PHIL 6014: Spring 2023
Special Topics in Philosophy
Philosophy of Inductive-Statistical Inference
Wed. 4-6:30 McBryde 223
Prof. Deborah Mayo

This course is an introduction to the philosophy of inductive-statistical inference in relation to general problems of philosophy of science (e.g., falsification vs confirmation, underdetermination, science vs pseudoscience) and to current controversies regarding uncertain inference in scientific practice (e.g., statistical significance tests, Bayesian vs frequentist methods, replication crisis, and science and values in evidence policy). We will study examples of statistical evidence in the law, psychology, medicine and physics. You do not need to have a statistical or a philosophical background, only an interest in learning about philosophy of statistics (PhilStat) in its relations to problems of philosophy of science and statistical epistemology.

Who is this course is directed to?

Anyone wishing to

- do original research in some aspect of philosophy of science or epistemology involving uncertain inference
- understand the current “crisis of replication” in fields that use statistics, data science, machine learning

- Courses in research methods in the *social sciences* allow for an impressive array of statistical methods and models,

- but using them successfully requires reacting to challenges regarding their legitimate use and interpretation.

- —often subject of philosophical controversies (although it is typically not recognized)

Researchers may protest: “There’s nothing philosophical about our criticisms of statistical significance tests; the problem is that small P-values are taken as giving low probability to a null hypothesis (of no effect)”

The reverse problem often arises in courses in philosophy of science:

- Without statistical understanding, tackling problems about uncertain evidence are often out of touch with abilities and problems of inference tools actually used

- Practitioners consulting texts in philosophy of science are typically at a loss to see how they are relevant to their problems.

Why is it important to address these issues to those without (as well as with) a statistical background?

- With the rise of big data, machine learning, and data science, high powered methods make the computations invisible to most users
- Methodological advocacy is directed at being nontechnical or requiring very minimal understanding of technical complexities
- We need to be able to critically analyze the debates about methods

- By the time you finish this course, you will be able to comprehend and critically evaluate the debates, disagreements, and controversies now taking place

- You will be beyond the typical audience to which the popular, “non-technical” arguments are directed.

- Using my book as a primary text, along with philosophical and statistical resources, I propose to pull back the cover on the central debates, disagreements and arguments

- You will also understand the historical, statistical and personality backdrop to the issues
- I'm keen to write a new edition and/or companion to the book—with your help
- I am posting outlines of the chapters ("tours") to help you cover the material

- You can skip any technical sections in your reading—but bring to class any unclear concepts and notation, and ideally some indication of what you skipped

- But please read the Preface!

- The first excursion jumps into the main debates with the promise of going much more slowly through them.

What is the Philosophy of Statistics (PhilStat)?

Statistical inference uses data ***x*** to learn about aspects of processes or mechanisms that produce them, along with a qualification of uncertainty—how good a job did it do?

- Claims inferred are statistical generalizations, parameters in theories and models, general predictions rather than specific events.

- The inference goes beyond the data—ampliative or inductive (7, 10)

What is the Philosophy of Statistics (PhilStat)?

At one level of analysis, statisticians and philosophers of science ask many of the same questions:

- *What should be observed and what may justifiably be inferred from the resulting data?*
- *How well do data confirm or fit a model?*
- *What is a good test?*

- *How can spurious relationships be distinguished from genuine regularities? from causal regularities?*
- *How can we infer more accurate and reliable observations from less accurate ones?*
- *When does a fitted model account for regularities in the data?*

- These very general questions are entwined with long standing debates in philosophy of science (even though they approach them very differently)

- Statistics is a kind of “applied philosophy of science” (Kempthorne, 1976).

"A Statistical Scientist Meets a Philosopher of Science"

Sir David Cox: "Deborah, in some fields foundations do not seem very important, but we both think foundations of statistical inference are important; why do you think that is?"

Mayo: "…in statistics …we invariably cross into philosophical questions about empirical knowledge and inductive inference." (Cox and Mayo 2011)

Sir David Cox: 15 July 1924 - 18 January 2022

Statistics → Philosophy

3 ways statistical accounts are used in philosophy of science

(1) Model Scientific Inference—capture the actual or rational ways to arrive at evidence and inference

(2) Solve (or reconstruct) Philosophical Problems about scientific inference, observation, experiment; (problem of induction, objectivity, underdetermination).

(3) Perform a Metamethodological Critique
-scrutinize methodological rules (e.g., novel data)
Analytic: what do we mean by evidence? reliable?

Philosophy → Statistics

job is to help resolve the conceptual, logical, and methodological discomforts of scientists

In tackling the statistics wars one also makes progress on:

the philosopher's problems of induction, objective evidence, underdetermination.

“2-way street”

Start from the preface:
The Statistics Wars

Bayesian-Frequentist Wars

The Goal of Frequentist Inference: Construct procedures with frequentist guarantees

good long-run performance

The Goal of Bayesian Inference: Quantify and manipulate your degrees of beliefs

belief probabilism

Larry Wasserman (p. 24)

But now we have marriages and reconciliations (pp. 25-8)

- End of foundations? (Unifications and Eclecticism—”we use whatever works”)
- Long-standing battles still simmer below the surface (agreement on numbers)
- We wouldn’t have American Statistical Association Task Forces, as in 2022, were it not that concepts are more confused than ever

- I will not be proselytizing for a given school; they all have shortcomings insofar as you can identify a “school”

- The goal is to unlock the mysteries that are leaving many skeptical statistical consumers in the dark about a crucial portion of science (12)

Let’s brush the dust off the pivotal debates, walk into the museums to hear the founders: Fisher, Neyman, Pearson, Savage and many others in relation to today’s statistical crisis in science (xi)

A metastatistical tool: Statistical inference as severe testing

- Main source of the statistical crisis in science?
- We set sail with a simple tool: you don't evidence for a claim if little or nothing has been done to rule out how it can be false
- You needn't accept this principle to use it to excavate the statistics wars

A claim is warranted to the extent it passes severely

- We have evidence for a claim only to the extent that it has been subjected to and passes a test that would probably have found it flawed or specifiably false, just if it is
- This probability is the stringency or severity with which it has passed the test

A philosophical excursion

"Taking the severity principle, along with the aim that we desire to find things out… let's set sail on a philosophical excursion to illuminate statistical inference." (8)

"…and engage with a host of tribes marked by family quarrels, peace treaties, and shifting alliances" (xiv)

Revisit some taboos: problems of induction & falsification, science vs. pseudoscience

Excursion 1 How to Tell What's True About Statistical inference

Tour I: Beyond Probabilism and Performance

Most findings are false?

"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. …

It can be proven that most claimed research findings are false." (John Ioannidis 2005, 0696)


OMG!

R.A. Fisher

"[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 1947, 14)

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;

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

Example of Lady Tasting Tea.

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*

Neyman-Pearson (N-P) tests:

A null* and alternative hypotheses H_0, H_1 that are exhaustive*

H_0: “no effect” vs. H_1: “some positive effect”

Type 1 error (mistakenly rejecting) and
Type 2 error (mistakenly failing to reject)

*test hypothesis

Despite personality conflicts & jealousies

- They both fall under tools for "appraising and bounding the probabilities (under respective hypotheses) of seriously misleading interpretations of data" (Birnbaum 1970, 1033)–*error probabilities*
- Can place all under the rubric of ***error statistics***
- Confidence intervals, N-P and Fisherian tests, resampling, randomization
- Details of tests do not enter until Excursion 3

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

Severity Requirement (weak):

If the test had little or no capability of finding flaws with *H* (even if *H is* incorrect), then agreement between data $\boldsymbol{x}_0$ and *H* provides poor (or no) evidence for *H*

- Such a test fails a *minimal requirement* for a stringent or severe test
- N-P and Fisher did not put it in these terms but our severe tester does

This alters the role of probability (typically just two):

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

(e.g., Bayesian, likelihoodist, Fisher (at times))

Performance (more apt than frequentist*). Ensure long-run reliability of methods, coverage probabilities (frequentist, behavioristic Neyman-Pearson, Fisher (at times))

- There are roles for both, but neither “probabilism” nor “performance” directly captures the idea of error probing capacity
- high degree of belief (in the sense of personalism) can’t suffice: even where a claim is known to be true, it can be poorly tested

- Good long-run performance is a necessary, not a sufficient, condition for severity

    — example: 2 weighing machines

- What bothers you with selective reporting, cherry picking, stopping when the data look good, P-hacking
- Not problems about long-runs—

We cannot say the case at hand has done a good job of avoiding the sources of misinterpreting data

A claim *C* is not warranted _____

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

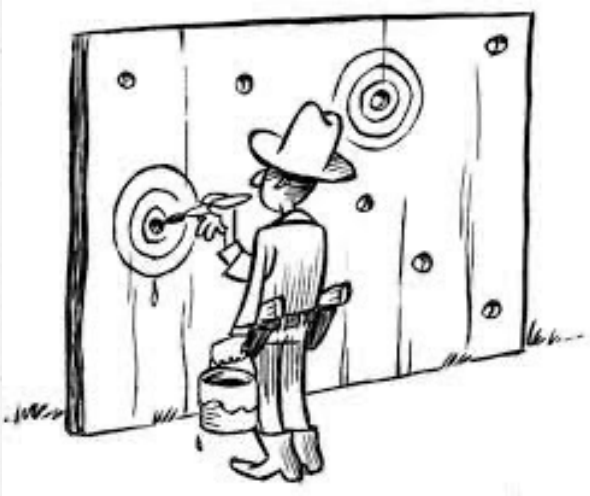
- ***Performance***: unless it stems from a method with low long-run error

- ***Probativism (severe testing)*** unless something (a fair amount) has been done to probe ways we can be wrong about *C*

Informal Examples

1. Bad Evidence No Test (BENT): Texas marksman (19), Pickrite method of investing (19)

(Having drawn a bull's eye around tightly clustered shots, claims marksman ability)



Informal Examples

2. Strong Severity (argument from coincidence to weight gain):

- Lift-off (p. 14)
- arguing from error (p. 16)

An informal example of a severe Test

To test if I've gained weight between the start of the pandemic and now, I use a series of well-calibrated and stable scales, both in March and now.

All show an over 3 lb. gain, none shows a difference in weighing EGEK, I'm forced to infer:

H: I've gained at least 3 pounds

- Properties of the scales are akin to the properties of statistical tests (performance).
- No one claims the justification is merely long run, and can say nothing about my weight.
- We infer something about the source of the readings from the high capability to reveal if any scales were wrong

STOPPED HERE Jan 18

The severe tester is assumed to be in a context of wanting to find things out

- I could insist all the scales are wrong—they work fine with weighing known objects—but this would prevent correctly finding out about weight….. (rigged alternative)

Popper vs logics of induction/ confirmation

Severity was Popper's term, and (though he never cashed it out adequately), the debate between Popperian falsificationism and inductive logics of confirmation/ support parallel those in statistics.

Notation Pr(*x*; H)

Pr(***x***;*H*): the probability of ***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

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 are less probable under H_0 *than under* H_1

- The **likelihood** of H_0 is less than H_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

"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)

A methodological probability, Popper

Don’t we want the probabilities to be assigned to statistical hypotheses and theories?

Statistical hypotheses assign probabilities to data Pr(***x***;*H*), 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 Bayesian might assign degrees of belief: betting

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)

Bayes Rule to Pr(*H*|***x***)

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

- We'll break this down shortly…

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

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?