Day #2 JULY 29

Quick Sum-up Day #1: Excursion 1

(full slides on a page on the Summer Seminar Blog)

- To get beyond today's statistics wars, we need to understand the jumble of philosophical, statistical, historical and other debates
- These are largely hidden in today's debates & the reforms put forward for restoring integrity
- There are ago-old: significance test controversy, Bayes-Frequentist battles
- Newer debates: reconciliations (default vs. subjective vs. empirical) Bayesians
- To evaluate the consequences of reforms need to excavate the jungle

- As impossible as it seems, I set out to do this
- I begin with a simple tool, underwritten by today's handwringing: the minimal requirement for evidence

(evidence for *C* only if it has been subjected to and passes a test it probably would have failed if false)

- Biasing selection effects make it easy to find impressive-looking effects erroneously
- They alter a method's error probing capacities
- They may not alter evidence (in *traditional* probabilisms): Likelihood Principle
- Members of different tribes talk past each other
- To the LP holder: worry about what could have happened but didn't is to consider "imaginary data" and "intentions"

- To the severe tester, probabilists are robbed from a main way to block spurious results
- Constructive role of replication crisis:
  Biasing selection effects *impinge on* error probabilities
  Error probabilities *impinge on* well-testedness
- Goes beyond 2 main ways of using probability: Performance and Probabilism
  Probativism (that's what we mean by viewing statistical inference as severe testing):
- It directs the reinterpretation of significance tests and other methods
- Probabilists may block inferences without appeal to error probabilities: high prior to H_0 (no effect) can result in a high posterior probability to H_0

- Gives a life-raft to the P-hacker and cherry picker; puts blame in the wrong place
- Severe probing (formal or informal) must take place at every level: from data to statistical hypothesis; from there to substantive claims
- A silver lining to distinguishing highly probable and highly probed–can use different methods for different contexts
- Some Bayesian tribes may find their foundations in error statistics
- Last excursion: (probabilist) foundations lost; (probative) foundations found

Excursion 2: Taboos of Induction and Falsification

Tour I: Induction and Confirmation (p. 59)

Contemporary philosophy of science presents us with some taboos: Thou shalt not try to find solutions to problems of induction, falsification, and demarcating science from pseudoscience.

It's impossible to understand rival statistical accounts, let alone get beyond the statistics wars, without first exploring how these came to be "lost causes".

Museum of StatSci/PhilSci

- Tour I begins with the traditional Problem of Induction, then to Carnapian confirmation and a brief look at contemporary formal epistemology.

- Tour II visits Popper, falsification and demarcation, moving into Fisherian tests and the replication crisis.

Having conceded loss in the battle for justifying induction, philosophers appeal to logic to capture scientific method

Inductive Logics	Logic of falsification
"Confirmation Theory" Rules to assign degrees of probability or confirmation to hypotheses given evidence e	Methodological falsification Rules to decide when to "prefer" or accept hypotheses
Carnap C(H,e) Inductive Logicians	**Popper** Deductive Testers

Inductive Logicians	**Deductive Testers**
we can build and try to justify "inductive logics" straight rule: Assign degrees of confirmation/credibility	we can reject induction and uphold the "rationality" of preferring or accepting H if it is "well tested"
Statistical affinity	*Statistical affinity*
Bayesian (and likelihoodist) accounts	Fisherian, Neyman-Pearson methods: probability enters to ensure reliability and severity of tests with these tests.

2.1 Traditional Problem of Induction

Inductive argument. With an inductive argument, the conclusion goes beyond the premises. So it's logically possible for all the premises to be true and the conclusion false: invalid.

The traditional problem of induction seeks to justify *enumerative induction (EI)* (*straight rule* of induction).

Infer from past cases of A's that were B's to all or most A's will be B's:

(EI) All observed A_1, A_2, ...A_n have been B's
Therefore, *H*: all A's are B's.

The premises might be experimental outcomes or data points:

X_1,X_2,…X_n (e.g., light deflection observations, drug reactions, radiation levels in fish)

Even if all observed cases had a property, or followed a law, there's no logical contradiction in the falsity of the generalization:

H: All *E*'s are *F*

This is also true for a statistical generalization:

90% in this class have property Q

Thus, *H*: 90% of all people have property Q.

***H* agrees with the data, but it's possible to have such good agreement even if *H* is false.**

(Asymmetry with falsification)

Exhibit (i). *Justifying Induction is Circular.* The traditional problem of induction is to justify the conclusion:

Conclusion: (EI) is rationally justified, it's a reliable rule.

We need an argument for concluding (EI) is reliable.

That's what is meant by "rationally" justifying it.

…….

_____

Conclusion: The 'inductive method' (EI) is reliable (it will work in the future): inferring from past to future success is a reliable method.

Little Bit of Logic

Argument:

A group of statements, one of which (the conclusion) is claimed to follow from one or more others (the premises), which are regarded as supplying evidence for the truth of that one.

This is written:

$$P_1, P_2, \ldots P_n / \therefore C.$$

In a 2-value logic, any statement *A* is regarded as true or false.

A deductively valid argument: *if* the premises are all true *then*, necessarily, the conclusion is true.

To use the "⊨" (double turnstile) symbol:

$$P_1, P_2, \ldots P_n \models C.$$

Note: Deductive validity is a *matter of form*—any argument with the same form or pattern as a valid argument is also a valid argument.

(Simple truth tables provide an algorithm or computable method to determine validity; no knowledge of context needed)

EXAMPLES (listing premises followed by the conclusion)

Modus Ponens	**Modus Tollens**
If *H* then *E*	If *H* then *E*
H	Not-*E*
∴ *E*	∴ Not-*H*

If (*H*) GTR, then (*E*) deflection effect.

(not-*E*) No light deflection observed.

∴ (not-*H*) GTR **is false**

(falsification)

These results depend on the English meaning of "if then" and of "not".

Disjunctive syllogism:

(1) Either the (A) experiment is flawed or (B) GTR is false
(2) GTR is true (i.e., not-B).
Conclusion: Therefore, (A) experiment is flawed.

If it's either *A* or *B*, and it's not *A*, then it must be *B*. Since if it's not *B*, you can't also hold the two premises—without contradiction. (soup or salad)

Either A or B. (disjunction)
Not-B
Therefore, A

Deductively Valid Argument (argument form)*:*

Three equivalent definitions:

- An argument where if the premises are all true, then necessarily, the conclusion is true. (i.e., if the conclusion is false, then (necessarily) one of the premises is false.)
- An argument where it's (logically) impossible for the premises to be all true and the conclusion false. (i.e., to have the conclusion false with the premises true leads to a logical contradiction: *A* & not-*A*.)
- An argument that maps true premises into a true conclusion with 100% reliability. (i.e., if the premises are all true, then 100% of the time the conclusion is true).

The categorical syllogism version of modus ponens

In the above, A and B are statements, here it's a property (but arguments in the real world blend them)

All A's are B's.
x is an A. Therefore x is a B.

All swans are white x is a swan.
Therefore, x is white.

True or False?

If an argument is deductively valid, then its conclusion must be true.

Here's an instance of the valid form:

All philosophers can fly.
Mayo is a philosopher.
Therefore Mayo can fly.

To detach the conclusion of a deductively valid argument as true, the premises must be true.

(Deductively) **Sound argument**: deductively valid + premises are true/approximately true.

<u>Invalid Argument</u>: Consider this argument:

If *H* then *E*
E
∴ *H*

If all numbers are even then 4 is even. 4 is even.
Therefore all numbers are even.

Affirming the consequent.

Invalid argument: An argument where it's possible to have all true premises and a false conclusion without contradiction.

Invalid arguments let us go from all true premises to false conclusions.

If the premises are contradictory is the argument valid or invalid? (pertains to next argument):

If (*H*) all swans are white, then the next swan I see will be white.

The next swan I see x is black.

Therefore (not-*H*) Not all swans are white (falsification)

(prefer the above to ex (1) on p. 60)

(H) all swans are white.

<u>The next swan I see x is black.</u>

Therefore (not-*H*) Not all swans are white

Logic of Simple Significance Tests: Statistical Modus Tollens

(Statistical analogy to the deductively valid pattern *modus tollens*)

If the hypothesis H_0 is correct then, with high probability, 1-p, the data would *not* be statistically significant at level p.

x_0 is statistically significant at level p.

Thus, x_0 is evidence against H_0, or x_0 indicates the falsity of H_0.

Going back to where we left off:

Exhibit (i). *Justifying Induction is Circular.* The traditional problem of induction is to justify the conclusion:

Conclusion: (EI) is rationally justified, it's a reliable rule.

We need an argument for concluding (EI) is reliable.

…….

Conclusion: The 'inductive method' (EI) is reliable (it will work in the future): inferring from past to future success is a reliable method.

What will the premises be?

a. Use an inductive argument to justify induction:

Premise: The inductive method has worked, has been reliable, in the past.

Conclusion: The inductive method is reliable (it will work in the future), i.e., inferring from past cases to future cases is a reliable method.

Problem: circular. (It uses the method in need of justification to justify that method.)

b. Use a deductive argument to justify induction:

Premise: If a method has worked (been reliable) in the past, then it will work in the future.

Premise: The inductive method has worked, has been reliable, in the past.

Conclusion: the inductive method is reliable (it will work in the future), i.e., inferring from past cases to future cases is a reliable method.

Problem*: we cannot say it's a sound argument.*

In order to infer the truth of the conclusion of a deductively valid argument, the premises must be true, i.e., the argument must be sound.

You'd have to know the very thing which the argument was supposed to justify!

We need to deductively infer EI will be reliable *in general*: the known cases only refer to the past and present, not the future.

Alternatively put in terms of assuming the **uniformity of nature**

Logical problem of induction: can't use logic to solve it.

Attempts to dissolve the problem (not in SIST)

- It's asking for justification beyond where it's appropriate,
- It's converting induction to deduction (i.e., it's asking for a certainly true conclusion from true premises)
- That's just what we mean by rational.

To see what's wrong with the last, consider my friend the crystal gazer…

Counterinductive method:

Infer from All A's have been B's in the past to the next A will not be a B.

In terms of a method:

Infer from the fact that a method M has worked poorly (been unreliable) in the past that M will work well in the future. (The crystal gazing)

If a method M has worked poorly (been unreliable) in the past then M will work well in the future.

M has worked poorly (been unreliable) in the past

Therefore, M will work well in the future.

So, unless we allow this justification of counterinduction, we should not allow the appeal to 'that's what we mean by rational'

Some argue that although an attempted justification is circular it is not *viciously* circular. (An excellent source is Skyrms 1986.)

You might say, you just can't argue with someone who accepts counterinduction as rational.

You can't convince them that induction is superior to counterinduction, but all we care about is showing it's rational *for us to accept scientific induction*

Exhibit (ii). *Probabilistic Affirming the Consequent.* Enter logics of confirmation.
They didn't renounce enumerative induction (EI), they sought logics that embodied it (EI):

If *H* (all A's are B's), then all observed A's (A_1, A_2, ….A_n) are B's.
All observed A's (A_1, A_2, ….A_n) are B's
Therefore, *H*: all A's are B's.

This is *affirming the consequent*.

Probabilistic affirming the consequent says only the conclusion gets a boost in confirmation or probability–a *B-boost*.

It's in this sense that Bayes' Theorem is often taken to justify (EI): it embodies probabilistic affirming the consequent.
How do we obtain the probabilities?

Rudolf Carnap tried to deduce them from the logical structure of a particular (first order) language.

The degree of probability, a rational degree of belief, would hold between two statements, a hypothesis and the data.

C(*H*,***x***) symbolizes "the confirmation of *H*, given ***x***".

Once you chose the initial assignments to core states of the world, calculating degrees of confirmation is a formal or syntactical matter, much like deductive logic.

Goal: measure the *degree of implication* or confirmation that ***x*** affords *H*.
Where do the initial assignments come? (63-4)

> Carnap has stated that the ultimate justification of the axioms is inductive intuition. I do not consider this answer an adequate basis for a concept of rationality. Indeed, I think that *every* attempt, including those by Jaakko Hintikka and his students, to ground the concept of rational degree of belief in logical probability suffers from the same unacceptable *apriorism*. (Salmon 1988, p. 13).

What does a highly probable claim, according to a particular inductive logic, have to do with the real world? How can it provide "a guide to life?" (E.g., Kyburg 2003, Salmon 1966.)

The hankering for an inductive logic remains.

It's behind the appeal of the default Bayesianism of Harold Jeffreys, and other attempts to view probability theory as extending deductive logic.

Exhibit (iii). Hacking announced (1.4): "there is no such thing as a logic of statistical inference" (1980, p. 145),

Not only did all attempts fail; he recognized the project is "founded on a false analogy with deductive logic" (ibid.). He follows Peirce:

> In the case of analytic [deductive] inference we know the probability of our conclusion (if the premises are true), but in the case of synthetic [inductive] inferences we only know the degree of trustworthiness of our proceeding (Peirce 2.693).

In ampliative, or inductive reasoning, the conclusion should go beyond the premises; probability enters to qualify the overall "trustworthiness" of the method.

My argument from coincidence to weight gain in (1.3) inferred *H*: I've gained at least 4 pounds

- The inference is qualified by the detailed data (group of weighings), and information on how capable the method is at blocking erroneous pronouncements of my weight.
- *What is being qualified probabilistically is the inferring or testing process.*
- By contrast, in a probability or confirmation logic what is generally detached is the probability of *H*, given data. It is a *probabilism*.

Take note of the quote by Fisher (p. 66)

> In deductive reasoning all knowledge obtainable is already latent in the postulates. Rigour is needed to prevent the successive inferences growing less and less accurate as we proceed. The conclusions are never more accurate than the data. In inductive reasoning we are performing part of the process by which new knowledge is created. The conclusions normally *grow more and more accurate* as more data are included. It should never be true, though it is still often said, that the conclusions are no more accurate than the data on which they are based. (Fisher 1935, p. 54; my emphasis)

2.2 *Is Probability a Good Measure of Confirmation?* *(p. 66)*

It is often assumed that the degree of confirmation of x by y must be the same as the (relative) probability of x given y, i.e., that C(x,y) = Pr(x,y). My first task is to show the inadequacy of this view. (Popper 1959, p. 396; Pr for P).

The most familiar (Bayesian) interpretation is this:

H is confirmed by ***x*** if ***x*** gives a boost to the probability of *H*, *incremental* confirmation.

The components of C(*H*,***x***) can be any statements, no reference to a probability model is required

There is typically a background variable k, so that ***x*** confirms *H* relative to k: to the extent that Pr(*H*|***x*** and k) > Pr(*H* and k).

I drop the explicit inclusion of k.

If H entails $\boldsymbol{x}$, then assuming $\Pr(\boldsymbol{x}) \neq 1$, $\Pr(H) \neq 0$, we have $\Pr(H|\boldsymbol{x}) > \Pr(H)$.

This is an instance of probabilistic affirming the consequent. (Note: if $\Pr(H|\boldsymbol{x}) > \Pr(H)$ then $\Pr(\boldsymbol{x}|H) > \Pr(\boldsymbol{x})$. Note 4, p. 69)

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

Simple way to see this (viewing H and $\boldsymbol{x}$ as statements*)

Pr(first & second) = Pr(first)Pr(second|first)

$\Pr(\boldsymbol{x} \,\&\, H) = \Pr(\boldsymbol{x})\ \Pr(H|\boldsymbol{x})$

$\Pr(H \,\&\, \boldsymbol{x}) = \Pr(H)\ \Pr(\boldsymbol{x}|H)$

& is commutative, so $\Pr(\boldsymbol{x} \,\&\, H) = \Pr(H \,\&\, \boldsymbol{x})$

Thus

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

Divide both sides by $\Pr(\boldsymbol{x})\Pr(H)$

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

(1) *Incremental* (B-boost)

H is confirmed by $\boldsymbol{x}$ iff Pr($H|\boldsymbol{x}$) > Pr(H),

H is disconfirmed iff Pr($H|\boldsymbol{x}$) < Pr(H).

(2) *Absolute*: *H* is confirmed by $\boldsymbol{x}$ iff Pr($H|\boldsymbol{x}$) is high, at least greater than Pr($\sim H|\boldsymbol{x}$).

Since Pr($\sim H|\boldsymbol{x}$) = 1 – Pr($H|\boldsymbol{x}$),

(2) is the same as defining $\boldsymbol{x}$ confirms *H*: Pr($H|\boldsymbol{x}$) > .5.

Airport alarm (p. 66)

From (1), ***x*** (the alarm) *dis*confirms the hypothesis *H*: the bag is clean, because its probability has gone down however slightly.

Yet from

(2) ***x*** confirms *H*: bag is clean, as Pr(*H*) is high to begin with.

At the very least, we must distinguish between an incremental and an absolute measure of confirmation for *H.*

Incremental confirmation as generally used in current Bayesian epistemology: Confirmation is a B-boost.

A simple B-boost would report the ratio R: $\frac{\Pr(H|x)}{\Pr(H)}$

What's your intuition?

Popper's first point: to identify confirmation and probability "C = Pr" leads to this type of conflict (incremental or absolute)

(you can read): Popper's example is a single toss of a fair die:

H: a 6 will occur; x: an even # occurs (2,4,6)
Pr(*H*) = 1/6, Pr(*x*) = ½.

The probability of *H* is increased by data *x*, from 1/6 to 1/3
However, Pr(*H*|*x*) < Pr(~*H*|*x*) (1/3 < 2/3)

So *H* is less well confirmed given *x* than is ~*H*, in the sense of (2).

Popper's second point is that "*the probability of a statement…simply does not express an appraisal of the severity of the tests a theory has passed, or of the manner in which it has passed these tests*" (p. 395).

Any H that entails $\boldsymbol{x}$ gets a B-boost, unless x already had probability 1, since $\Pr(\boldsymbol{x}|H) = 1$

Exhibit (iv). *Paradox of Irrelevant Conjunctions* (Glymour 1980) "tacking paradox" (69-70)

If ***x*** confirms *H*, then ***x*** also confirms (*H* & *J*), even if hypothesis *J* is just "tacked on" to *H*
(Fitelson 2002, Hawthorne and Fitelson 2004)

> *J* is an *irrelevant conjunct* to *H*, with respect to evidence ***x***: Pr(***x***|*J*) = Pr(***x***|*J* & *H*).

For instance, ***x*** might be radioastronomic data in support of:

> *H*: the GTR deflection of light effect is 1.75" and
> *J*: the radioactivity of the Fukushima water being dumped in the Pacific Ocean is within acceptable levels.

(A) If $\boldsymbol{x}$ confirms H, then $\boldsymbol{x}$ confirms (H & J), where Pr($\boldsymbol{x}$|H & J) = Pr($\boldsymbol{x}$|H) for any J consistent with H.

The reasoning is as follows:

(i) Pr($\boldsymbol{x}$|H)/Pr($\boldsymbol{x}$) > 1 ($\boldsymbol{x}$ Bayesian confirms H)

(ii) Pr($\boldsymbol{x}$|H & J) = Pr($\boldsymbol{x}$|H) (J's irrelevance is given)

Substituting (ii) into (i) we get [Pr($\boldsymbol{x}$|H & J)/Pr($\boldsymbol{x}$)]> 1

Therefore $\boldsymbol{x}$ Bayesian confirms (H & J)

In fact, the conjunction gets just as much of a boost as does H, if we measure confirmation with R: $\frac{\Pr(H|x)}{\Pr(H)}$

p. 70 typo in the denominator of R (terrible!)

However, granting confirmation is an incremental B-boost doesn't commit you to measuring it by R.

The conjunction (*H* & *J*) gets less of a confirmation boost than does *H* if we use a different measure of the boost.

But aren't they uncomfortable with (A), allowing (*H* & *J*) to be confirmed by ***x***?

I agree with Glymour, we're not happy with an account that tells us deflection of light data confirms that GTR is true and GTR and the radioactivity of the Fukushima water is within acceptable levels, while assuring us that ***x*** does not confirm the Fukushima water having acceptable levels of radiation? (70)

Glymour goes further:

It is plausible to hold what philosophers call the "special consequence" condition (p. 70):

If ***x*** confirms W, and W entails J, then ***x*** confirms J.

In particular, let W be (*H* & *J*),

(B) If ***x*** confirms (*H* & *J*), then ***x*** confirms *J*.

(B) gives the second piece of the argument.

From (A) and (B) we have, if ***x*** confirms *H*, then ***x*** confirms *J* for any irrelevant *J* consistent with *H* (neither *H* nor *J* have probabilities 0 or 1).

Most Bayesian epistemologists reject (B) (special consequence), so avoid this.

If confirmation is a B-boost, special consequence doesn't follow.

What Does the Severity Account Say? (71)

Our account of inference disembarked way back at (1): that ***x*** confirms *H* so long as Pr(*H*|***x***) > Pr(*H*).

We reject probabilistic affirming the consequent.

The simplest case has *H* entail ***x***, and ***x*** is observed.
(We assume the probabilities are well defined, and *H* doesn't already have probability 1.)

H gets a B-boost, but there are many other "explanations" of ***x***.

It's the same reason we reject the Law of Likelihood (LL).

Unless they have been probed, finding an *H* that fits ***x*** is not difficult to achieve even if *H* is false: *H* hasn't passed severely.

Now the confirmation theorist is only saying a B-boost suffices for some evidence.

To us, to have *any* evidence, or even the weaker notion of an "indication", requires a minimal threshold of severity be met.

How about tacking? As always, the error statistician needs to know the relevant properties of the test procedure

Just handing me the *H*'s, ***x***'s, and relative probabilities do not suffice.

The process of tacking–one form–is once you have an incrementally confirmed *H* with data ***x***, tack on any consistent *J* and announce "***x*** confirms (*H* & *J*)".

Let's allow that (*H* & *J*) fits or accords with ***x*** (since GTR entails or renders probable the deflection data ***x***).

But "(*H* & *J*) is confirmed by ***x***" has been subjected to a radically non-risky test.

Nothing has been done to measure the radioactivity of the Fukushima water being dumped into the ocean.

What They Call Confirmation We Call Mere "Fit" or "Accordance"

> In opposition to [the] inductivist attitude, I assert that C(*H*,***x***) must not be interpreted as the degree of corroboration of *H* by ***x***, unless ***x*** reports the results of *our sincere efforts to overthrow H*. The requirement of sincerity cannot be formalized–no more than the inductivist requirement that ***x*** must represent our total observational knowledge (Popper, 1959, p. 418, *H* for h; ***x*** for *e*.)

There are many other famous paradoxes of confirmation theory (e.g., the white shoe confirming all ravens are black and the grue paradox)

We laugh at them, but they often contain a puzzle of relevance for statistical practice.

There are two reasons the tacking paradox above is of relevance to us: There is a large-scale theory *T* that predicts ***x***, and we want to discern which portion of *T* to credit.

Severity says: do not credit those portions that could not have been found false, even if they're false. They are poorly tested.

Second, the question of measuring support with a Bayes boost or with posterior probability arises in Bayesian statistical inference

When you hear that what you want is some version of probabilism, be sure to ask if it's a boost (and if so which kind) or a posterior probability, a likelihood ratio or something else.

Now statisticians might say, we don't go around tacking on hypotheses like this.

True, the Bayesian epistemologist invites trouble by not clearly spelling out corresponding statistical models

They seek a formal logic, holding for statements about radiation, deflection, fish or whatnot: a mistake, I say.

We can have a general account for statistical inference, it just won't be purely formal.

(Popper never saw how to use statistics to cash out his intuitions about severity)

Statistical Foundations Need Philosophers of Statistics

> The idea of putting probabilities over hypotheses delivered to philosophy a godsend, an entire package of superficiality." (Glymour 2010, p. 334).

Given a formal epistemology, the next step is to use it to represent or justify intuitive principles of evidence (meta-methodology, meeting #1)

The problem to which Glymour is alluding is: you can start with the principle you want your confirmation logic to reflect, and then *reconstruct* it using probability.