Mementos from Excursion 2 Tour II: Falsification, Pseudoscience, Induction 2.3-2.7 (first installment, Nov. 17, 2018)[1]

Sketch of Tour: Tour II visits Popper, falsification, corroboration, Duhem's problem (what to blame in the case of anomalies) and the demarcation of science and pseudoscience (2.3). While Popper comes up short on each, the reader is led to improve on Popper's notions (live exhibit (v)). Central ingredients for our journey are put in place via souvenirs: a framework of models and problems, and a post-Popperian language to speak about inductive inference. Defining a severe test, for Popperians, is linked to when data supply novel evidence for a hypothesis: family feuds about defining novelty are discussed (2.4). We move into Fisherian significance tests and the crucial requirements he set (often overlooked): isolated significant results are poor evidence of a genuine effect, and statistical significance doesn't warrant substantive, e.g., causal inference (2.5). Applying our new demarcation criterion to a plausible effect (males are more likely than females to feel threatened by their partner's success), we argue that a real revolution in psychology will need to be more revolutionary than at present. Whole inquiries might have to be falsified, their measurement schemes questioned (2.6). The Tour's pieces are synthesized in (2.7), where a guest lecturer explains how to solve the problem of induction now, having redefined induction as severe testing.

Mementos from 2.3

There are four key, interrelated themes from Popper:

(1) Science and Pseudoscience. For a theory to be scientific it must be testable and falsifiable.
(2) Conjecture and Refutation. We learn not by enumerative induction but by trial and error: conjecture and refutation.
(3) Observations Are Not Given. If they are at the "foundation," it is only because there are apt methods for testing their validity. We dub claims observable *because* or to the extent that they are open to stringent checks.
(4) Corroboration Not Confirmation, Severity Not Probabilism. Rejecting probabilism, Popper denies scientists are interested in highly probable hypotheses (in any sense). They seek bold, informative, interesting conjectures and ingenious and severe attempts to refute them.

These themes are in the spirit of the error statistician. Considerable spade-work is required to see *what to keep and what to revise*, so bring along your archeological shovels.

The severe tester revises Popper's Demarcation of Science (Live Exhibit (vi)):
What he should be asking is not whether a theory is unscientific, but *When is an inquiry into a theory, or an appraisal of claim H, unscientific*? We want to distinguish meritorious modes of inquiry from those that are BENT. If the test methods enable ad hoc maneuvering, sneaky face-saving devices, then the inquiry – the handling and use of data – is unscientific. Despite being logically falsifiable, theories can be rendered immune from falsification by means of questionable methods for their testing.

Greater Content, Greater Severity. The severe tester accepts Popper's central intuition in (4): if we wanted highly probable claims, scientists would stick to low-level observables and not seek generalizations, much less theories with high explanatory content.A highly explanatory, high-content theory, with interconnected tentacles, has a higher probability of having flaws discerned than low-content theories that do not rule out as much. Thus, when the bolder, higher content, theory stands up to testing, it may earn higher overall severity than the one with measly content.

It is the fuller, unifying, theory developed in the course of solving interconnected problems that enables severe tests.

Methodological Probability. *P*robability in learning attaches to a method of conjecture and refutation, that is to testing: it is *methodological probability*. An error probability is a special case of a methodological probability. We want methods with a high probability of teaching us (and machines) how to distinguish approximately correct and incorrect interpretations of data. That a theory is plausible is of little interest, in and of itself; what matters is that it is *im*plausible for it to have passed these tests were it false or incapable of adequately solving its set of problems.

Methodological falsification: We appeal to methodological rules for when to regard a claim as falsified.

- Inductive-statistical falsification proceeds by methods that allow ~H to be inferred with severity. A first step is often to infer an anomaly is real, by falsifying a "due to chance" hypothesis.
- Going further, we may corroborate (i.e., infer with severity) effects that count as falsifying hypotheses. A *falsifying hypothesis* is a hypothesis inferred in order to falsify some other claim. Example: the pathological proteins (prions) in mad cow disease infect without nucleic acid. This falsifies: all infectious agents involve nucleic acid.

Despite giving lip service to testing and falsification, many popular accounts of statistical inference do not embody falsification – even of a statistical sort.

However, the falsifying hypotheses that are integral for Popper also necessitate an evidence-transcending (inductive) statistical inference.

The Popperian (Methodological) Falsificationist Is an Error Statistician
When is a statistical hypothesis to count as falsified? Although extremely rare events may occur, Popper notes:

> such occurrences would not be physical effects, because, on account of their immense improbability, *they are not reproducible at will* ... If, however, we find *reproducible deviations from a macro effect .. . deduced from a probability estimate ... then we must assume that the probability estimate is falsified. (Popper 1959, p. 203)*

In the same vein, we heard Fisher deny that an "isolated record" of statistically significant results suffices to warrant a reproducible or genuine effect (Fisher 1935a, p. 14).

In a sense, the severe tester 'breaks' from Popper by solving his key problem: Popper's account rests on severe tests, tests that would probably falsify claims if false, but he cannot warrant saying a method is probative or severe, because that would mean it was reliable, which makes Popperians squeamish. It would appear to concede to his critics that Popper has a "whiff of induction" after all. But it's not inductive enumeration. Error statistical methods (whether from statistics or informal) can supply the severe tests Popper sought.

A scientific inquiry (a procedure for finding something out) for a severe tester:

- blocks inferences that fail the minimal requirement for severity:

- *must be able to embark on a reliable probe to pinpoint blame for anomalies (and use the results to replace falsified claims and build a repertoire of errors).*

The parenthetical remark isn't absolutely required, but is a feature that greatly strengthens scientific credentials.
The reliability requirement is: infer claims just to the extent that they pass severe tests. There's no sharp line for demarcation, but when these requirements are absent, an inquiry veers into the realm of questionable science or pseudoscience.

2.4 Novelty and Severity

There is a tension between the drive for a logic of confirmation and our strictures against practices that lead to poor tests and *ad hoc* hypotheses.

Adhering to the former downplays or blocks the ability to capture the latter, which demands we go beyond the data and hypotheses; need to know something about the *history* of the hypothesis: Was the hypothesis developed as a result of deliberate and *ad hoc* attempts to spare one's theory from refutation?

When holders of the Likelihood Principle (LP) wonder why data can't speak for themselves, they're echoing the logical empiricist (1.4)"

> According to modern logical empiricist orthodoxy, in deciding whether hypothesis *h* is confirmed by evidence *e*, …we must consider only the statements *h* and *e*, and the logical relations between them. It is quite irrelevant whether *e* was known first and *h* proposed to explain it, or whether *e* resulted from testing predictions drawn from *h*. (Musgrave 1974, p. 2)

Logics of confirmation ran into problems because they insisted on purely formal or syntactical criteria of confirmation that, like deductive logic, "should contain no reference to the specific subject matter" (Hempel, 1945, p. 9) in question. The Popper-Lakatos school attempts to avoid these shortcomings by means of novelty requirements:

> *Novelty Requirement*: for data to warrant a hypothesis *H* requires not just that
> (i) *H* agree with the data, but also (ii) the data should be novel or surprising or the like.

Types of novelty: There's (1) *temporal novelty*–the data were not already available before the hypothesis was erected (Popper, early); (2) *theoretical novelty*–the data were not already predicted by an existing hypothesis (Popper, Lakatos), and (3) *use-novelty*–the data were not used to construct or select the hypothesis.

Severe Testers: What matters is not novelty, in any of the senses, but severity in the error statistical sense. Even where our intuition is to prohibit use-novelty violations, the requirement is murky. We should instead consider specific ways that severity can be violated.

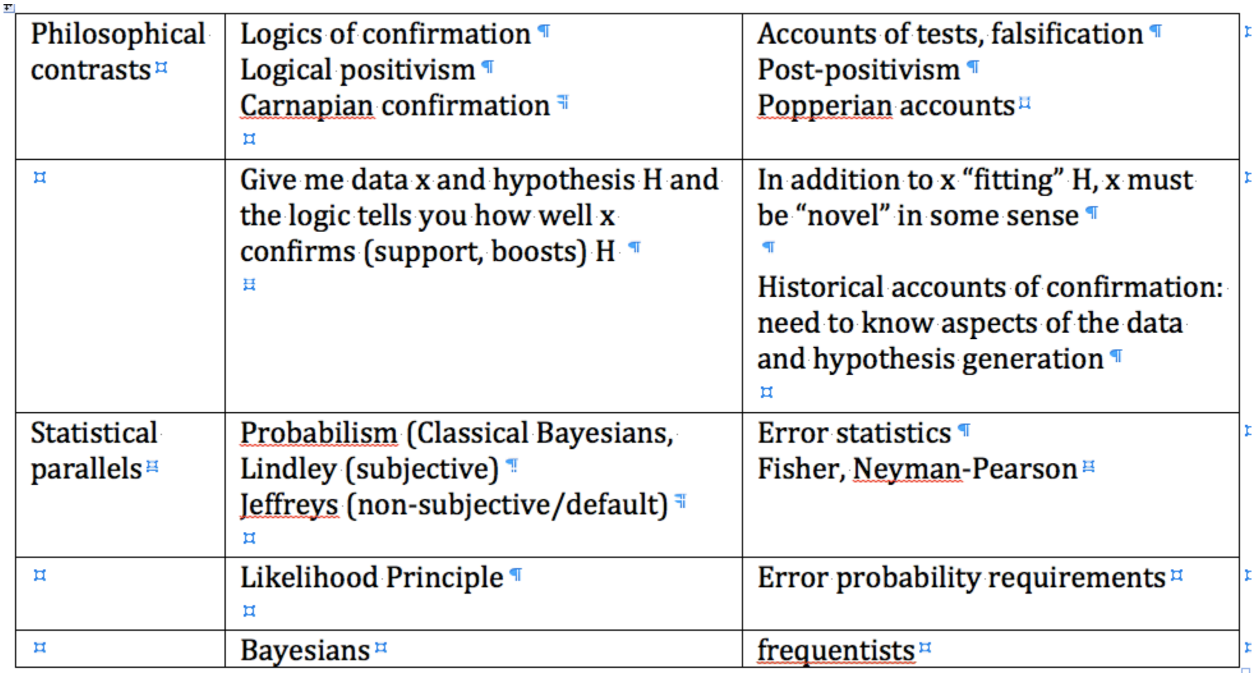
> *Biasing selection effects*: when data or hypotheses are selected or generated (or a test criterion is specified), in such a way that the minimal severity requirement is violated, seriously altered, or incapable of being assessed.

Cherry picking, fishing, hunting, significance seeking, searching for the pony, trying and trying again, data dredging, monster-barring, look elsewhere effect, P-hacking, multiple testing.

Putting severity in the form of the Popper-Lakatos school:

> *Severity Requirement:* for data to warrant a hypothesis *H* requires not just that
> (S-1) *H* agree with the data (*H* passes the test), but also (S-2) with high probability, *H* would not have passed the test so well, were *H* false.

This describes corroborating a claim, it's "strong" severity. Weak severity denies *H* is warranted if the test method would probably have passed *H* even if false.



Philosophical contrasts	Logics of confirmation Logical positivism Carnapian confirmation	Accounts of tests, falsification Post-positivism Popperian accounts
	Give me data x and hypothesis H and the logic tells you how well x confirms (support, boosts) H	In addition to x "fitting" H, x must be "novel" in some sense Historical accounts of confirmation: need to know aspects of the data and hypothesis generation
Statistical parallels	Probabilism (Classical Bayesians, Lindley (subjective) Jeffreys (non-subjective/default)	Error statistics Fisher, Neyman-Pearson
	Likelihood Principle	Error probability requirements
	Bayesians	frequentists

2.5 Fallacies of Rejection and an Animal Called NHST

fallacies of rejection.

1. The reported (nominal) statistical significance result is *spurious* (it's not even an actual P-value). This can happen in two ways: biasing selection effects, or violated assumptions of the model.
2. The reported statistically significant result is genuine, but it's an isolated effect not yet indicative of a genuine experimental phenomenon. (Isolated low P-value ≠> *H*: statistical effect)
3. There's evidence of a genuine statistical phenomenon but either (i) the magnitude of the effect is less than purported, call this a *magnitude error*, or (ii) the substantive interpretation is unwarranted. (*H* ≠> *H**)

- An *audit* of a P-value: a check of any of these concerns, generally in order, depending on the inference
- Until audits are passed, the relevant statistical inference is to be reported as "unaudited".
- Until #2 is ruled out, it's a mere "indication", perhaps, in some settings, grounds to get more data.

Criticisms of significance tests are based on an animal that goes by the acronym NHST (null hypothesis significance testing).

> *If NHST permits going from a single small P-value to a genuine effect, it is illicit; and if it permits going directly to a substantive research claim it is doubly illicit!*

- We can add: if it permits biasing selection effects it's triply guilty.
- Drop the term NHST; statistical tests will do.

2.6 The Reproducibility Revolution (Crisis) in Psychology

The "replication revolution in psychology" won't be nearly revolutionary enough until they subject to testing the methods and measurements intended to link statistics with what they really want to know. A hypothesis to be considered must always be: *the results point to the inability of the study to severely probe the phenomenon of interest*. The goal would be to build up a body of knowledge on closing existing loopholes when conducting a type of inquiry. *The scientific status of an inquiry is questionable* if it cannot or will not distinguish the correctness of inferences from problems stemming from a poorly run study. What must be subjected to grave risk are assumptions that the experiment was well run.

2.7 How to Solve the Problem of Induction Now

Viewing inductive inference as severe testing, the problem of induction is transformed into the problem of showing the existence of severe tests and methods for identifying insevere ones. The trick isn't to have a formal, context free method–as with the traditional problem of induction; the trick is to have methods that alert us when an application is shaky.

What enables induction (as severe testing) to work: Informal, Quasi-formal, and formal: assorted strategies for amplifying and learning from types of errors and mistakes.

What Warrants Inferring a Hypothesis that Passes Severe Tests?

Even with a strong argument from coincidence akin to my weight gain showing up on myriad calibrated scales, there is no logical inconsistency with invoking a hypothesis from *conspiracy*: all these instruments conspire to produce results as if *H* were true but in fact *H* is false. The ultra-skeptic may invent a *rigged* hypothesis R:

R: Something else other than *H* actually explains the data

without actually saying what this something else is. If someone is bound to discount a strong argument for *H* by rigging, then she will be adopting a highly unreliable method. Even with claims that are true, or where problems are solved correctly, she would have no chance of finding this out. I began with the stipulation that we wish to learn. Inquiry that blocks learning is pathological. *This leads severe testers to go beyond weak, to strong, severity.*

[1] I’m sure to revise these over the course of the winter semester, so please check back. Please note corrections on my blog.