Excursion 1: How to Tell What's True about Statistical Inference

Tour I: Beyond Probabilism and Performance

(1.1) If we're to get beyond the statistics wars, we need to understand the arguments behind them. Disagreements about the roles of probability in statistical inference–holdovers from long-standing frequentist-Bayesian battles–still simmer below the surface of current debates on scientific integrity, irreproducibility, and questionable research practices. Striving to restore scientific credibility, researchers, professional societies, and journals are getting serious about methodological reforms. Some–disapproving of cherry picking and advancing preregistration–are welcome. Others might create obstacles to the critical standpoint we seek. Without understanding the assumptions behind proposed reforms, their ramifications for statistical practice remain hidden. (1.2) Rival standards reflect a tension between using probability (i) to constrain a method's ability to avoid erroneously interpreting data (*performance*), and (ii) to assign degrees of support, confirmation, or plausibility to hypotheses (*probabilism*). We set sail with a tool for telling what's true about statistical inference: If little has been done to rule out flaws in taking data as evidence for a claim, then that claim has not passed a *severe test*. From this minimal severe-testing requirement, we develop a statistical philosophy that goes beyond probabilism and performance. (1.3) We survey the current state of play in statistical foundations.

Excursion 1 Tour I: Keywords

Error statistics, severity requirement: weak/strong, probabilism, performance, probativism, statistical inference, argument from coincidence, Life-off (vs drag down), sampling distribution, cherry-picking

Excursion 1 Tour II: Error Probing Tools vs. Logics of Evidence

Core battles revolve around the relevance of a method's error probabilities. What's distinctive about the severe testing account is that it uses error probabilities evidentially: to assess how severely a claim has passed a test. Error control is necessary but not sufficient for severity. Logics of induction focus on the relationships between given data and hypotheses–so outcomes other than the one observed drop out. This is captured in the Likelihood Principle (LP). Tour II takes us to the crux of central wars in relation to the Law of Likelihood (LL) and Bayesian probabilism. (1.4) Hypotheses deliberately designed to accord with the data can result in minimal severity. The likelihoodist tries to oust them via degrees of belief captured in prior probabilities. To the severe tester, such gambits directly alter the evidence by leading to inseverity. (1.5) If a tester tries and tries again until significance is reached–optional stopping–significance will be attained erroneously with high probability. According to the LP, the stopping rule doesn't alter evidence. The irrelevance of optional stopping is an asset for holders of the LP, it's the opposite for a severe tester. The warring sides talk past each other.

Excursion 1 Tour II: Keywords

Statistical significance: nominal vs actual, Law of likelihood, Likelihood principle
Inductive inference, Frequentist/Bayesian, confidence concept, Bayes theorem, default/non-subjective Bayesian, stopping rules/optional stopping, argument from intentions

Excursion 1. Tour I: Notes

Notes from Section 1.1 Severity Requirement: Bad Evidence, No Test (BENT)
1.1 Terms (quick looks, to be crystalized as we journey on)

1. *epistemology*: The general area of philosophy that deals with knowledge, evidence, inference, and rationality.
2. *severity requirement*. In its weakest form it supplies a minimal requirement for evidence: severity requirement (weak): One does not have evidence for a claim if little if anything has been done to rule out ways the claim may be false. If data x agree with a claim C but the method used is practically guaranteed to find such agreement, and had little or no capability of finding flaws with C even if they exist, then we have bad evidence, no test (BENT).
3. *error probabilities of a method*: probabilities it leads or would lead to erroneous interpretations of data. (We will formalize this as we proceed.)
4. *error statistical account*: one that revolves around the control and assessment of a method's error probabilities. An inference is qualified by the error probability of the method that led to it.
   (This replaces common uses of "frequentist" which actually has many other connotations.)
   error statistician: one who uses error statistical methods.
5. *severe testers*: a proper subset of error statisticians: those who use error probabilities to assess and control severity. (They may use them for other purposes as well.)

   The severe tester also requires reporting what has been poorly probed and inseverely tested, Error probabilities can, but don't necessarily, provide assessments of the capability of methods to reveal or avoid mistaken interpretations of data. When they do, they may be used to assess how severely a claim passes a test.

6. *methodology and meta-methodology*: Methods we use to study statistical methods may be called our meta-methodology – it's one level removed.

We can keep to testing language as part of the meta-language we use to talk about formal statistical methods, where the latter include estimation, exploration, prediction, and data analysis.

There's a difference between finding *H* poorly tested by data ***x***, and finding ***x*** renders *H* improbable – in any of the many senses the latter takes on.

> *H*: Isaac knows calculus.
> ***x***: results of a coin flipping experiment

Even taking *H* to be true, data ***x*** has done nothing to probe the ways in which *H* might be false.

R.A. Fisher, against isolated statistically significant results (p. 4).

> [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 1935b/1947, p. 14)

Notes from section 1.2: How to get beyond the stat wars

7. *statistical philosophy* (associated with a statistical methodology): core ideas that direct its principles, methods, and interpretations.
   two main philosophies about the roles of probability in statistical inference : performance (in the long run) and probabilism.

   (i) performance: probability functions to control and assess the relative frequency of erroneous inferences in some long run of applications of the method
   (ii) probabilism: probability functions to assign degrees of belief, support, or plausibility to hypotheses. They may be non-comparative (a posterior probability) or comparative (a likelihood ratio or Bayes Factor)

Severe testing introduces a third:

(iii) probativism: probability functions to assess and control a methods' capability of detecting mistaken inferences, i.e., the severity associated with inferences.

- Performance is a necessary but not a sufficient condition for probativeness.
- Just because an account is touted as having a long-run rationale, it does not mean it lacks a short run rationale, or even one relevant for the particular case at hand.

8. *Severity strong (argument from coincidence):*
   We have evidence for a claim *C* just to the extent it survives a stringent scrutiny. If *C* passes a test that was highly capable of finding flaws or discrepancies from *C*, and yet no or few are found, then the passing result, ***x***, is evidence for *C*.

lift-off vs drag down

(i) lift-off: an overall inference can be more reliable and precise than its premises individually.
(ii) drag-down: An overall inference is only as reliable/precise as is its weakest premise.

- Lift-off is associated with convergent arguments, drag-down with linked arguments.
- Statistics is the science par excellence for demonstrating lift-off!

9. *arguing from error:* there is evidence an error is absent to the extent that a procedure with a high capability of signaling the error, if and only if it is present, nevertheless detects no error.

Bernoulli (coin tossing) model: we record success or failure, assume a fixed probability of success θ on each trial, and that trials are independent. (P-value in the case of the Lady Tasting tea, pp. 16-17).

Error probabilities can be readily invalidated due to how the data (and hypotheses!) are generated or selected for testing.

10. *computed (or nominal) vs actual error probabilities:* You may claim it's very difficult to get such an impressive result due to chance, when in fact it's very easy to do so, with selective reporting (e.g., your computed P-value can be small, but the actual P-value is high.)

Examples: Peirce and Dr. Playfair (a law is inferred even though half of the cases required Playfair to modify the formula after the fact.) Texas marksman (shooting prowess inferred from shooting bullets into the side of a barn, and painting a bull's eye around clusters of bullet holes); Pickrite stock portfolio (Pickrite's effectiveness at stock picking is inferred based on selecting those where the "method" did best)

• We appeal to the same statistical reasoning to show the problematic cases as to show genuine arguments from coincidence.
• A key role for statistical inference is to identify ways to spot egregious deceptions and create strong arguments from coincidence.

11. *Auditing a P-value (one part)* checking if the results due to selective reporting, cherry picking, trying and trying again, or any number of other similar ruses.

- ***Replicability isn't enough:*** Example. observational studies on Hormone Replacement therapy (HRT) reproducibly showed benefits, but had little capacity to unearth biases due to "the healthy women's syndrome."

Souvenir A. [ii] ***Postcard to Send: the 4 fallacies from the opening of 1.1.***

• We should oust mechanical, recipe-like uses of statistical methods long lampooned,
• But simple significance tests have their uses, and shouldn't be ousted simply because some people are liable to violate Fisher's warnings.
• They have the means by which to register formally the fallacies in the postcard list. (Failed statistical assumptions, selection effects alter a test's error probing capacities).
• Don't throw out the error control baby with the bad statistics bathwater.

12. *severity requirement (weak):* If data $\boldsymbol{x}$ agree with a claim C but the method was practically incapable of finding flaws with C even if they exist, then $\boldsymbol{x}$ is poor evidence for C.
severity (strong): If C passes a test that was highly capable of finding flaws or discrepancies from C, and yet no or few are found, then the passing result, $\boldsymbol{x}$, is an indication of, or evidence for, C.

Notes from Section 1.3: **The Current State of Play in Statistical Foundations: A View From a Hot-Air Balloon**

The Bayesian versus frequentist dispute parallels disputes between probabilism and performance.

- Using Bayes' Theorem doesn't make you a Bayesian.
- Subjective Bayesianism and non-subjective (default) Bayesians

13. *Advocates of unifications are keen to show that* (i) default Bayesian methods have good performance in a long series of repetitions – so probabilism may yield

performance; or alternatively, (ii) frequentist quantities are similar to Bayesian ones (at least in certain cases) – so performance may yield probabilist numbers. Why is this not bliss? Why are so many from all sides dissatisfied?

It had long been assumed that only subjective or personalistic Bayesianism had a shot at providing genuine philosophical foundations, but some Bayesians have come to question whether the widespread use of methods under the Bayesian umbrella, however useful, indicates support for subjective Bayesianism as a foundation.

Marriages of Convenience? The current frequentist–Bayesian unifications are often marriages of convenience:

- some are concerned that methodological conflicts are bad for the profession.
- frequentist tribes have not disappeared; scientists still call for error control.
- Frequentists' incentive to marry: Lacking a suitable epistemic interpretation of error probabilities – significance levels, power, and confidence levels – frequentists are constantly put on the defensive.

Eclecticism and Ecumenism. Current-day eclecticisms have a long history – the dabbling in tools from competing statistical tribes has not been thought to pose serious challenges.

Decoupling. On the horizon is the idea that statistical methods may be decoupled from the philosophies in which they are traditionally couched (e.g., Andrew Gelman and Cosma Shalizi 2013). The concept of severe testing is suffciently general to apply to any of the methods now in use.

Why Our Journey? To disentangle the jungle. Being hesitant to reopen wounds from old battles does not heal them. They show up in the current problems of scientific integrity, irreproducibility, questionable research practices, and in the swirl of methodological reforms and guidelines that spin their way down from journals and reports.

How it occurs: the new stat scrutiny (arising from failures of replication) collects from:

- the earlier social science "significance test controversy"
- the traditional frequentist and Bayesian accounts, and corresponding frequentist-Bayesian wars
- the newer Bayesian–frequentist unifications (non-subjective, default Bayesianism)

This jungle has never been disentangled.

Excursion 1 Tour II: Notes

1.4 The Law of Likelihood and Error Statistics: Key Items

Ian Hacking (1965) – the Law of Likelihood.

Law of Likelihood (LL): Data $\boldsymbol{x}$ are better evidence for hypothesis H_1 than for H_0 if $\boldsymbol{x}$ is more probable under H_1 than under H_0.

- Likelihoods are defined and several examples are given.
- Likelihoods of hypotheses should not be confused with their probabilities.
- The Law of Likelihood (LL) is seen to fail the minimal severity requirement – at least if it is taken as an account of inference.

Gellerized hypotheses: maximally fitting, but minimally severely tested, hypotheses.

We observe one outcome, but we can consider that for any outcome, unless it makes H_0 maximally likely, we can find an H_1 that is more likely.

A severity assessment is one level removed: you give me the rule, and I consider its latitude for erroneous outputs.

Sampling distribution.

Richard Royall: He distinguishes three questions: belief, action, and evidence:

1. What do I believe, now that I have this observation?
2. What should I do, now that I have this observation?
3. How should I interpret this observation as evidence regarding [H_0] versus [H_1]?

Exhibit (i): Law of Likelihood Compared to a Significance Test.

Why the LL Reject Composite Hypotheses

Royall holds that all attempts to say whether $\boldsymbol{x}$ is good evidence for H, or even if $\boldsymbol{x}$ is better evidence for H than is $\boldsymbol{y}$, are futile. Similarly,

> "What does the [LL] say when one hypothesis attaches the same probability to two different observations? It says absolutely nothing . . . [it] applies when two different hypotheses attach probabilities to the same observation" (Royall 2004, p. 148).

The severe tester distinguishes the evidential warrant for one and the same hypothesis H in two cases: one where it was constructed post hoc, cherry picked, and so on, a second where it was predesignated.

Souvenir B: Likelihood versus Error Statistical

To the Likelihoodist, points in favor of the LL are:

- The LR offers "a precise and objective numerical measure of the strength of statistical evidence" for one hypotheses over another; it is a frequentist account and does not use prior probabilities (Royall 2004, p. 123).
- The LR is fundamentally related to Bayesian inference: the LR is the factor by which the ratio of posterior probabilities is changed by the data.
- A Likelihoodist account does not consider outcomes other than the one observed, unlike P-values, and Type I and II errors. (Irrelevance of the sample space.)
- Fishing for maximally fitting hypotheses and other gambits that alter error probabilities do not affect the assessment of evidence; they may be blocked by moving to the "belief" category.

To the error statistician, problems with the LL include:

- LRs do not convey the same evidential appraisal in different contexts.
- The LL denies it makes sense to speak of how well or poorly tested a single hypothesis is on evidence, essential for model checking; it is inapplicable to composite hypothesis tests.
- A Likelihoodist account does not consider outcomes other than the one observed, unlike P-values, and Type I and II errors. (Irrelevance of the sample space.)
- Fishing for maximally fitting hypotheses and other gambits that alter error probabilities do not affect the assessment of evidence; they may be blocked by moving to the "belief" category.

Notice, the last two points are identical for both. What's a selling point for a Likelihoodist is a problem for an error statistician.

Notes 1.5 Trying and Trying again: Key Items

" trying and trying again" to achieve statistical significance, stopping rules and their relevance/irrelevance

- Edwards, Lindman, and Savage (E, L, & S, 1963).
- Simmons, Nelson, and Simonsohn

The Likelihood Principle (LP).

> *Weak Repeated Sampling Principle.* (Cox and Hinkley 1974, p. 51). "[W]e should not follow procedures which for some possible parameter values would give, in hypothetical repetitions, misleading conclusions most of the time" (ibid., pp. 45– 6).

The 1959 Savage Forum

Arguments from Intentions:

- Error Probabilities Violate the LP
- Problem of " known (or old) evidence" made famous by Clark Glymour (1980).

Souvenir C. A Severe Tester's Translation Guide [i]