t seems very strange that a frequentist could not analyze a given set of data, such as $(x_1, \ldots, x_n)$ if the stopping rule is not given . . . [D]ata should be able to speak for itself. (Berger and Wolpert 1988, p. 78)

Inference by Bayes' Theorem satisfies this intuition, which sounds appealing; but for our severe tester, data no more speak for themselves in the case of stopping rules than with cherry picking, hunting for significance, and the like.

SIST p.49

- *Default Bayesian Reforms are touted as free of selection effects*

"…*Bayes factors can be used in the complete absence of a sampling plan…*"

(Bayarri, Benjamin, Berger, Sellke 2016, 100)

Colloquium announcement

"Reproducibility of science: p-values, multiple testing and optional stopping"

Presented by
James Berger, Duke University

Thursday, April 4, 2019
313 Seitz Hall
3:30 p.m.

Abstract: Three of the statistical causes for the lack of reproducibility of science will be discussed, along with a suggested cure. The first cause is the common misinterpretation of p-values; the second is the frequent lack of sufficient adjustment for multiple testing; the third is the common ignoring of optional stopping when testing. The suggested cure for all three is to use odds of hypotheses as the basic inference tool. Surprisingly, this can be done in a way that simultaneously accommodates both frequentist and Bayesian reasoning.

Recent push to “redefine statistical significance” is based on the odds ratios/ Bayes Factors (BF) (Benjamin et al 2017)

$$\frac{\Pr(H_0|x)}{\Pr(H_1|x)} = \frac{\mathbf{Pr}(\boldsymbol{x}|\boldsymbol{H_0})\,\Pr(H_0)}{\mathbf{Pr}(\boldsymbol{x}|\boldsymbol{H_1})\Pr(H_1)}$$

Old & new debates: P-values exaggerate based on Bayes Factors or Likelihood Ratios

- Testing of a point null hypothesis, a lump of prior probability given to H_0

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

- The criticism is the posterior probability on H_0 can be larger than the P-value.

- So if you interpret a P-value as a posterior on H_0 (a fallacy), you'd be saying the $Pr(H_0|x)$ is low

Valen Johnson (2013a,b) offers a way to bring the likelihood ratio more into line with what counts as strong evidence, according to a Bayes factor. He begins with a review of "Bayesian hypotheses tests." "The posterior odds between two hypotheses H_1 and H_0 can be expressed as"

$$\frac{\Pr(H_1|\boldsymbol{x})}{\Pr(H_0|\boldsymbol{x})} = \mathrm{BF}_{10}(\boldsymbol{x}) \times \frac{\Pr(H_1)}{\Pr(H_0)}.$$

Like classical statistical hypothesis tests, the tangible consequence of a Bayesian hypothesis test is often the rejection of one hypothesis, say H_0, in favor of the second, say H_1. In a Bayesian test, the null hypothesis is rejected if the posterior probability of H_1 exceeds a certain threshold. (Johnson 2013b, pp. 1720–1)

SIST p. 261

Table 4.2 Upper Bounds on the Comparative Likelihood

P-value: one-sided	z_α	$\text{Lik}(\mu_{\max})/\text{Lik}(\mu_0)$
0.05	1.65	3.87
0.025	1.96	6.84
0.01	2.33	15
0.005	2.58	28
0.0005	3.29	227

Do P-values Exaggerate Evidence?

- *Don't reject the null until the BF is 20 or 28*

- *Significance testers balk at allowing highly significant results to be interpreted as no evidence against the null– or even evidence for it*!

Bad Type II error

- Lump of prior on the null conflicts with the idea that all nil nulls are false)

- P-values can also equal the posterior! (without the spiked prior)–even though they measure very different things.

- When you reject with p = 0.005, you can assign a posterior of .97 to μ ≥ observed sample mean (the max likely value), but there's terrible evidence for this!!!

- This is like using a confidence level of .5
- The severity is .5
- To base the argument on lowering P-values on BFs –if you don't do inference by means of BFs– is not innocuous

- SIST p. 266 Accepting the LP, Johnson's convinced the problem is merely using p-vaues not low enough

BF Advocates Forfeit Their Strongest Criticisms

- Daryl Bem (2011): subjects did better than chance at predicting the picture to be shown in the future, credit to ESP (SIST P. 283)

Admits data dredging

- Keen to show we should trade significance tests for BFs, critics relinquish their strongest criticism
- Resort to a default Bayesian prior to make the null hypothesis comparatively more probable than a chosen alternative

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 analysis that yields strong evidence in support of the experimental hypothesis can be contradicted by a misguided Bayesian analysis that concludes that the same data are more likely under the null.* (Bem et al., 717)

Bayes-Fisher Disagreement or Jeffreys-Lindley "Paradox"

- With-a lump of prior given to the point null, and the rest appropriately spread over the alternative, an α significant result can correspond to

$$\Pr(H_0 \mid x) = (1 - \alpha)!$$ (e.g., 0.95)

"spike and smear"

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

Admittedly, significance tests require supplements to avoid overestimating (or underestimating) effect sizes

That's what severity does

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

Normal testing: H_0: $\mu \leq \mu_0$ 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 $\mathbf{d_0}$ would have been larger than it is, even if $\mu = \mu_1$, then $\mathbf{d_0}$ is *not* a good indication $\mu > \mu_1$: SEV($\mu > \mu_1$) is low.

SIN: (Severity Interpretation for Negative results)

- (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.

Hidden Controversies

1. **The Statistics Wars**: Age-old abuses, fallacies of statistics (Bayesian-frequentist, Fisher-N-P debates) simmer below todays debates; reformulations of frequentist methods are ignored
2. **Replication Paradoxes**: While a major source of handwringing stem from biasing selection effects—cherry picking, data dredging, trying and trying again, some reforms and preferred alternatives conflict with the needed error control.

3. **Underlying assumptions** and their violations

- A) statistical model assumptions
- B) links from experiments, measurements & statistical inferences to substantive questions

- Recognize different roles of probability: **probabilism**, long run **performance**, **probativism** (severe testing)
- Move away from cookbook stat–long deplored–but don't expect agreement on numbers from evaluations of different things (P-values, posteriors)
- Recognize criticisms & reforms are often based on rival underlying philosophies of evidence
- The danger is that some reforms may enable rather than directly reveal illicit inferences due to biasing selection effects
- Worry more about model assumptions, and whether our experiments are teaching us about the phenomenon of interest