Methodological Probability. A valuable idea to take from Popper is that probability in learning attaches to a method (p. 80)

An error probability is a special case of a methodological probability.

We want methods with a high probability of teaching us (and machines) how to distinguish approximately correct and incorrect interpretations of data

“epistemology” vs “variability” shoehorn:
The typical choices for probability : “in here” (beliefs ascertained by introspection) or “out there” (frequencies in long-runs, or chance mechanisms).

We reject this (Excur 1, Souv. (D), p. 54 Reid and Cox).

Falsification is rarely deductive (outside of philosophy classes: all swans are white)

Though there are analogous cases, I will argue, in testing assumptions (null is: IID is satisfied)

However, we can erect reliable rules for falsifying claims with severity (corroborate their denials, p. 81)

Popperian falsificationists must suppose ”ampliative” inference

We will call it “inductive” even thought not the unreliable enumerative induction (EI)

We say that a theory is falsified only if we have accepted basic statements which contradict it…This condition is necessary, but not sufficient; for we have seen that non-reproducible single occurrences are of not sigf]nificance to science…We shall take it as falsified **only if we discover a reproducible effect which refutes the theory. …we only accept the falsification if a low level empirical hypothesis which describes such an effect is proposed and corroborated.** (Popper 1959, 86)
(see also pp 82-3 in SIST)

(~ to Fisher's "we need not an isolated result")

We need a <u>falsifying hypothesis</u> (typically statistical): a hypothesis inferred to falsify some other claim

Kuru (which means “shaking”) SIST p. 81

Widespread among the Fore people of New Guinea,1960s.

- Kuru, and (what we now know to be) related diseases, e.g., Mad Cow, Crutzfield Jacobs, scrapie) are “spongiform” diseases: the brains have a spongy appearance.

- Turns out transmission was through mortuary cannibalism

Falsifying the central dogma of biology (infection requires nucleic acid) involved moving from the bottom up (not a series of conjunctions p. 84).

- prions are not eradicated with techniques known to kill viruses and bacteria, but they are deactivated with substances known to kill proteins
- If it were a mistake to regard prions as having no nucleic acid, then at least one of these known agents would have eradicated it

The discovery of prions led to a “revolution” in molecular biology, Prusiner gets a Nobel prize in 1997.

Souvenir E. (p. 87) An array of questions, problems, models

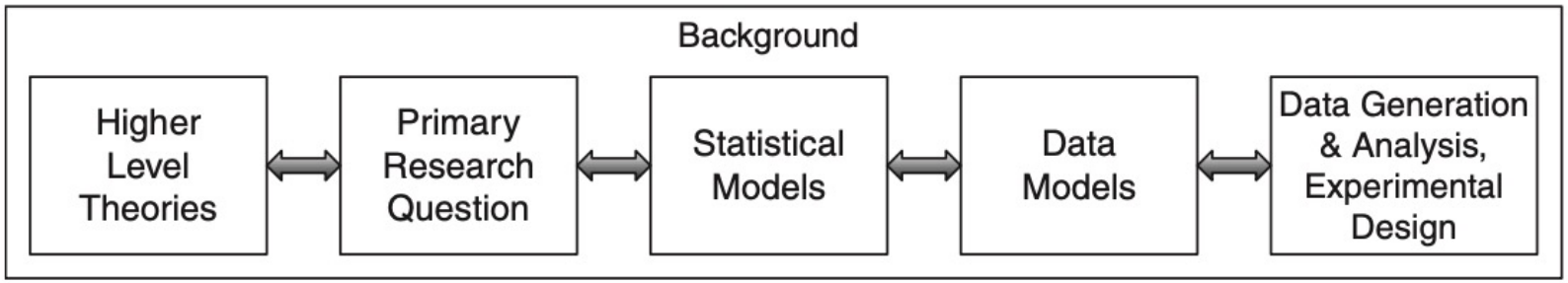


Figure 2.1 Array of questions, problems, models.

What should we say about demarcation?

Forty years ago, Larry Laudan's famous (1983) paper declared the demarcation problem taboo: "The Demise of the Demarcation Problem".

This is a highly unsatisfactory situation for philosophers of science, especially with today's statistical replication crisis.

Exhibit (vi) Revisiting Popper's Demarcation

Popper confuses things by putting the weight on the theory to be scientific rather than the inquiry or test

His own construal of severity as satisfying a type of novelty goes against this (2.4).

- We want to distinguish meritorious modes of inquiry from those that are BENT.
- If the test methods enable biasing selection effects, *ad hoc* face-saving devices, then the inquiry is unscientific. (biasing selection effects p. 92)
- Despite being logically falsifiable, theories can be *rendered immune from falsification*.

From my current blogpost

Some areas have so much noise and/or flexibility that they can't or won't distinguish warranted from unwarranted explanations of failed predictions.

It does not suffice– for an inquiry to be scientific– that there is criticism of methods and models.

The criticism must be constrained by what's actually responsible for any alleged problems. It may be correct to criticize an inference to a hypothesis H, but it may be for the wrong reason.

Demarcating scientific inquiry (4 requirements)

A severe tester says: A scientific inquiry or test must be able:

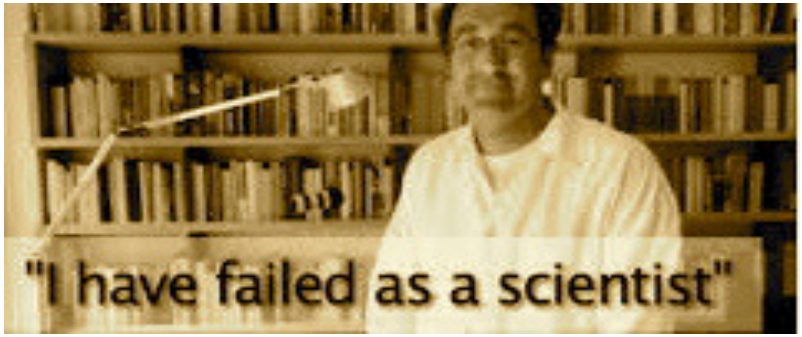
(a) to block inferences that fail the minimal requirement for severity

(b) to embark on a reliable probe to pinpoint blame for anomalies

(c) (from (a)) to directly pick up on altered error probing capacities due to biasing selection effects, optional stopping, cherry picking, data-dredging etc.

(d) (from (b)) to test and falsify claims.
So we get four requirements for an inquiry to be scientific

Notes on the replication revolution in psychology (p. 97)



- Diederik Stapel, the social psychologist who fabricated his data (2011)
- Investigating Stapel revealed a culture of verification bias
- A string of high profile cases followed, as did replication research

The Committee Investigating Stapel:

"One of the most fundamental rules of scientific research is that an investigation must be designed in such a way that facts that might refute the research hypotheses are given at least an equal chance of emerging as do facts that confirm the research hypotheses." (Levelt Committee., p. 48)

This is the gist of our minimal requirement for evidence

“I see a train-wreck looming,” Daniel Kahneman, calls for a “daisy chain” of replication in Sept. 2012

OSC
OPEN SCIENCE
COLLABORATION

OSC: Reproducibility Project: Psychology: 2011-15 (Led by Brian Nozek, U. VA)

Failed Replication

- Failed replication: Results found statistically significant are not found significant when an independent group tries to replicate the finding with new subjects, and more stringent protocols

- Note on terminology (replicating vs reproducing p. 97)

Recall from the First Session of this seminar:

"Several methodologists have pointed out that the high rate of nonreplication of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. …

(John Ioannidis 2005, 0696)

Simple significance tests (Fisher)

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

…we find a function $d(\boldsymbol{X})$ of the data, the **test statistic**, such that

- the larger the value of $d(\boldsymbol{X})$ the more inconsistent are the data with H_0;
- Must be able to compute $\Pr(d(\boldsymbol{X}) \geq d(\mathbf{x}); H_0)$

See also Cox (1977), pp. 93-4

Differences are expected, so when is it large enough for evidence against H_0 ?

Example of Lady Tasting Tea (wouldn't be impressed if she got only 9 of 16 correct)

Lady Tasting Tea (16-17)

point just now is this: so long as lacking ability is sufficiently like the canonical "coin tossing" (Bernoulli) model (with the probability of success at each trial of 0.5), we can learn from the test procedure. In the Bernoulli model, we record success or failure, assume a fixed probability of success θ on each trial, and that trials are independent. If the probability of getting even more successes than she got, merely by guessing, is fairly high, there's little indication of special tasting ability. The probability of at least 9 of 16 successes, even if $\theta = 0.5$, is 0.4. To abbreviate, Pr(at least 9 of 16 successes; H_0: $\theta = 0.5$) = 0.4. This is the *P*-value of the observed difference; an unimpressive 0.4. You'd expect as many or even more "successes" 40% of the time merely by guessing. It's also the *significance level attained* by the result. (I often use *P*-value as it's shorter.) Muriel Bristol-Roach pledges that if her

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

"Suppose that we ere 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)

- Small P-value indicates *some* underlying discrepancy from H_0 because **very probably you would have seen a less impressive** difference d than observed d_{0bs} were H_0 true.

- Usually require .05, .025, .01
- Tool to avoid being fooled by randomness
- She actually is said to have gotten them all right
- Still not evidence of a substantive scientific hypothesis H*

Problems: *fallacies of rejection p. 94*

- The reported (nominal) statistical significance result is *spurious* (it's not even an actual P-value).

(This can happen in two ways: biasing selection effects, or violated assumptions of the model.)

Problems: 3 *fallacies of rejection p. 94*

- The reported statistically significant result is genuine, but it's an isolated effect not yet indicative of a genuine experimental phenomenon. (Isolated low P-value ≠> *H*: statistical effect)

- There's evidence of a genuine statistical phenomenon but either (i) the magnitude of the effect is less than purported, call this a *magnitude error*, or (ii) the substantive interpretation is unwarranted. (*H* ≠> *H**)

An *audit* of a P-value: a check of any of these concerns, generally in order, depending on the inference.

So I place the background information for auditing throughout our ‘series of models’ representation (figure 2.3, p. 87.

Meehl blames Fisher p. 93

- "[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 which will rarely fail to give us a statistically significant result." (Fisher 1935, 14) **(low P-value ≠> H: statistical effect)**

- "[A]ccording to Fisher, rejecting the null hypothesis is not equivalent to accepting the efficacy of the cause in question. The latter...requires obtaining more significant results when the experiment, or an improvement of it, is repeated at other laboratories or under other conditions." (Gigerentzer 1989, 95-6) **(*H* ≠> *H**)**

Biasing selection effects:

One function of severity is to identify problematic selection effects (not all are) p. 92

- ***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, severity is seriously altered or incapable of being assessed**

SEV is also applied quasi-formally and informally

- If flaws in the substantive alternative *H** have not been probed by, the inference from a statistically significant result to *H** fails to pass with severity
- Generally goes beyond statistics
- Largely ignored in today's replication research

People may want to believe claims (for political or ethical reasons)

- Diederik Stapel says he always read the research literature extensively to generate his hypotheses.

- *"So that it was believable and could be argued that this was the only logical thing you would find*." (E.g., eating meat causes aggression.)
- (In "The Mind of a Con Man," NY Times, April 26, 2013[4])

To return to the OSC: Reproducibility Project:

- Crowd sourced: Replicators chose 100 articles from three journals (2008) to try and replicate using the same method as the initial research: **direct replication**

Does a negative replication mean the original was a false positive?

- Preregistered, avoid P-hacking, designed to have high power
- Free of “perverse incentives” of usual research: guaranteed to be published

But there may also be biases

- But might they be influenced by replicator's beliefs in the claim?
- Replicator's attitude toward the methodology
- Subjects (often students) often told the purpose of the experiment
- Kahneman calls for a "new etiquette" (getting approval from original researcher) p.99

- One of the non-replications: cleanliness and morality: *Do cleanliness primes make you less judgmental?*

"*Ms. Schnall had 40 undergraduates unscramble some words.* ***One group unscrambled words that suggested cleanliness*** (pure, immaculate, pristine), while the **other group unscrambled neutral words.** ***They were then presented with a number of moral dilemmas, like whether it's cool to eat your dog after it gets run over by a car****."*

Turns out it did. Subjects who had unscrambled clean words weren't as harsh on the guy who chows down on his chow."
(Bartlett, *Chronicle of Higher Education*)

- By focusing on the P-values, they ignore the larger question of the methodological adequacy of the leap from the statistical to the substantive.

- Are they even measuring the phenomenon they intend? Is the result due to the "treatment"?

Nor is there discussion of the multiple testing in the original study

- Only 1 of the 6 dilemmas even in the original study showed statistically significant differences in degree of wrongness–not the dog one
- No differences on 9 different emotions (relaxed, angry, happy, sad, afraid, depressed, disgusted, upset, and confused)
- Many studies are coming into experimental philosophy: philosophers of science need to critique them

Skip Researcher degrees of freedom in the Schnall study

After the priming task, participants rated six moral dilemmas : ‘‘Dog’’ (eating one’s dead dog), ‘‘Trolley’’ (switching the tracks of a trolley to kill one workman instead of five), ‘‘Wallet’’ (keeping money inside a found wallet), ‘‘Plane Crash’’ (killing a terminally ill plane crash survivor to avoid starvation), ‘‘Resume’’ (putting false information on a resume), and ‘‘Kitten’’ (using a kitten for sexual arousal). Participants rated how wrong each action was from 0 (perfectly OK) to 9 (extremely wrong).

We should worry much more about

- Links from experiments to inferences of interest
- I've seen plausible hypotheses poorly tested
- Macho men, self esteem measures p. 101-4
- p. 101 Psychometrician Joel Michell castigates psychology for having bought the operationalist Stevens' (1946, p. 667) definition of measurement

Macho Men

H: partner's success lowers self-esteem in men*

I have no doubts that certain types of men feel threatened by the success of their female partners, wives or girlfriends

I've even known a few.

Can this be studied in the lab? Ratliff and Oishi (2013) did:

.

H:* "men's implicit self-esteem is lower when a partner succeeds than when a partner fails."

Not so for women

Treatments: Subjects are randomly assigned to five "treatments": think, write about a time your partner succeeded, failed, succeeded when you failed (partner beats me), failed when you succeeded (I beat partner), and a typical day (control).

Effects: a measure of "self esteem"

Explicit: "How do you feel about yourself?"

Implicit: a test of word associations with "me" versus "other".

None showed statistical significance in explicit self-esteem, so consider just implicit measures

Some null hypotheses: The average self-esteem score is no different (these are statistical hypotheses) (p. 102)

a)when partner succeeds (rather than failing)

b)when partner beats (surpasses) me or I beat her

c)control: when she succeeds, fails, or it's a regular day

There are at least double this, given self-esteem could be "explicit" or "implicit" (others too, e.g., the area of success)

Only null (a) was rejected statistically!

Should they have taken the research hypothesis as disconfirmed by negative cases?

Or as casting doubt on their test?

Or should they just focus on the null hypotheses that were rejected, in particular null (a), for implicit self esteem.

They opt for the third.

It's not that they should have regarded their research hypothesis *H** as disconfirmed much less falsified.

This is precisely the nub of the problem! I'm saying the hypothesis that the study isn't well-run needs to be considered

- Is the artificial writing assignment sufficiently relevant to the phenomenon of interest? (look at proxy variables)
- Is the measure of implicit self esteem (word associations) a valid measure of the effect? (measurements of effects)

Clearly they expected to reject the null in b), that “she beat me in X” would have a greater negative impact on self-esteem than “she succeeded at X”.

Still, they could view it as lending “some support to the idea that men interpret ‘my partner is successful’ as ‘my partner is more successful than me” (p. 698),

….as do the authors.

That is, *any success of hers is always construed by Macho man as, she beat me.*

Bending over Backwards

For the stringent self-critic, this skirts too close to viewing the data through the theory, a kind of "self-sealing fallacy".

"I'm talking about a specific, extra type of integrity...bending over backwards to show how you're maybe wrong, that you ought to have when acting as a scientist." (R. Feynman 1974)

I'm describing what's needed to show "sincerely trying to find flaws" under the austere account I recommend

The most interesting information was never reported! Perhaps it was never even looked at: *what they wrote about*.

Replication Research in Psychology Under an Error Statistical Philosophy

Replication problems can't be solved without correctly understanding their sources

Biggest sources of problems in replication crises

(a) *Stat H ->research H** and (b) biasing selection effects:

Reasons for (a): focus on P-values and Fisherian tests ignoring N-P tests (and the illicit NHST that goes directly ***H–> H****)

(b) Biases are exacerbated by accounts that ignore selection effects, e.g., probabilisms that embrace the likelihood principle LP

What's replicable? discrepancies that are severely warranted

There's no point in raising thresholds for significance if your methodology does not pickup on biasing selection effects.

Solving the problem of induction now (p. 107)

Viewing inductive inference as severe testing, the problem of induction is transformed into the problem of showing the existence of severe tests and methods for identifying insevere ones.

The trick isn't to have a formal, context free method that you can show is reliable… the trick is to have methods that alert us when an application is shaky.

Build a Repertoire of Errors of Inquiry

To identify fields and inquiries where inference problems are solved efficiently, and how obstacles are overcome–or not.

Calls upon material from the entire book

Builds on lift-off, convergent arguments (from coincidence), pinpointing blame (Duhem's problem) and falsification.

108: The problem has always been rather minimalist: to show at least some reliable methods exist: Just find me one.

The imaginary talk on solving induction proceeds by 4 questions

1. What warrants inferring a hypothesis that stands up to severe tests?
2. What enables induction (as severe testing) to work?
3. What is Neyman's quarrel with Carnap?
4. Neyman's empirical justification for using statistical models

1. What warrants inferring a hypothesis that stands up to severe tests? (108)

Ruling out rigging (by weak severity) and the desire to learn leads to strong severity

2. What enables induction (as severe testing) to work? (109-110)

- Local arguments from error, experimental knowledge
- Use of *i-assumption*
- Repertoire of errors & mistakes

2. What enables induction (as severe testing) to work? (109-110)

- Local arguments from error, experimental knowledge
- Use of *i-assumption*
- Repertoire of errors & mistakes

3. What is Neyman's quarrel with Carnap?

- Enumerative induction ignores the need for a (Binomial model based on Bernouilli trials)

4. Neyman's empirical justification for using statistical models (p. 111)

"There are real experiments that "even if carried out repeatedly with the utmost care to keep conditions constant, yield varying results" (Neyman 1952, p. 25). …roulette wheels (electrically regulated), tossing coins with a special machine ..the number of disintegrations per minute in a quantity of radioactive matter, and the tendency for properties of organisms to vary despite homogeneous breeding."

“Whenever we succeed in arranging” the data generation such that the relative frequencies adequately approximate the mathematical probabilities in the sense of the LLN (Law of Large Numbers), we can say that the probabilistic model “adequately represents the method of carrying out the experiment” (ibid. p.19).

[then] we are warranted in describing the results of real experiments as random samples from the population given by the probability model.

From C. S. Peirce:

All we need is that mysterious “supernal powers withhold their hands and let me alone, and that no mysterious uniformity…interferes with the action of chance” (Peirce 2.749) to justify induction.

Bernouilli Model:
Recall our example:

In general, with this x_0,

$\text{Lik}(\theta; x_0) = \Pr(1,1,0,1; \theta) = (\theta)(\theta)(1-\theta)(\theta) = \theta^3(1-\theta)$

order doesn't matter

This is *one* way of getting 3 out of 4 successes, how many are there in general

<1,1,0,1>, <1,1,1,0> , <1,0,1,1>, <0,1,1,1>

They are mutually exclusive outcomes, so by the addition rule:

Pr(3 out of 4 successes) = $4 \times \theta^3(1 - \theta)$

Binomial distribution

p.111 Pr(k out of n successes) =
(n choose k) x $\theta^k(1 - \theta)^{n - k}$

mean = n θ