Formal Epistemology and the Foundations of Statistics: Is the Error Statistician a Formal Epistemologist?

Deborah G. Mayo

FEW May, 2011

I do appeal to quantitative and in particular, probabilistic methods and models in my work in philosophy of science

Formal methods ⟶ Philosophy

Conversely, I think philosophers (at least philosophers of science) should help solve foundations problems in formal modeling with data, learning, and inference

(2-way street)

But whether that places me in the current camp of formal epistemologists is far from clear…

On the face of it, we might regard a formal epistemologist as appealing to statistical and formal methods to:

1. Model inference, characterize learning from data

2. Solve problems about evidence, inductive inference (underdetermination, Duhem’s problem)

3. Perform a metamethodological critique (novel evidence, avoidance of double counting, varying evidence, randomized control trials)

(These are of course interrelated)

Probabilistic methods enter for all these goals, but its role can be thought to measure:

1. degrees or probability or belief in a hypothesis or claim (a priori, subjective)

2. how reliably a hypothesis was tested, how well-tested, corroborated or severely probed it is (objective-empirical, frequentist)

A first reason my work might not be seen to fall under formal epistemology is that where formal epistemologists tend to look to the first (sometimes both); whereas, I look only to the second.

To infer hypotheses that have been highly probed—very different from saying they are highly probable

The appeal to probability to assess well-testedness while less common in philosophy does have a pedigree (Peirce, Popper, Braithwaite):

Charles Peirce:

"The theory here proposed does not assign any probability to the inductive or hypothetic conclusion, in the sense of undertaking to say how frequently *that conclusion* would be found true. It does not propose to look through all the possible universes, and say in what proportion of them a certain uniformity occurs; such a proceeding, were it possible, would be quite idle. The theory here only says how frequently, in this universe, the special form of induction or hypothesis would lead us right. The probability given by this theory is in every way different—in meaning, numerical value, and form—from that of those who would apply to ampliative inference the doctrine of inverse chances.[e.g., Bayesians]" (Peirce, CP, 2.748)

… in the case of analytic inference we know the probability of our conclusion (if the premises are true), but in the case of synthetic inferences we only know the degree of trustworthiness of our proceeding. (2.693)

Peirce's Self-Correcting Rationale for Inductive Methods:

- The justification for statistical methods is to learn from error: error correcting
- Not getting closer to the truth but literally learning from standard types of mistakes
- The goal is not trying to measure strength of subjective belief, but avoiding being misled by beliefs—mine and yours
- Peirce is perhaps the one philosopher who defined <u>induction</u> as the experimental test of a theory and the inference of whatever was severely tested
- His recognition: Induction (understood as severe testing) is better at correcting its errors than deduction
- Requires two things: *predesignation* (take account of alternations to severity from data dependent methods)
- *Random sampling* (more generally, establishing a sound probabilistic connection between the data and some underlying experimental phenomena)

Empirical but Objective/Normative

Severity Requirement (weakest): An agreement between data **x** and *H* fails to count as evidence for a hypothesis or claim *H* if the test would (with high probability certainly) yield so good an agreement even if *H* is false.

- Because such a test procedure had little or no ability to find flaws in *H*, finding none scarcely counts in *H*'s favor.
- The severity requirement reflects a central intuition about evidence, I do not regard it as some kind of primitive: it can be justified in terms of learning goals
- To flout it would not only permit being wrong with high probability—*a long-run behavior rationale* (necessary but not sufficient).
- In any particular case, little has been done to rule out the ways agreement between data and hypothesis can come about, even where the hypothesis is false
- One can get considerable mileage even with weak severity I also accept:

Severity Principle (Full): Data $\mathbf{x_0}$ provide a good indication of or evidence for hypothesis *H* (just) to the extent that test *T* severely passes *H* with $\mathbf{x_0}$.

Reason (confession?) #2: ***I deny that formal decision theory is a good place to look for knowledge/evidence/inference—epistemology***

- We want to know what the data are saying, what is warranted to infer from evidence, which is very different from what is best to do, from costs/benefits/risks.
- We want to apply evidence obviously, but by mixing together what the data are saying, with how much I stand to lose if the side effects of this drug are known is a mistake.
- In learning contexts, we do not want to have to delineate all of the possible claims we might wish to evaluate on the basis of data (beyond broad directions of alternative answers to a given problem).
- Whether people are or ought to maximize expected utility in decision contexts, it does not well model the appraisal of evidence in science.
- (Even where evidence is immediately feeding into policy, I distinguish "acceptable evidence" from "Acceptable Evidence" (coedited with R. Hollander)).

Reason #3: Ascertainability/applicability: An adequate epistemology (of statistics or other) would, as I see it, needs to be applicable:

- Rational reconstructions alone will not do.

“The idea of putting probabilities over hypotheses delivered to philosophy a godsend and an entire package of superficiality” (Glymour 2010, 334).

- Plausible principles of evidence might be supposed to be well captured until one asks if the computational components are ascertainable with the kinds of information scientists have in a learning context.
- If an account requires an exhaustive set of hypotheses and probability assignments to each, this would be relevant to its assessment as an adequate way to capture how we actually get knowledge).
- Most importantly, one forfeits the use of statistical and other formal tools to solve problems epistemologists care about.
- To be fair, analytic epistemology has always aimed to give a conceptual analysis, e.g., S knows that *H* iff *H* is true, obtained reliably, and whatever it takes to avoid Gettier problems.
- If formal epistemology is a way to do analytic epistemology using probabilistic notions, then my approach would not fall under a formal epistemology.

Reason #4: Relevance to Foundational Problems in Practice.

- A frequentist error statistical epistemologist would appeal to formal concepts/methods from error statistics in modeling inference and solving philosophical problems, but I could not do this without solving the foundational problems in practice
- Frequentist methods have long faced foundational criticisms going back to the late 1930s and I have worked to solve them over many years
- Ironically, philosophers of science regularly mixed with statisticians in the 80’s and early 90’s
- Coincidentally, in the late 90’s traditional Bayesian foundations began to rupture

- Some (Bayesians!) even suggest that “the idea of Bayesian inference as inductive, culminating in the computation of the posterior probability of scientific hypotheses has had malign effects on statistical practice” (Gelman 2010).
- “I cannot remember ever seeing a non-trivial Bayesian analysis which actually proceeded according to the usual Bayes formalism (Goldstein 2006, *Bayesian Analysis*).”
- Frequentist methods did not disappear (as some predicted).
- “The (arguably correct) view that science should embrace subjective statistics falls on deaf ears; they come to statistics in large part because they wish it to provide objective validation of their science.” (p. 388).
- Subjective elicitation of priors is mostly abandoned as unreliable or even impossible, and has been largely replaced by “nonsubjective” priors.
- Having conceded that “there is no such thing as uninformative priors” (Bernardo) the default priors are “references” for computations whose interpretation is unclear.

- While the key role for the prior was formally to incorporate an agent's degree of belief in hypotheses, apart from the data statistical model, the nonsubjective priors are model-dependent, are not even intended to represent beliefs (they are often not even probabilities.)

- Since in practice it is toward reference or default Bayesian priors that many look (Kass and Wasserman 1996) any philosophical problems these face would seem to be relevant to Bayesian epistemology.

If none of this matters to formal epistemology then my conception, again, would not fall under a formal epistemology

It seems to me that Bayesian epistemologists should help untangle these foundational problems in Bayesian practice…

Or discovering much practice is actually disinterring frequentist roots, try out an error statistical epistemology.

Despite all these reasons, it would seem possible to erect a “formal epistemology” based on frequentist error statistics

An Error Statistical Epistemology (parallel to a Bayesian epistemology) that reflects the learning from error idea; is ascertainable and applicable to practice; uses probability to capture how severely and inseverely claims are passed; etc.

My hope is to convince some of you to try and pursue this….

Epistemology of Error Statistics

We conceive of statistics very broadly: the conglomeration of tools for collecting, modeling and drawing inferences from data, including purely ‘data analytic’ methods

Several different statistical hypotheses are used to link formal error statistical tools and data,

Models of:

- ***primary scientific hypotheses***
- ***experimental model***
- ***data models***

A **'secondary' set of hypotheses** arise to check assumptions of other models in the network.

Statistical hypotheses offer ways to couch conjectured flaws in inference:

- mistaking spurious from genuine correlations, mistaken directions of effects,
- mistaken values of parameters,
- mistakes about causal factors,
- mistakes about assumptions of statistical models
- mistakes in linking statistical and substantive models

Knowledge goal: to find out what is (truly) the case about aspects of phenomena, as modeled, but *the hypotheses erected in the actual processes of finding things out are generally approximations and may even be deliberately false.*

[Probing these errors turns out to parallel the Coxian taxonomy of types of null hypotheses.]

Error Statistical Methods

Probability arises (in inference) to quantify how frequently methods are capable of discriminating between alternative hypotheses and how reliably they detect errors.

These probabilistic properties of procedures are *error frequencies* or *error probabilities*

Questions are addressed by means of hypotheses framed within a statistical model giving the probability distribution (or density) of **X**,

- an approximate or idealized representation of the data generating process.
- statistical hypotheses are typically in terms of unknown parameter θ which governs the probability distribution (or density) of **X**.
- Hypotheses assign probabilities to various outcomes "computed under the supposition that H_i is correct": P(**x**; H_i).
  or P(**x**; θ)

(not a conditional probability, that would assume there is a prior probability for H_i)

- Fisherian tests, Neyman-Pearson tests and confidence intervals (perhaps others)

Distance function: $d(\boldsymbol{X})$, the *test statistic*, reflects how well the data $\mathbf{x}_0 = (x_1, \ldots, x_n)$ fit the hypothesis H_0 —

the larger the value of $d(\mathbf{x}_0)$ the further the outcome is from what is expected under H_0 in the direction of alternative H_1 , with respect to the particular question being asked.

(Note: **X** is a random variable, and $\mathbf{x}_0$ is a fixed value of **X**; bolding the random variable indicates it is a vector.)

Adequate tests:
need to be able to calculate the probability of $\{d(\mathbf{X}) > \mathrm{d}(\mathbf{x}_0)\}$ under the assumption that H_0 is adequate

as well as under various discrepancies from H_0 is contained in the composite alternative $H_{1.}$

p-value: $P(d(\mathbf{X}) > d(\mathbf{x}_0); H_0)$

The hypotheses exhaust the space of parameter values.

Key: The probabilities derived from the modeled phenomena are equal to or are close to the actual relative frequencies of results in applying the method.

In a little known article (“**The Problem of Inductive Inference”** Neyman, 1955) that I found in my attic, Jerzy Neyman responds to Carnap’s (critical) depiction of “Neyman’s frequentism”:

> “When Professor Carnap* criticizes some attitudes which he represents as consistent with my (“frequentist”) point of view, I readily join him in his criticism without, however, accepting the responsibility for the criticized paragraphs”. (p. 13)
>
> (*in *Logical Foundations of Probability*, 1950)

“I am concerned with the term ‘degree of confirmation’ introduced by Carnap. …We have seen that the application of the locally best one-sided test to the data…failed to reject the hypothesis [that the 26 observations come from a source in which the null hypothesis is true]. The question is: does this result ‘confirm’ the hypothesis that H_0 is true of the particular data set]? ” (pp. 40-1).

Best one-sided Test T (embedded null)

A sample $\mathbf{X} = (X_1, \ldots, X_n)$ each X_i is Normal, $N(\mu,\sigma^2)$, (NIID), σ assumed known;

$$H_0\text{: } \mu \leq \mu_0 \text{ against } H_1\text{: } \mu > \mu_0\text{, say } \mu_0=0,$$

test statistic $d(\mathbf{X})$ is the sample mean $\bar{x}$

standardized $d^*(\mathbf{X}) = (\bar{x}-\mu_0)/\sigma_{\mathbf{x}}$,

Test result is not statistically significantly great enough to "reject" the null, $d^*(\mathbf{x}_0) \leq c_\alpha$

α is the significance level

p- value: $P(d^*(\boldsymbol{X}_0) > d^*(\mathbf{x}_0) \text{ ; } H_0)$

-probability of finding small p-value when should, less than when shouldn't:

$P(d^*(\boldsymbol{X}_0) > d^*(\mathbf{x}_0) \text{ ; } H_0) < P(d^*(\boldsymbol{X}_0) > d^*(\mathbf{x}_0) \text{ ; } H_i)$ alternatives H_i

Neyman continues:

"The attitude described (taking the non-significant difference as confirmation for the null) is dangerous.

…the chance of detecting the presence [of discrepancy δ from the null], when only [this number] of observations are available, is extremely slim, even if [δ is present].

"One may be confident in the absence of the discrepancy only if the power to detect it were high".

(1) $P(d(X) > c_\alpha; \mu = \mu_0 + \delta)$ Power to detect δ

- Just missing the cut-off c_α is the worst case
- It is more informative to look at the probability of getting a worse fit than you did

(2) $P(d(X) > d(x_0); \mu = \mu_0 + \delta)$ "attained power"

- a measure of the **severity** (or degree of corroboration) the inference $\mu < \mu_0 + \delta$

(1) $P(d(X) > c_\alpha; \mu = \mu_0 + \delta)$ Power to detect δ

(2) $P(d(X) > d(x_0); \mu = \mu_0 + \delta)$ "attained power"

Although (1) may be low, (2) may be high.

(2) could also be made out viewing the p-value as a random variable, calculating its distribution for various alternatives (Cox 2006, p. 25).

> Neyman's critique of Carnap's confirmation goes toward avoiding common fallacious construals of insignificant results (e.g., in risk assessment)
>
> "No evidence of increased risk is not evidence of no increased risk"

Frequentist Evidence (FEV) (Mayo and Cox 2010, p. 256):

FEV(a) A moderate p-value is evidence of the absence of a discrepancy δ from H_0 only if there is a high probability the test would have given a worse fit with H_0 (i.e., smaller p value) were a δ to exist.

FEV(b) **x** is evidence of a discrepancy δ from H_0 iff H_0 were a correct description of the mechanism generating **x**, then, with high probability a less discordant result would have occurred.

It is important to see that it is only in the case of a negative result that severity for various inferences is in the same direction as power.

In the case of significant results, d($\mathbf{x}_0$) in excess of the cut-off, the opposite concern arises—namely the test is too sensitive.

Severity is always relative to a statistical model of experiment E, a test couched within model E, outcomes of the test, and a hypothesis or claim to be considered (as possibly warranted or not):

$$\text{SEV}(\text{E}, \mathbf{x}, H)$$

I am not claiming it is part of the N-P school,

I recommend moving away from the idea that we are to “sign up” for N-P or Fisherian “paradigms”

As a philosopher of statistics I see:

Our error statistical epistemologist would supply the tools with an interpretation and an associated philosophy of learning….

Error Statistical Philosophy

What is key on the statistics side: The probabilities refer to the distribution of statistic $d(\mathbf{X})$ (sampling distribution)

—probability applied to events

This alone is at odds with Bayesian methods where consideration of outcomes other than the one observed is disallowed (likelihood principle)

Neyman-Pearson hypothesis testing violates the likelihood principle, because the event either happens or does not; and hence has probability one or zero." (Kadane, 2011)

What is key on the philosophical side: *e*rror probabilities may* be used to quantify probativeness or severity of tests (for a given inference)

*they do not always or automatically give us this.

(2) For the vast majority of cases we deal with, satisfying the N-P (behavioristic) long-run desiderata, leads to a uniquely appropriate test, that simultaneously satisfies Cox's (Fisherian) focus on minimal sufficient statistics ~ SEV construal

The standpoint of the severe prober, or the severity principle, directs us to obtain error probabilities that are relevant to determining well-testedness.

If one wants a post-data measure, one can write:

SEV($\mu < \bar{x}_{\mathbf{0}} + \delta\sigma_x$) to abbreviate:

The severity with which test T+ with data $\mathbf{x}_0$ passes claim:

($\mu < \bar{x}_{\mathbf{0}} + \delta\sigma_x$).

One can consider a series of upper severity bounds…

$\mathrm{SEV}(\mu < \bar{x}_{\mathbf{0}} + 0\sigma_x) = .5$

$\mathrm{SEV}(\mu < \bar{x}_{\mathbf{0}} + .5\sigma_x) = .7$

$\mathrm{SEV}(\mu < \bar{x}_{\mathbf{0}} + 1\sigma_x) = .84$

$\mathrm{SEV}(\mu < \bar{x}_{\mathbf{0}} + 1.5\sigma_x) = .93$

$\mathrm{SEV}(\mu < \bar{x}_{\mathbf{0}} + 1.98\sigma_x) = .975$

RELATION TO CONFIDENCE INTERVALS?

It will be noticed these bounds are identical to the corresponding upper confidence interval bounds for estimating μ.

There's a duality between confidence intervals and tests: the confidence interval contains the parameter values that would not be rejected *by the test at the specified level of significance*.

It follows that the (1-α) one sided interval corresponding to test *T*+ is:

$$\mu > \bar{X} - c_{\alpha}(\sigma/\surd n).$$

It would not be the assertion of the upper bound, as in our interpretation of negative results.

The 97.5% CI estimator corresponding to test *T*+ is:

$$\mu > \bar{X} - 1.96(\sigma/\surd n).$$

We were led to the upper bounds in the midst of a severity interpretation of results.

Quick Note:

This is different from what I call the

Rubbing off construal: The procedure is rarely wrong, therefore, the probability it is wrong in this case is low.

Still too *behavioristic*

The long-run reliability of the rule is a necessary but not a sufficient condition to infer *H* (with severity)

The reasoning instead is counterfactual:

$$H: \quad \mu \leq \bar{x}_0 + 1.96\sigma_{\mathbf{x}}$$

passes severely because were this inference false,

and the true mean $\mu > \bar{x}_0 + 1.96\sigma_{\mathbf{x}}$

then, very probably, we would have observed a larger sample mean

But am I not just using this as another way to say how probable each claim is?

No. This would lead to inconsistencies, moreover, probability gives the wrong logic for "how well-tested" (or "corroborated") a claim is.

Probability can be used to assess severity, but SEV logic is not probability logic.

It may be that data ***x*** provide poor grounds for *H*, and also poor grounds for ~*H*
Other points— (e.g., problem of irrelevant conjunctions, tacking paradox)

If SEV(Test *E*, data ***x***, claim *H*) is high, but *J* is not probed in the least by the experiment *E*, then SEV (*E*, ***x***, *H* & *J*) = very low or minimal

H: GTR and *J*: Kuru is transmitted through funerary cannibalism,
and data $\mathbf{x}_0$ be a value of the observed deflection in accordance with GTR.

- The two hypotheses do not make reference to the same data models or experimental outcomes, so it would be odd to conjoin them.

Note: we must distinguish ***x*** severely passing *H* and *H* being severely passed on all evidence in science at a time

Similarly,

- SEV(GTR gravity in solar system) = high in 1960s
- SEV(all of GTR) = low

We can apply the severity idea because the condition "given *H* is false" (even within a larger theory) means "given *H* is false with respect to what T says about *this particular* effect or phenomenon"

The goal of severity underscores a fundamental strategy used to probe high-level theories—<u>partitioning theories</u> to severely probe given effects or given errors (e.g., different values of λ, the deflection of light)

- *We do exhaust the space, but of <u>relevant rivals</u> to H—they must be at "the same level" as H.*
- *We learn a lot from considering what has passed inseverely—the logic should reflect this.*

The claims to be considered in a general error statistical epistemology would go beyond the statistical ones

- We have already seen they would go beyond predesignated ones, but rather would consider discrepancies from them

But claims need not be limited even to these

- The strongest SEV arguments do not call for statistics at all, but we learn a lot from the way statistics self-consciously probes errors

However, features from the formal system obviously constrain the more general logic one might develop, so it is useful to see some differences (e.g. between posterior probabilities and error-statistical quantities like p-values).

Highly Probable vs. Highly Probed Hypotheses: P-values vs. Bayesian Posteriors

Certain choices for prior probabilities in the null and alternative hypothesis shows that a small p-value is consistent with a much higher posterior probability in null hypothesis,

- The alternative hypothesis would, in such cases, pass severely, even though the null hypothesis has a high posterior (Bayesian) probability.

(Two-sided) Test of a mean of a Normal Distribution, Test T_2

H_0: $\mu = 0$ vs. H_1: $\mu \neq 0$

The difference between p-values and posteriors is less marked with one-sided tests

A random sample $\mathbf{X} = (X_1, \ldots, X_n)$ is taken where each X_i is $N(0, \sigma^2)$.

H_0: there are no increased risks (or benefits) associated with HRT in women who have taken them for 10 years.

"If n = 50 one can classically 'reject H_0 at significance level p = .05,' although P(H_0|**x**) = .52 (which would actually indicate that the evidence favors H_0)." (Berger and Sellke, 1987, p. 113).

- An error statistical tester takes the low significance level as evidence for H_1 — the very weak claim that there is some non-zero difference in risk rates

		***n* (sample size)**				
p	***t***	***n=10***	***n=20***	***n=50***	***n=100***	***n=1000***
.10	**1.645**	**.47**	**.56**	**.65**	**.72**	**.89**
.05	**1.960**	**.37**	**.42**	**.52**	**.60**	**.82**
.01	**2.576**	**.14**	**.16**	**.22**	**.27**	**.53**
.001	**3.291**	**.024**	**.026**	**.034**	**.045**	**.124**

What warrants the prior?

(1) Some claim the prior of .5 is a warranted frequentist assignment:

- H_0 was randomly selected from an urn in which 50% are true
- (*) Therefore P(H_0) = p
- What should go in the urn of hypotheses?

In the context under consideration it is agreed that either H_0 is true or not about this "one universe." (Peirce: "universes are not as plentiful as raspberries drawn from an urn.")

My main point: even if we agreed that there was a 50%chance of randomly selecting a true null hypothesis from a given pool of nulls, .5 would still not give the error statistician a frequentist prior probability of the truth of this hypothesis.

Fallacy of Probabilistic Instantiation:

50% of the null hypotheses in a given pool of nulls are true.
This particular null hypothesis H_0 was randomly selected from this pool.
Therefore, P(H_0 is true) = .5.

(2) The Assumption of "Objective" Bayesian Priors

More commonly it is assumed that the "objective" Bayesian prior gives 0.5 to the null, the remaining 0.5 probability being spread out over the alternative parameter space.

- Many take this to be an “impartial” or “uninformative” Bayesian prior probability, as recommended by Jeffreys, 1939.

- Far from impartial, the “spiked concentration of belief in the null” gives high weight to the null and seems at odds with the role of null hypotheses in testing; indeed, it is commonly claimed that “all nulls are false”.

- The error statistical tester would balk at the fact that use of the recommended priors can result in highly significant results often being construed as no evidence against the null—or even evidence for it!

The Bayesian significance tester wishes to start with a fairly high prior to the null because otherwise a rejection of the null merely claims a fairly improbable hypothesis has become more improbable (Berger and Sellke, 1987, p. 115).

- By contrast, it is informative for an error statistical tester to reject a null, even assuming it is not precisely true, because we can learn how false it is.

- "Default" or reference Bayesians disagree here: some use the spiked prior of .5, others reject this.

- The default Bayesian position on tests seems to be in flux.

The Analogous Criticism Among the Bayesian Epistemologist:

These statistical facts are often exported to launch the analogous indictment of the severity account (e.g., Howson 1997, Achinstein 2010, 2011).

Achinstein (2010, p. 187) "My response to the probabilistic fallacy charge is to say that it would be true if the probabilities in question were construed as relative frequencies. However, … I am concerned with epistemic probability."

So he grants:

p% of the hypotheses in a given pool of hypotheses are true (or a character holds for p%).

This particular hypothesis H_i was randomly selected from this pool.

Therefore, the objective epistemic probability P(H_i is true) = p.

We are to imagine a student, Isaac, who has taken a battery of tests and achieved very high scores, which is very rarely achieved by those who are not college-ready, and the hypothesis:

H(I): Isaac is college-ready.

H’(I) Isaac is not college-ready, i.e., he is deficient.

The probability for such good results, given a student is college-ready is extremely high, so that
$P(s\ |H(\mathrm{I}))$ is practically 1,

while very low, given that he is not college-ready

Imagine Isaac was randomly selected from Fewready Town – in which college readiness is rare, say one out of one thousand, he infers that

(┤) $P(H(\mathrm{I})) = .001$.

If so, then the posterior probability that Isaac is college-ready, given his high test results, would be very low

$p(H(\mathrm{I}) \mid s)$ is very low,

even though the posterior probability has increased from the prior in (┤).

To the error statistician, while the probability of a randomly selected high school student from Fewready Town is .001, *it does not follow that Isaac, the one we happened to select, has a probability of .001 of being college-ready.*

For Achinstein's epistemic probabilist it does.

The example considers some artificial ingredients, all of which I grant:

- only two outcomes: the high scores *s*, or lower scores, < *s*.
- only the two hypotheses, ready or not, whereas a proper statistical hypothesis would consider the extent of readiness

With only two outcomes: the high scores *s*, or lower scores, <*s*:
Clearly a lower grade gives even less evidence of readiness; that is,

$P(H'(\mathrm{I})|{<}\mathrm{s}) > P(H'(I)|\mathrm{s})$.

Therefore, whether Isaac scored as high as *s* or lower, Achinstein's epistemic probabilist would appear to report justified high belief that Isaac is not ready.

- Even if he claims he is merely blocking evidence for Isaac's readiness the analysis is open to problems: the probability of Achinstein finding evidence of Isaac's readiness even if in fact he is ready (*H* is true) is low if not zero.

- Erich Lehmann: the test has not done its job which was to discriminate…

Other Bayesians might interpret things differently, noting that since the posterior for readiness has increased, the test scores provide at least some evidence for *H*(I)---but then the example scarcely demonstrates a conflict between a frequentist and Bayesian assessment.

- <u>Fewdeficient Town:</u> If the epistemic probabilist learns that in fact Isaac was selected randomly, not from Fewready Town, but from a population where college readiness is common, Fewdeficient Town, the same score now warrants strong objective epistemic belief in Isaac's readiness.

- A high school student from Fewready Town would need to have scored quite a bit higher on these tests than one from Fewdeficient Town for his scores to be considered evidence for his readiness.

- When we move from hypotheses like "Isaac is college-ready" to generalizations, the difficulty for the epistemic probabilist becomes far worse if we are to obtain epistemic probabilities via the probabilistic instantiation rule.

- **We** need not preclude that *H*(I) has a legitimate frequentist prior; the frequentist probability that Isaac is college-ready might refer to genetic and environmental factors that determine the chance of his deficiency – although I do not have a clue how one might compute it.

- The main thing is that it is not given by the probabilistic instantiation.

- These examples invariably shift the meaning from one kind of experimental outcome—a randomly selected student has the property “college ready” —to another, a genetic and environmental “experiment” concerning Isaac and the outcomes are ready or not ready.

- This points out the flaw in trying to glean reasons for epistemic belief by means of just any conception of “low frequency of error.”

  If we declared “unready” for any member of Fewready, we would rarely be wrong, but in each case the “test” has failed to discriminate the particular student’s readiness from his un-readiness.

- Moreover, were we really interested in the probability that a student randomly selected from a town is college ready, and have the requisite probability model (e.g., Bernouilli), then there is nothing to stop the frequentist error statistician from inferring the conditional probability.

- However, there is nothing Bayesian in this relative frequency calculation.

CONCLUSION

An Error Statistical Epistemology

An *error statistical epistemology*: the application of error statistical tools and their interpretation to problems of knowledge:

- to model scientific inference (actual or rational),
- to scrutinize principles of inference (e.g., prefer novel results, double-counting, no evidence of risk is not evidence of no risk),
- to frame and tackle philosophical problems about evidence and inference (how to warrant data, pinpoint blame for anomalies, test models and theories).

I have used the severity account to address all of these, but the question of an overarching "severity logic" if there be one, is unanswered.

Severity Principle as Arguing From Error. Data $\mathbf{x}_0$ provides evidence for ruling out mistakes in regarding data as evidence for hypothesis *H* (just) to the extent test T would have detected the error, if present (but instead lets us continually generate data that accords with the absence of the error).

An adequate account must be able to assess, scrutinize and control error probabilities to this end.

Requires considering the procedures for generating data and specifying hypotheses to evaluate.

While the logic is complex, they may be organized systematically:

Inferences can err in the data-experiment-primary hypothesis triad if:

- the SEV assessment in the experimental model does not hold of the actual data,
- that claims pass inseverely,
- that the claim that passes severely is not the substantive one of interest

If it is granted that these have fundamental roles in learning about the world, our epistemologies should reflect this, formal or informal.

EXTRA

DOUBLE COUNTING/PREFERENCE FOR USE-NOVELTY

Severity criteria provides niches to pick up on “selection effects” and double-counting.

It seems clear that if one is allowed to search through several factors and report just those that show (apparently) impressive correlations, there is a high probability of erroneously inferring a real correlation.

But, it is equally clear that we can reliably use the same data both to arrive at and warrant:

- Measured parameters (e.g., my weight gain in LA)
- The cause or source of a fingerprint (e.g., a particular criminal)

It is the severity, stringency, or probativeness of the test—or lack of it—that should determine if a double-use of data is permissible

i.e., the rationale for use-novelty is severity.

Rather than a total prohibition, we need to adjust the SEV when “selection effects” alter the error probing capacity of the test.