SECOND SESSION 6014 JAN 25

Excursion 1 Tour II: Error Probing Tools vs. Logics of Evidence (p. 30)

To understand the stat wars, start with the holy grail–a purely formal (syntactical) logic of evidence

It should be like deductive logic but with probabilities

Engine behind probabilisms (e.g., Carnapian confirmation theories Likelihood accounts, Bayesian posteriors)

Notation Pr(x; H)

Pr($\boldsymbol{x}$;H): the probability of $\boldsymbol{x}$ computed under the assumption. In the Lady Tasting Tea example, the hypothesis under test was that the probability of success on each trial is .5, in a Bernoulli model

We can write $\boldsymbol{x}_0$ to indicate a specific fixed value of x

$X = 1$, to symbolize the result is a ‘success’

$X = 0$, for ‘failure’

Comparative Logic of Support

- **Ian Hacking (1965)** “Law of Likelihood”: ***x*** support hypothesis H_0 less well than H_1 if,

  $\Pr(\boldsymbol{x};H_0) < \Pr(\boldsymbol{x};H_1)$

(rejects in 1980)

The data support H_0 less well than they support H_1 if ***x*** *is is less* probable under H_0 *than under* H_1

- The “**likelihood”** of H_0 is less than H_1

- ***x*** supports hypothesis H_0 less well than H_1 if,

  $\Pr(\boldsymbol{x};H_0)/\Pr(\boldsymbol{x};H_1) < 1$

Or,

- ***x*** support hypothesis H_1 better than H_0 if,

  $\Pr(\boldsymbol{x};H_1)/\Pr(\boldsymbol{x};H_0) > 1$

H_0: Lady is guessing (Pr(Success on each trial) = .5) or $\theta = .5$

H_1: Lady can always tell if tea or milk is first ($\theta = 1$)

$$\Pr(\boldsymbol{x};H_1)/\Pr(\boldsymbol{x};H_0) = 2$$

Likelihood Principle (LP)

In probabilisms, the import of the data is via the ratios of *likelihoods* of hypotheses

$$\mathrm{Pr}(\boldsymbol{x}_0;H_0)/\mathrm{Pr}(\boldsymbol{x}_0;H_1)$$

The data $\boldsymbol{x}_0$ are fixed, while the hypotheses vary

A pivotal disagreement in the philosophy of statistics battles

Trick Deck (p. 38)

x: draws an ace of spades

H_0: *normal deck;* H_1: *trick deck (all aces of spades)*

$\Pr(\boldsymbol{x}_0;H_1)/\Pr(\boldsymbol{x}_0;H_0) = 1/\ 1/52$

H_1 is better supported than H_0 *by a factor of 52*

No matter what card is drawn, the Law implies the corresponding trick deck is much better supported (52 cards just like the one observed)

Royall bites the bullet

Pr(LR favors trick deck; normal deck) = 1

Why N-P Introduced error-probabilities

“there *always* is such a rival hypothesis [H_1] *viz.*, that things just had to turn out the way they actually did” (Barnard 1972, 129).

- Pr(H_0 is less well supported than H_1; H_0) is high

for some H_1 or other

- Pr(Test would yield better (comparative) support for some H_1; even though H_0) = high

Souvenir C, p. 52—a different example but the same point

Error Probability:

“In order to fix a limit between ‘small’ and ‘large’ values of [the likelihood ratio] *we must know how often such values appear when we deal with a true hypothesis*.” (Pearson and Neyman 1967, 106)

Requires considering the general behavior of the tool, how it would perform with other data

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)

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)

Souvenir B Likelihood vs error statistics (41)

Last bullet: what’s a plus for the likelihoodist (fishing for maximally fitting alternatives does not alter the evidence) is a problem for the error statistician

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)

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

- Sufficient finagling—cherry-picking, significance seeking, multiple testing, post-data subgroups, trying and trying again—may practically guarantee a preferred claim *H* gets support, even if it's unwarranted by evidence
- You might report a "nominally" small P-value even though the probability of mistakenly inferring a "real effect" is high.

Optional Stopping:

Error probing capacities are altered not just by data dredging, but also via data dependent stopping rules: *EXAMPLE. Testing claims about the mean* μ *of a normal distribution*

H_0: no effect vs. H_1: some effect

2-sided H_0: μ = 0 vs. H_1: μ ≠ 0.

(use data on observed benefits on average (M) among treated to learn about population mean μ)

With optional stopping, instead of fixing the sample size *n* in advance, *n* is determined by a *stopping rule*:

- Keep sampling until H_0 is rejected at (“nominal”) 0.05 level

Keep sampling until sample mean M differs from 0 from some amount (2SE)

- *Trying and trying again*: Having failed to rack up a statistically significant difference after 10 trials, go to 20, 30 and so on until obtaining a 2 SE difference

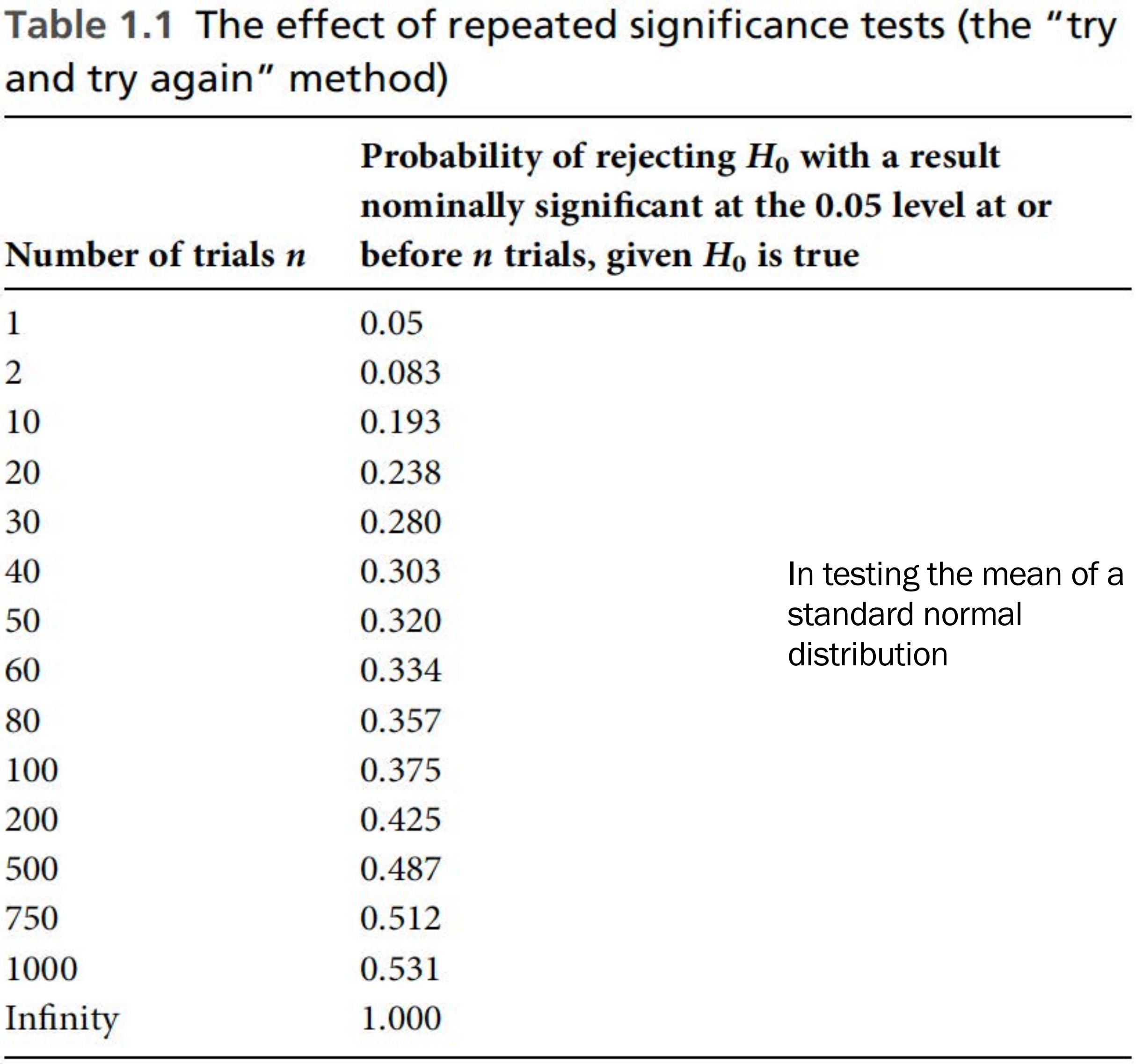

Table 1.1 The effect of repeated significance tests (the "try and try again" method)

Number of trials n	Probability of rejecting H_0 with a result nominally significant at the 0.05 level at or before n trials, given H_0 is true
1	0.05
2	0.083
10	0.193
20	0.238
30	0.280
40	0.303
50	0.320
60	0.334
80	0.357
100	0.375
200	0.425
500	0.487
750	0.512
1000	0.531
Infinity	1.000

In testing the mean of a standard normal distribution

***Nominal* vs. *Actual* significance levels :**

- 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)

Optional Stopping (43)

- “if an experimenter uses this [optional stopping] procedure, then with probability 1 he will eventually reject any sharp null hypothesis, even though it be true”
  (Edwards, Lindman, and Savage 1963, 239)

- Understandably, they observe, the significance tester frowns on this, or at least requires adjustment of the P-values

Imagine instead if an account advertised itself as ignoring stopping rules (43)

- “[the] irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design that had been lost by classical emphasis on significance levels (in the sense of Neyman and Pearson).” (Edwards, Lindman, and Savage 1963, 239)

These same authors who warn that to ignore stopping rules is to guarantee rejecting the null hypothesis even if it’s true” (43) declare it irrelevant.

What counts as cheating depends on statistical philosophy

- Are they contradicting themselves?
- "No. It is just that what looks to be, and indeed is, cheating from the significance testing perspective is not cheating from [*their*] Bayesian perspective." (43)

At odds with reforms to block irreplication: 21 Word Solution

- Replication researchers (re)discovered that data-dependent hypotheses and stopping are a major source of spurious significance levels.
- Statistical critics, Simmons, Nelson, and Simonsohn (2011) place at the top of their list the need to block flexible stopping

“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).

Competing Intuitions

The error here would be mistaking chance variability as genuine

- You've scarcely bent over backwards to block that mistake by trying and trying again (at least using this test) and failing to report this
- On the other hand, why should intentions to stop alter the import of the evidence? (what if she always intended to go to 100 trials, say)
- Inference by Bayes Theorem says it should not: (45, derived in 1 page). (stopping rule principle)

Bayesian sequential (adaptive) analysts say:

"The [regulatory] requirement of type I error control for Bayesian adaptive designs causes them to lose many of their philosophical advantages, such as compliance with the likelihood principle" (Ryan et al. 2020, radiation oncology)

Errorstatistics.com: https://errorstatistics.com/2021/08/21/should-bayesian-clinical-trialists-wear-error-statistical-hats/

The authors admit

that the type I error was inflated in the Bayesian adaptive designs …that allowed early stopping for efficacy and without adjustments to account for multiplicity. …[This] may violate the weak repeated sampling principle [Cox and Hinkley 1974] …

Whilst the long-run frequency behaviour of sequential testing procedures is irrelevant from the strict Bayesian perspective, long-run properties have been established as being important in the clinical trial setting, … One of the core functions of health-care regulators is to prevent those that market interventions from making spurious claims of benefit. For this reason, adequate control of type I error is one of the perennial concerns when appraising the results of confirmatory clinical trials.

Still….

- The requirement of type I error control for Bayesian adaptive designs …creates a design that is inherently frequentist.

Questions for Session #2 (1/25)

1. What’s the difference between weak and strong severity (14, 23)? Although a skeptic could stop after weak severity (why?), the severe tester accepts strong severity as well. Why?

2**. Explain the Likelihood Principle in contrast to Error Statistical Principles. Why does Cox say that optional stopping violates the Weak Repeated Sampling principle?**

Further: What’s your *current* intuition about this highly controversial issue?

Probabilists can still block intuitively unwarranted inferences (without error probabilities)?

- Supplement with subjective beliefs: What do I believe? As opposed to What is the evidence? (Royall's delineation)
- Likelihoods + prior probabilities

Richard Royall

1. What do I believe, given x

2. What should I do, given x

3. How should I interpret this observation x as evidence? (comparing 2 hypotheses)

For #1–degrees of belief, Bayesian posteriors

For #2–frequentist performance

For #3–LL

p. 33

Hypotheses vs observable events

- Statistical hypotheses assign probabilities to data Pr($\boldsymbol{x}_0$;H_1), but it's rare to assign frequentist probabilities to hypotheses

Hypotheses

- The deflection of light due to the sun is 1.75 degrees
- Deficiency in vitamin D increases chances of fractures
- HRT decreases age-related dementia
- IQ is more variable in men than women

A subjective might assign degrees of belief: betting

Bayes Rule (next time)

They might look at Pr(*H*|***x***) vs. Pr(*H)*

$$\Pr(H|\mathbf{x}) = \frac{\Pr(\mathbf{x}|H)\Pr(H)}{\Pr(\mathbf{x})}$$

(p. 24)

$$\Pr(H|\mathbf{x}) = \frac{\Pr(\mathbf{x}|H)\Pr(H)}{\Pr(\mathbf{x}|H)\Pr(H) + \Pr(\mathbf{x}|{\sim}H)\Pr({\sim}H)}$$

Bayesians may (indirectly) block implausible inferences

- With a low prior degree of belief on *H* (e.g., *real effect*), the Bayesian can block inferring *H*

Concerns:

- Doesn’t show what has gone wrong—it’s the multiplicity
- The believability of post hoc hypotheses is what makes them so seductive
- Claims can be highly probable (or even known) while poorly probed.
- Additional source of flexibility

Current State of Play in Bayesian-Frequentist Wars

1.3 View from a Hot-Air Balloon (p. 23)

How can a discipline, central to science and to critical thinking, have two methodologies, two logics, two approaches that frequently give substantively different answers to the same problems? Is complacency in the face of contradiction acceptable for a central discipline of science? (Donald Fraser 2011, p. 329)

Most Bayesians (last decade) use "default" priors: unification

- 'Eliciting' subjective priors too difficult, scientists reluctant for subjective beliefs to overshadow data

> "[V]irtually never would different experts give prior distributions that even overlapped" (J. Berger 2006, 392)

- Default priors are supposed to prevent prior beliefs from influencing the posteriors–data dominant

Marriages of Convenience

Why?

For subjective Bayesians, to be less subjective

For frequentists, to have an inferential or epistemic interpretation of error probabilities

How should we interpret them?

- “The priors are not to be considered expressions of uncertainty, ignorance, or degree of belief. Conventional priors may not even be probabilities…” (Cox and Mayo 2010, 299)

- No agreement on rival systems for default/non-subjective priors.

- No such thing as uninformative priors (2000?)

Some Bayesians reject probabilism (Falsificationist Bayesians)

- *“[C]rucial parts of Bayesian data analysis, such as model checking, can be understood as ‘error probes’ in Mayo’s sense” which might be seen as using modern statistics to implement the Popperian criteria of severe tests.*

  (Andrew Gelman and Cosma Shalizi 2013, 10).

Decoupling

Break off stat methods from their traditional philosophies

Can Bayesian methods find a new foundation in error statistical ideas? (p. 27)

Excursion 6: (probabilist) foundations lost; (probative) foundations found

Questions for Session #2* (1/25)

1. What's the difference between weak and strong severity (14, 23)? Although a skeptic could stop after weak severity (why?), the severe tester accepts strong severity as well. Why?

2. Explain the Likelihood Principle in contrast to Error Statistical Principles. Why does Cox say that optional stopping violates the Weak Repeated Sampling principle? Further: What's your *current* intuition about this highly controversial issue?

3. Explain and contrast error statistical, likelihoodist, and Bayesian positions just to the extent discussed in Excursion 1.
Cla activity

From First Session: P-value

You don’t need a strict cut-off to evaluate a particular result:

“Suppose that we were to accept the available data as evidence against H_0. Then we would be bound to accept all data with a larger value of [d] as even stronger evidence.

Hence [the observed P-value] is the probability that we would mistakenly declare there to be evidence against H_0, were we to regard the data under analysis as just decisive against H_0.” (Type 1 error)

Cox and Hinkley (1974, 66)