Excursion 2 Taboos of Induction and Falsification

Itinerary

Tour I Induction and Confirmation

> Cox: [I]n some fields foundations do not seem very important, but we both think that foundations of statistical inference are important; why do you think that is?
>
> Mayo: I think because they ask about fundamental questions of evidence, inference, and probability . . . we invariably cross into philosophical questions about empirical knowledge and inductive inference. (Cox and Mayo 2011, p. 103)

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." I am not talking of ancient history here: these problems were alive and well when I set out to do philosophy in the 1980s. I think we gave up on them too easily, and by the end of Excursion 2 you'll see why. Excursion 2 takes us into the land of "Statistical Science and Philosophy of Science" (StatSci/PhilSci). Our Museum Guide gives a terse thumbnail sketch of Tour I. Here's a useful excerpt:

> Once the Problem of Induction was deemed to admit of no satisfactory, non-circular solutions (~1970s), philosophers of science turned to building formal logics of induction using the deductive calculus of probabilities, often called Confirmation Logics or Theories. A leader of this Confirmation Theory movement was Rudolf Carnap. A distinct program, led by Karl Popper, denies there is a logic of induction, and focuses on Testing and Falsification of theories by data. At best a theory may be accepted or corroborated if it fails to be falsified by a severe test. The two programs have analogues to distinct methodologies in statistics: Confirmation theory is to Bayesianism as Testing and Falsification are to Fisher and Neyman–Pearson.

Tour I begins with the traditional Problem of Induction, then moves to Carnapian confirmation and takes a brief look at contemporary formal epistemology. Tour II visits Popper, falsification, and demarcation, moving into Fisherian tests and the replication crisis. Redolent of Frank Lloyd Wright's Guggenheim Museum in New York City, the StatSci/PhilSci Museum is

arranged in concentric sloping oval floors that narrow as you go up. It's as if we're in a three-dimensional Normal curve. We begin in a large exposition on the ground floor. Those who start on the upper floors forfeit a central Rosetta Stone to decipher today's statistical debates.

2.1 The Traditional Problem of Induction

Start with the *asymmetry of falsification and confirmation*. One black swan falsifies the universal claim that *C*: all swans are white. Observing a single white swan, while a *positive instance* of *C*, wouldn't allow inferring generalization *C*, unless there was only one swan in the entire population. If the generalization refers to an infinite number of cases, as most people would say about scientific theories and laws, then no matter how many positive instances observed, you couldn't infer it with certainty. It's always possible there's a black swan out there, a *negative instance*, and it would only take one to falsify *C*. But surely we think enough positive instances of the right kind might warrant an argument for inferring *C*. Enter the problem of induction. First, a bit about arguments.

Soundness versus Validity

An *argument* is a group of statements, one of which is said to follow from or be supported by the others. The others are premises, the one inferred, the conclusion. A deductively *valid* argument is one where if its premises are all true, then its conclusion must be true. Falsification of "all swans are white" follows a deductively valid argument. Let ~*C* be the denial of claim *C*.

(1) *C*: All swans are white.
x is a swan but is black.
Therefore, ~*C*.

We can also infer, validly, what follows if a generalization *C* is true.

(2) *C*: All swans are white.
x is a swan.
Therefore, *x* is white.

However, validity is not the same thing as *soundness*. Here's a case of argument form (2):

(3) All philosophers can fly.
Mayo is a philosopher.
Therefore, Mayo can fly.

Validity is a matter of form. Since (3) has a valid form, it is a valid argument. But its conclusion is false! That's because it is *unsound*: at least one of its premises is false (the first). No one can stop you from applying deductively valid arguments, regardless of your statistical account. Don't assume you will get truth thereby. Bayes' Theorem can occur in a valid argument, within a formal system of probability:

(4) If $\Pr(H_1), \ldots, \Pr(H_n)$ are the prior probabilities of an exhaustive set of hypotheses, and $\Pr(\boldsymbol{x}|H_i)$ the corresponding likelihoods.
Data $\boldsymbol{x}$ are given, and $\Pr(H_1|\boldsymbol{x})$ is defined.
Therefore, $\Pr(H_1|\boldsymbol{x}) = p$.[1]

The conclusion is the posterior probability $\Pr(H_1|\boldsymbol{x})$. It can be inferred only if the argument is *sound*: all the givens must hold (at least approximately). To deny that all of statistical inference is reducible to Bayes' Theorem is not to preclude your using this or any other deductive argument. What you need to be concerned about is their soundness. So, you will still need a way to vouchsafe the premises.

Now to the traditional philosophical problem of induction. What is it? Why has confusion about induction and the threat of the traditional or "logical" problem of induction made some people afraid to dare use the "I" word? The traditional problem of induction seeks to justify a type of argument: one taking a form of *enumerative induction* (EI) (or the *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 $A_1, A_2, \ldots, A_n$ have been B's.
Therefore, *H*: all A's are B's.

It is not a deductively valid argument, because clearly its premises can all be true while its conclusion false. It's *invalid*, as is so for any inductive argument. As Hume (1739) notes, nothing changes if we place the word "probably" in front of the conclusion: it is justified to infer from past A's being B's that, *probably*, all or most A's will be B's. To "rationally" justify induction is to supply a reasoned argument for using EI. The traditional problem of induction, then, involves trying to find an argument to justify a type of argument!

Exhibit (i): Justifying Induction Is Circular. In other words, the traditional problem of induction is to justify the conclusion:

[1] i.e., $p = \dfrac{\Pr(\boldsymbol{x}|H_1)\Pr(H_1)}{\Pr(\boldsymbol{x}|H_1)\Pr(H_1) + \cdots + \Pr(\boldsymbol{x}|H_n)\Pr(H_n)}$

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

We need an argument for concluding EI is reliable. Using an inductive argument to justify induction lands us in a circle. We'd be using the method we're trying to justify, or begging the question. What about a deductively valid argument? The premises would have to be things we know to be true, otherwise the argument would not be sound. We might try:

Premise 1: EI has been reliable in a set of observed cases.

Trouble is, this premise can't be used to deductively infer EI will be reliable *in general*: the known cases only refer to the past and present, not the future. Suppose we add a premise:

Premise 2: Methods that have worked in past cases will work in future cases.

Yet to assume Premise 2 is true is to use EI, and thus, again, to beg the question.

Another idea for the additional premise is in terms of assuming nature is uniform. We do not escape: to assume the *uniformity of nature* is to assume EI is a reliable method. Therefore, induction cannot be rationally justified. It is called the *logical* problem of induction because logical argument alone does not appear able to solve it. All attempts to justify EI assume past successes of a rule justify its general reliability, which is to assume EI – what we're trying to show.

I'm skimming past the rest of a large exhibition on brilliant attempts to solve induction in this form. Some argue that although an attempted justification is circular it is not *viciously* circular. (An excellent source is Skyrms 1986.)

But wait. Is inductive enumeration a rule that has been reliable even in the past? No. It is reasonable to expect that unobserved or future cases will be very different from the past, that apparent patterns are spurious, and that observed associations are not generalizable. We would only want to justify inferences of that form if we had done a good job ruling out the many ways we know we can be misled by such an inference. That's not the way confirmation theorists see it, or at least, saw it.

Exhibit (ii): Probabilistic (Statistical) Affirming the Consequent. Enter logics of confirmation. Conceding that we cannot justify the inductive method (EI), philosophers sought logics that represented apparently plausible inductive reasoning. The thinking is this: never mind trying to convince a skeptic of the inductive method, we give up on that. But we know what we mean. We need only to make sense of the habit of applying EI. True to the logical positivist spirit of the 1930s–1960s, they sought evidential relationships

between statements of evidence and conclusions. I sometimes call them evidential-relation (E-R) logics. They didn't renounce enumerative induction, they sought logics that embodied it. Begin by fleshing out the full argument behind 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.

The premise that we added, the first, is obviously true; the problem is that the second premise can be true while the conclusion false. The argument is deductively *invalid* – it even has a name: *affirming the consequent*. However, its probabilistic version is weaker. *Probabilistic affirming the consequent* says only that the conclusion is probable or gets a boost in confirmation or probability – a *B-boost*. It's in this sense that Bayes' Theorem is often taken to ground a plausible confirmation theory. It probabilistically justifies EI in that it embodies probabilistic affirming the consequent.

How do we obtain the probabilities? Rudolf Carnap's audacious program (1962) had been to assign probabilities of hypotheses or statements by deducing them from the logical structure of a particular (first order) language. These were called *logical probabilities*. The language would have a list of properties (e.g., "is a swan," "is white") and individuals or names (e.g., i, j, k). The task was to assign equal initial probability assignments to states of this mini world, from which we could deduce the probabilities of truth functional combinations. The degree of probability, usually understood as a rational degree of belief, would hold between two statements, one expressing a hypothesis and the other the data. $C(H,\boldsymbol{x})$ symbolizes "the confirmation of *H*, given $\boldsymbol{x}$." Once you have chosen the initial assignments to core states of the world, calculating degrees of confirmation is a formal or syntactical matter, much like deductive logic. The goal was to somehow measure the *degree of implication* or confirmation that $\boldsymbol{x}$ affords *H*. Carnap imagined the scientist coming to the inductive logician to have the rational degree of confirmation in *H* evaluated, given her evidence. (I'm serious.) Putting aside the difficulty of listing all properties of scientific interest, from where do the initial assignments come?

Carnap's first attempt at a C-function resulted in no learning! For a toy illustration, take a universe with three items, i, j, k, and a single property B. "Bk" expresses "k has property B." There are eight possibilities, each called a *state description*. Here's one: {Bi, ~Bj, ~Bk}. If each is given initial probability of ⅛, we have what Carnap called the logic $c^{\dagger}$. The degree of confirmation that j will be black given that i was white = ½, which is the same as the initial confirmation of Bi (since it occurs in four of eight state descriptions). Nothing

has been learned: $c^{\dagger}$ is scrapped. By apportioning initial probabilities more coarsely, one could learn, but there was an infinite continuum of inductive logics characterized by choosing the value of a parameter he called λ (λ continuum). λ in effect determines how much uniformity and regularity to expect. To restrict the field, Carnap had to postulate what he called "inductive intuitions." As a logic student, I too found these attempts tantalizing – until I walked into my first statistics class. I was also persuaded by philosopher Wesley Salmon:

> 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).

This program, still in its heyday in the 1980s, was part of a general logical positivist attempt to reduce science to observables plus logic (no metaphysics). Had this reductionist goal been realized, which it wasn't, the idea of scientific inference being reduced to particular predicted observations might have succeeded. Even with that observable restriction, the worry remained: 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 epistemology is restricted to inner coherence and consistency. However much contemporary philosophers have gotten beyond logical positivism, the hankering for an inductive logic remains. You could say it's behind the appeal of the default (non-subjective) Bayesianism of Harold Jeffreys, and other attempts to view probability theory as extending deductive logic.

Exhibit (iii): A Faulty Analogy Between Deduction and Induction. When we heard Hacking announce (Section 1.4): "there is no such thing as a logic of statistical inference" (1980, p. 145), it wasn't only the failed attempts to build one, but the recognition that the project is "founded on a false analogy with deductive logic" (ibid.). The issue here is subtle, and we'll revisit it through our journey. I agree with Hacking, who is agreeing with C. S. 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 getting new knowledge, in ampliative or inductive reasoning, the conclusion should go beyond the premises; probability enters to qualify the overall "trustworthiness" of the method. Hacking not only retracts his Law of Likelihood (LL), but also his earlier denial that Neyman–Pearson statistics is

inferential. "I now believe that Neyman, Peirce, and Braithwaite were on the right lines to follow in the analysis of inductive arguments" (Hacking 1980, p. 141). Let's adapt some of Hacking's excellent discussion.

When we speak of an inference, it could mean the entire argument including premises and conclusion. Or it could mean just the conclusion, or statement inferred. Let's use "inference" to mean the latter – the claim detached from the premises or data. A statistical procedure of *inferring* refers to a method for reaching a statistical inference about some aspect of the source of the data, together with its probabilistic properties: in particular, its capacities to avoid erroneous (and ensure non-erroneous) interpretations of data. These are the method's error probabilities. My argument from coincidence to weight gain (Section 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. I argue that, very probably, my scales would not produce the weight data they do (e.g., on objects with known weight) were *H* false. 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*. Hacking's diagnosis in 1980 is that this grows out of an abiding logical positivism, with which he admits to having been afflicted. There's this much analogy with deduction: In a deductively valid argument: if the premises are true then, necessarily, the conclusion is true. But we don't attach the "necessarily" to the conclusion. Instead it qualifies the entire argument. So mimicking deduction, why isn't the inductive task to qualify the method in some sense, for example, report that it would probably lead to true or approximately true conclusions? That would be to show the reliable performance of an inference method. If that's what an inductive method requires, then Neyman–Pearson tests, which afford good performance, are inductive.

My main difference from Hacking here is that I don't argue, as he seems to, that the warrant for the inference is that it stems from a method that very probably gets it right (so I may hope it is right this time). It's not that the method's reliability "rubs off" on this particular claim. I say inference *C* may be detached as *indicated* or *warranted*, having passed a severe test (a test that *C* probably would have failed, if false in a specified manner). This is the central point of Souvenir D. The logician's "semantic entailment" symbol, the double turnstile: "|=", could be used to abbreviate "entails severely":

$$\text{Data + capacities of scales } |=_{\text{SEV}} \text{ I've gained at least } k \text{ pounds.}$$

(The premises are on the left side of |=.) However, I won't use this notation.

Keeping to a deductive logic of probability, we never detach an inference. This is in sync with a probabilist such as Bruno de Finetti:

> *The calculus of probability can say absolutely nothing about reality* . . . As with the logic of certainty, the logic of the probable adds nothing of its own: it merely helps one to see the implications contained in what has gone before. (de Finetti 1974, p. 215)

These are some of the first clues we'll be collecting on a wide difference between statistical inference as a deductive logic of probability, and an inductive testing account sought by the error statistician. When it comes to inductive learning, we want our inferences to go beyond the data: we want lift-off. To my knowledge, Fisher is the only other writer on statistical inference, aside from Peirce, to emphasize this distinction.

> 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 1935b, p. 54)

2.2 Is Probability a Good Measure of Confirmation?

> 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; substituting Pr for P)

If your suitcase rings the alarm at an airport, this might slightly increase the probability of its containing a weapon, and slightly decrease the probability that it's clean. But the probability it contains a weapon is so small that the probability it's clean remains high, even if it makes the alarm go off. These facts illustrate a tension between two ways a probabilist might use probability to measure confirmation. A test of a philosophical confirmation theory is whether it elucidates or is even in sync with intuitive methodological principles about evidence or testing. Which, if either, fits with intuitions?

The most familiar interpretation is that H is confirmed by $\boldsymbol{x}$ if $\boldsymbol{x}$ gives a boost to the probability of H, *incremental* confirmation. The components of $C(H,\boldsymbol{x})$ are allowed to be any statements, and, in identifying C with Pr, no reference to a probability model is required. There is typically a background variable k, so that $\boldsymbol{x}$ confirms H relative to k: to the extent that $\Pr(H|\boldsymbol{x} \text{ and } k) > \Pr(H \text{ and } k)$. However, for readability, I will drop the explicit inclusion of k. More generally, if H entails $\boldsymbol{x}$, then assuming $\Pr(\boldsymbol{x}) \neq 1$ and $\Pr(H) \neq 0$, 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 $Pr(\boldsymbol{x}|H) > Pr(\boldsymbol{x})$.)

(1) *Incremental* (B-boost): H is confirmed by $\boldsymbol{x}$ iff $Pr(H|\boldsymbol{x}) > Pr(H)$, H is disconfirmed iff $Pr(H|\boldsymbol{x}) < Pr(H)$.

("iff" denotes if and only if.) Also plausible is an *absolute* interpretation:

(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: $Pr(H|\boldsymbol{x}) > 0.5$. From (1), $\boldsymbol{x}$ (the alarm) *dis*confirms the hypothesis H: the bag is clean, because its probability has gone down, however slightly. Yet from (2) $\boldsymbol{x}$ confirms H: bag is clean, because $Pr(H)$ is high to begin with.

There's a conflict. Thus, if (1) seems plausible, then probability, $Pr(H|\boldsymbol{x})$, isn't a satisfactory way to define confirmation. At the very least, we must distinguish between an incremental and an absolute measure of confirmation for H. No surprise there. From the start Carnap recognized that "the verb 'to confirm' is ambiguous"; Carnap and most others choose the "making firmer" or incremental connotation as better capturing what is meant than that of "making firm" (Carnap 1962, p. xviii). Incremental confirmation is generally used in current Bayesian epistemology. Confirmation is a B-boost.

The first point Popper's making in the epigraph is this: to identify confirmation and probability "C = Pr" leads to this type of conflict. His example is a single toss of a homogeneous die: The data $\boldsymbol{x}$: an even number occurs; hypothesis H: a 6 will occur. It's given that $Pr(H) = 1/6$, $Pr(\boldsymbol{x}) = 1/2$. The probability of H is increased by data $\boldsymbol{x}$, while $\sim H$ is undermined by $\boldsymbol{x}$ (its probability goes from 5/6 to 4/6). If we identify probability with degree of confirmation, $\boldsymbol{x}$ confirms H and disconfirms $\sim H$. However, $Pr(H|\boldsymbol{x}) < Pr(\sim H|\boldsymbol{x})$. So H is less well confirmed given $\boldsymbol{x}$ than is $\sim H$, in the sense of (2). Here's how Popper puts it, addressing Carnap: How can we say H is confirmed by $\boldsymbol{x}$, while $\sim H$ is not; but at the same time $\sim H$ is confirmed to a higher degree with $\boldsymbol{x}$ than is H? (Popper 1959, p. 390).[2]

[2] Let HJ be $(H \& J)$. To show: If there is a case where $\boldsymbol{x}$ confirms HJ more than $\boldsymbol{x}$ confirms J, then degree of probability cannot equal degree of confirmation.

(i) $C(HJ, \boldsymbol{x}) > C(J, \boldsymbol{x})$ is given.
(ii) $J = \sim HJ$ or HJ by logical equivalence.
(iii) $C(HJ, \boldsymbol{x}) > C(\sim HJ$ or $HJ, \boldsymbol{x})$ by substituting (ii) in line (i).

Since $\sim HJ$ and HJ are mutually exclusive, we have from the special addition rule for probability:

(iv) $Pr(HJ, \boldsymbol{x}) \leq Pr(\sim HJ$ or $HJ, \boldsymbol{x})$.

So if Pr = C, (iii) and (iv) yield a contradiction. (Adapting Popper 1959, p. 391)

Moreover, Popper continues, confirmation theorists don't use $\Pr(H|\boldsymbol{x})$ alone (as they would if $C = \Pr$), but myriad functions of probability to capture how much $\boldsymbol{x}$ has firmed up H. A number of measures offer themselves for the job. A simple B-boost would report the ratio R: $\Pr(H|\boldsymbol{x})/\Pr(H)$, which in Popper's example is 2. Or we can use the likelihood ratio of H compared to $\sim H$. Since I used LR in Excursion 1, where the two hypotheses are not exhaustive, let's write [LR] to denote

$$[\text{LR}]: \frac{\Pr(\boldsymbol{x}|H)}{\Pr(\boldsymbol{x}|\sim H)} = (1/0.4) = 2.5.$$

Many other ways of measuring the increase in confirmation that $\boldsymbol{x}$ affords H could do as well. (For some excellent lists see Popper 1959 and Fitelson 2002.)

What shall we say about the numbers like 2, 2.5? Do they mean the same thing in different contexts? Then there's the question of computing $\Pr(\boldsymbol{x}|\sim H)$, the *catchall factor*. It doesn't offer problems in this case because $\sim H$, the *catchall hypothesis*, is just an event statement. It's far more problematic once we move to genuine statistical hypotheses. Recall how Royall's Likelihoodist avoids the composite catchall factor by restricting his likelihood ratios to two simple statistical hypotheses.

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" (pp. 394–5). Ultimately, Popper denies that severity can be completely formalized by any C function. Is there nothing in between a pure formal-syntactical approach and leaving terms at a vague level? I say there is.

Consider for a moment philosopher Peter Achinstein – a Carnap student. Achinstein (2000, 2001) declared that scientists should not take seriously philosophical accounts of confirmation because they make it too easy to confirm. Furthermore, scientists look to empirical grounds for confirmation, whereas philosophical accounts give us formal (non-empirical) a priori measures. (I call it Achinstein's "Dean's problem" because he made the confession to a Dean asking about the relevance of philosophy – not usually the best way to keep funding for philosophy.) Achinstein rejects confirmation as increased firmness, denying it is either necessary or sufficient for evidence (rejects (1)).[3] He requires for H to be confirmed by $\boldsymbol{x}$ that the posterior of H given $\boldsymbol{x}$ be rather high, a version of (2): $\Pr(H|\boldsymbol{x}) \gg \Pr(\sim H|\boldsymbol{x})$, but that's not all. He requires that, before we apply

[3] Why is a B-boost not necessary for Achinstein? Suppose you know x: the newspaper says Harry won, and it's never wrong. Then a radio, also assumed 100% reliable, announces y: Harry won. Statement y, Achinstein thinks, should still count as evidence for H: he won. I agree.

confirmation measures, the components have an appropriate explanatory relationship to each other. Yet this requires an adequate way to make explanatory inferences before getting started. It's not clear how the formalism helped. He still considers himself a Bayesian epistemologist – a term that has replaced confirmation theorist – but the probabilistic representation threatens to be mostly a kind of bookkeeping for inferential work done in some other way.

Achinstein is right to object that (1) incremental confirmation makes it too easy to have evidence. After all, *J*: Mike drowns in the Pacific Ocean, entails $\boldsymbol{x}$: there is a Pacific Ocean; yet $\boldsymbol{x}$ does not seem to be evidence for *J*. Still the generally favored position is to view confirmation as (1) a B-boost.

Exhibit (iv): Paradox of Irrelevant Conjunctions. Consider a famous argument due to Glymour (1980). If we allow that $\boldsymbol{x}$ confirms *H* so long as $\Pr(H|\boldsymbol{x}) > \Pr(H)$, it seems everything confirms everything, so long as one thing is confirmed!

The first piece of the argument is the problem of irrelevant conjunctions – also called the "tacking paradox." If $\boldsymbol{x}$ confirms *H*, then $\boldsymbol{x}$ also confirms $(H \,\&\, J)$, even if hypothesis *J* is just "tacked on" to *H*. As with most of these chestnuts, there is a long history (e.g., Earman 1992, Rosenkrantz 1977) but I consider a leading contemporary representative, Branden Fitelson. Fitelson (2002) and Hawthorne and Fitelson (2004) define the statement "*J* is an *irrelevant conjunct* to *H*, with respect to evidence $\boldsymbol{x}$" as meaning $\Pr(\boldsymbol{x}|J) = \Pr(\boldsymbol{x}|J \,\&\, H)$. For instance, $\boldsymbol{x}$ might be radioastronomic data in support of

> *H*: the General Theory of Relativity (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) gives $\Pr(\boldsymbol{x}|H \,\&\, J)/\Pr(\boldsymbol{x}) > 1$.

> Therefore $\boldsymbol{x}$ Bayesian-confirms $(H \,\&\, J)$.[4]

[4] To expand the reasoning, first observe that $\Pr(H|\boldsymbol{x})/\Pr(H) = \Pr(\boldsymbol{x}|H)/\Pr(\boldsymbol{x})$ and $\Pr(H \,\&\, J|\boldsymbol{x})/\Pr(H \,\&\, J) = \Pr(\boldsymbol{x}|H \,\&\, J)/\Pr(\boldsymbol{x})$, both by Bayes' Theorem. So, when $\Pr(H|\boldsymbol{x})/\Pr(H) > 1$, we also have $\Pr(\boldsymbol{x}|H)/\Pr(\boldsymbol{x}) > 1$. This, together with $\Pr(\boldsymbol{x}|H \,\&\, J) = \Pr(\boldsymbol{x}|H)$ (given), yields $\Pr(\boldsymbol{x}|H \,\&\, J)/\Pr(\boldsymbol{x}) > 1$. Thus, we also have $\Pr(H \,\&\, J|\boldsymbol{x})/\Pr(H \,\&\, J) > 1$.

However, it is also plausible to hold what philosophers call the "special consequence" condition: If $\boldsymbol{x}$ confirms a claim W, and W entails J, then $\boldsymbol{x}$ confirms J. In particular:

(B) If $\boldsymbol{x}$ confirms $(H \,\&\, J)$, then $\boldsymbol{x}$ confirms J.

(B) gives the second piece of the argument. From (A) and (B) we have, if $\boldsymbol{x}$ confirms H, then $\boldsymbol{x}$ confirms J for any irrelevant J consistent with H (neither H nor J have probabilities 0 or 1).

It follows that if $\boldsymbol{x}$ confirms any H, then $\boldsymbol{x}$ confirms any J.

This absurd result, however, assumed (B) (special consequence) and most Bayesian epistemologists reject it. This is the gist of Fitelson's solution to tacking, updated in Hawthorne and Fitelson (2004). It is granted that $\boldsymbol{x}$ confirms the conjunction $(H \,\&\, J)$, while denying $\boldsymbol{x}$ confirms the irrelevant conjunct J. Aren't they uncomfortable with (A), allowing $(H \,\&\, J)$ to be confirmed by $\boldsymbol{x}$?

I'm inclined to agree with Glymour that we are not too happy with an account of evidence that tells us deflection of light data confirms the conjunction of the GTR deflection and the radioactivity of the Fukushima water is within acceptable levels, while assuring us that $\boldsymbol{x}$ does not confirm the conjunct, that the Fukushima water has acceptable levels of radiation (1980, p. 31). Moreover, suppose we measure the confirmation boost by

$$\text{R: } \Pr(H|\boldsymbol{x})/\Pr(\boldsymbol{x}).$$

Then, Fitelson points out, the conjunction $(H \,\&\, J)$ is just as well confirmed by $\boldsymbol{x}$ as is H!

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, instead of R, the likelihood ratio [LR] of H against $\sim H$:

$$\text{[LR]: } \Pr(\boldsymbol{x}|H)/\Pr(\boldsymbol{x}|\sim H).^{5}$$

This avoids the counterintuitive result, or so it is claimed. (Note: $\Pr(H|\boldsymbol{x}) > \Pr(H)$ iff $\Pr(\boldsymbol{x}|H) > \Pr(\boldsymbol{x})$, but measuring the boost by R differs from measuring it with [LR].)

[5] Recall that Royall restricts the likelihood ratio to non-composite hypotheses, whereas here $\sim H$ is the Bayesian catchall.

What Does the Severity Account Say?

Our account of inference disembarked way back at (1): that $\boldsymbol{x}$ confirms H so long as $\Pr(H|\boldsymbol{x}) > \Pr(H)$. That is, we reject probabilistic affirming the consequent. In the simplest case, H entails $\boldsymbol{x}$, and $\boldsymbol{x}$ 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 $\boldsymbol{x}$. It's the same reason we reject the Law of Likelihood (LL). Unless stringent probing has occurred, finding an H that fits $\boldsymbol{x}$ is not difficult to achieve even if H is false. H hasn't passed severely. Now severely passing is obviously stronger than merely finding some evidence for H, and 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 or rule, and just handing me the H's, $\boldsymbol{x}$'s, and relative probabilities will not suffice. The process of tacking, at least one form, is this – once you have an incrementally confirmed H with data $\boldsymbol{x}$, tack on any consistent J and announce "$\boldsymbol{x}$ confirms (H & J)." Let's allow that (H & J) fits or accords with $\boldsymbol{x}$ (since GTR entails or renders probable the deflection data $\boldsymbol{x}$). However, the very claim: "(H & J) is confirmed by $\boldsymbol{x}$" 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. B-boosters might reply, "We're admitting J is irrelevant and gets no confirmation," but our testing intuitions tell us then it's crazy to regard (H & J) as confirmed. They will point out other examples where this doesn't seem crazy. But what matters is that it's being permitted in general.

We should *punish* a claim to have evidence for H with a tacked-on J, when nothing has been done to refute J. Imagine the chaos. Are we to allow positive trial data on diabetes patients given drug D to confirm the claim that D improves survival of diabetes patients *and* Roche's artificial knee is effective, when there's only evidence for one? If the confirmation theorist simply stipulates that (1) defines confirmation, then it's within your rights to deny it captures ordinary notions of evidence. On the other hand, if you do accept (1), then why are you bothered at all by tacking? Many are not.

Patrick Maher (2004) argues that if B-boosting is confirmation, then there is nothing counterintuitive about data confirming irrelevant conjunctions; Fitelson should not even be conceding "he bites the bullet." It makes sense that (H & J) increases the probability assignment to $\boldsymbol{x}$ just as much as does H, for J the irrelevant conjunct. The supposition that this is problematic and that therefore one must move away from R: $\Pr(\boldsymbol{x}|H)/\Pr(\boldsymbol{x})$ sits uneasily with the fact

that R > 1 is just what confirmation boost means. Rather than "solve" the problem by saying we can measure boost so that (*H* & *J*) gets less confirmation than *H*, using [LR], why not see it as what's meant by an irrelevant conjunct *J: J* doesn't improve the ability to predict ***x***. Other philosophers working in this arena, Crupi and Tentori (2010), notice that [LR] is not without problems. In particular, if ***x*** *dis*confirms hypothesis *Q*, then (*Q* & *J*) isn't as badly disconfirmed as *Q* is, for irrelevant conjunct *J*. Just as (*H* & *J*) gets less of a B-boost than does *H*, (*Q* & *J*) gets less disconfirmation in the case where ***x*** disconfirms *J*. This too makes sense on the [LR] measure, though I will spare the details. Their intuitions are that this is worse than the irrelevant conjunction case, and is not solved by the use of [LR]. Interesting new measures are offered. Again, this seems to our tester to reflect the tension between Bayes boosts and good tests.

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, substituting *H* for *h*; ***x*** for *e*)

Sincerity! Popper never held that severe tests turned on a psychological notion, but he was at a loss to formalize severity. A fuller passage from Popper (1959) is worth reading if you get a chance.[6] All the measures of confirmation, be it R or LR, or one of the others, count merely as "fit" or "accordance" measures to Popper and to the severe tester. They may each be relevant for different problems – that there are different dimensions for fit is to be expected. These measures do not capture what's needed to determine if much (or anything) has been done to find *H* is flawed. What we need to add are the associated error probabilities. Error probabilities do not enter into these standard confirmation theories – which isn't to say they couldn't. If R is used and observed to be *r*, we want to compute Pr(R > *r*; ~(*H* & *J*)). Here, the probability of getting R > 1 is maximal (since (*H* & *J*) entails ***x***), even if ~(*H* & *J*) is true. So ***x*** is "bad evidence,

[6] "I must insist that *C*(*h*, *e*) can be interpreted as degree of corroboration only if *e* is *a report on the severest tests we have been able to design*. It is this point that marks the difference between the attitude of the inductivist, or verificationist, and my own attitude. The inductivist or verificationist wants *affirmation* for his hypothesis. He hopes to make it *firmer* by his evidence *e* and he looks out for '*firmness*' – for '*confirmation.*' . . . Yet if *e* is *not* a report about the results of our sincere attempts to overthrow *h*, then we shall simply deceive ourselves if we think we can interpret *C*(*h,e*) as degree of corroboration, or anything like it." (Popper 1959, p. 418).

no test" (BENT) for the conjunction.[7] It's not a psychological "sincerity" being captured; nor is it purely context free. Popper couldn't capture it as he never made the error probability turn.

Time prevents us from entering multiple other rooms displaying paradoxes of confirmation theory, where we'd meet up with such wonderful zombies as the white shoe confirming all ravens are black, and the "grue" paradox, which my editor banished from my 1996 book. (See Skyrms 1986.) Enough tears have been shed. Yet they shouldn't be dismissed too readily; they very often contain a puzzle of deep relevance for statistical practice. There are two reasons the tacking paradox above is of relevance to us. The first concerns a problem that arises for both Popperians and Bayesians. There is a large-scale theory T that predicts $\boldsymbol{x}$, 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. This may not be evident until long after the experiment. We don't want to say there is evidence for a large-scale theory such as GTR just because one part was well tested. On the other hand, it may well be that all relativistic theories with certain properties have passed severely.

Second, the question of whether to measure support with a Bayes boost or with posterior probability arises in Bayesian statistical inference as well. 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 rightly 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. I think this is a mistake. That doesn't preclude a general account for statistical inference; it just won't be purely formal.

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. The problem to which Glymour is alluding is this: you can start with the principle you want your confirmation logic to reflect, and then *reconstruct* it using probability. The task, for the formal epistemologist, becomes the problem of assigning priors and likelihoods that mesh with the principle you want to defend. Here's an example. Some think

[7] The real problem is that $\Pr(x; H \,\&\, J) = \Pr(x; H \,\&\, \sim J)$.

that GTR got more confirmation than a rival theory (e.g., Brans-Dicke theory) because the latter is made to fit the data thanks to adjustable parameters (Jefferys and Berger 1992). Others think the fact it had adjustable parameters does not alter the confirmation (Earman 1992). They too can reconstruct the episode so that Brans-Dicke pays no penalty. The historical episode can be "rationally reconstructed" to accord with either philosophical standpoint.

Although the problem of statistical inference is only a small part of what today goes under the umbrella of formal epistemology, progress in the statistics wars would advance more surely if philosophers regularly adopted the language of statistics. Not only would we be better at the job of clarifying the conceptual discomforts among practitioners of statistics and modeling, some of the classic problems of confirmation could be scotched using the language of random variables and their distributions.[8] Philosophy of statistics had long been ahead of its time, in the sense of involving genuinely interdisciplinary work with statisticians, scientists, and philosophers of science. We need to return to that. There are many exceptions, of course; yet to try to list them would surely make me guilty of leaving several out.

[8] For a discussion and justification of the use of "random variables," see Mayo (1996).