Tour II Falsification, Pseudoscience, Induction

We'll move from the philosophical ground floor to connecting themes from other levels, from Popperian falsification to significance tests, and from Popper's demarcation to current-day problems of pseudoscience and irreplication. An excerpt from our Museum Guide gives a broad-brush sketch of the first few sections of Tour II:

> Karl Popper had a brilliant way to "solve" the problem of induction: Hume was right that enumerative induction is unjustified, but science is a matter of deductive falsification. Science was to be demarcated from pseudoscience according to whether its theories were testable and falsifiable. A hypothesis is deemed severely tested if it survives a stringent attempt to falsify it. Popper's critics denied he could sustain this and still be a deductivist . . .
>
> Popperian falsification is often seen as akin to Fisher's view that "every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (1935a, p. 16). Though scientists often appeal to Popper, some critics of significance tests argue that they are used in decidedly non-Popperian ways. Tour II explores this controversy.

While Popper didn't make good on his most winning slogans, he gives us many seminal launching-off points for improved accounts of falsification, corroboration, science versus pseudoscience, and the role of novel evidence and predesignation. These will let you revisit some thorny issues in today's statistical crisis in science.

2.3 Popper, Severity, and Methodological Probability

Here's Popper's summary (drawing from Popper, *Conjectures and Refutations*, 1962, p. 53):

- [Enumerative] induction . . . is a myth. It is neither a psychological fact . . . nor one of scientific procedure.
- The actual procedure of science is to operate with conjectures. . .
- Repeated observation and experiments function in science as tests of our conjectures or hypotheses, i.e., as attempted refutations.

- [It is wrongly believed that using the inductive method can] *serve as a criterion of demarcation between science and pseudoscience.* . . . None of this is altered in the least if we say that induction makes theories only probable.

There are four key, interrelated themes:

(1) Science and Pseudoscience. Redefining scientific method gave Popper a new basis for demarcating genuine science from questionable science or pseudoscience. Flexible theories that are easy to confirm – theories of Marx, Freud, and Adler were his exemplars – where you open your eyes and find confirmations everywhere, are low on the scientific totem pole (ibid., p. 35). For a theory to be scientific it must be testable and falsifiable.

(2) Conjecture and Refutation. The problem of induction is a problem only if it depends on an unjustifiable procedure such as enumerative induction. Popper shocked everyone by denying scientists were in the habit of inductively enumerating. It doesn't even hold up on logical grounds. To talk of "another instance of an A that is a B" assumes a conceptual classification scheme. How else do we recognize it as another item under the umbrellas A and B? (ibid., p. 44). You can't just observe, you need an interest, a point of view, a problem.

The actual procedure for learning in science is to operate with conjectures in which we then try to find weak spots and flaws. Deductive logic is needed to draw out the remote logical consequences that we actually have a shot at testing (ibid., p. 51). From the scientist down to the amoeba, says Popper, we learn by trial and error: conjecture and refutation (ibid., p. 52). The crucial difference is the extent to which we constructively learn how to reorient ourselves after clashes.

> Without waiting, passively, for repetitions to impress or impose regularities upon us, we actively try to impose regularities upon the world. . . These may have to be discarded later, should observation show that they are wrong. (ibid., p. 46)

(3) Observations Are Not Given. Popper rejected the time-honored empiricist assumption that observations are known relatively unproblematically. 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. (Popper: "anyone who has learned the relevant technique can test it" (1959, p. 99).) Accounts of hypothesis appraisal that start with "evidence $\boldsymbol{x}$," as in confirmation logics, vastly oversimplify the role of data in learning.

(4) Corroboration Not Confirmation, Severity Not Probabilism. Last, there is his radical view on the role of probability in scientific inference. Rejecting probabilism, Popper not only rejects Carnap-style logics of confirmation, he 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. If one uses a logical notion of probability, as philosophers (including Popper) did at the time, the high content theories are highly improbable; in fact, Popper said universal theories have 0 probability. (Popper also talked of statistical probabilities as propensities.)

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

Demarcation and Investigating Bad Science

There is a reason that statisticians and scientists often refer back to Popper; his basic philosophy – at least his most winning slogans – are in sync with ordinary intuitions about good scientific practice. Even people divorced from Popper's full philosophy wind up going back to him when they need to demarcate science from pseudoscience. Popper's right that if using enumerative induction makes you scientific then anyone from an astrologer to one who blithely moves from observed associations to full blown theories is scientific. Yet the criterion of testability and falsifiability – as typically understood – is nearly as bad. It is both too strong and too weak. Any crazy theory found false would be scientific, and our most impressive theories aren't deductively falsifiable. Larry Laudan's famous (1983) "The Demise of the Demarcation Problem" declared the problem taboo. This is a highly unsatisfactory situation for philosophers of science. Now Laudan and I generally see eye to eye, perhaps our disagreement here is just semantics. I share his view that what really matters is determining if a hypothesis is warranted or not, rather than whether the theory is "scientific," but surely Popper didn't mean logical falsifiability sufficed. Popper is clear that many unscientific theories (e.g., Marxism, astrology) are falsifiable. It's clinging to falsified theories that leads to unscientific practices. (Note: The use of a strictly falsified theory for prediction, or because nothing better is available, isn't unscientific.) I say that, with a bit of fine-tuning, we can retain the essence of Popper to capture what makes an inquiry, if not an entire domain, scientific.

Following Laudan, philosophers tend to shy away from saying anything general about science versus pseudoscience – the predominant view is that there is no such thing. Some say that there's at most a kind of "family

resemblance" amongst domains people tend to consider scientific (Dupré 1993, Pigliucci 2010, 2013). One gets the impression that the demarcation task is being left to committees investigating allegations of poor science or fraud. They are forced to articulate what to count as fraud, as bad statistics, or as mere questionable research practices (QRPs). People's careers depend on their ruling: they have "skin in the game," as Nassim Nicholas Taleb might say (2018). The best one I know – the committee investigating fraudster Diederik Stapel – advises making philosophy of science a requirement for researchers (Levelt Committee, Noort Committee, and Drenth Committee 2012). So let's not tell them philosophers haven given up on it.

Diederik Stapel. A prominent social psychologist "was found to have committed a serious infringement of scientific integrity by using fictitious data in his publications" (Levelt Committee 2012, p. 7). He was required to retract 58 papers, relinquish his university degree and much else. The authors of the report describe walking into a culture of confirmation and verification bias. They could scarcely believe their ears when people they interviewed "defended the serious and less serious violations of proper scientific method with the words: that is what I have learned in practice; everyone in my research environment does the same, and so does everyone we talk to at international conferences" (ibid., p. 48). Free of the qualms that give philosophers of science cold feet, they advance some obvious yet crucially important rules with Popperian echoes:

> One of the most fundamental rules of scientific research is that an investigation must be designed in such a way that facts that might refute the research hypotheses are given at least an equal chance of emerging as do facts that confirm the research hypotheses. Violations of this fundamental rule, such as continuing to repeat an experiment until it works as desired, or excluding unwelcome experimental subjects or results, inevitably tend to confirm the researcher's research hypotheses, and essentially render the hypotheses immune to the facts. (ibid., p. 48)

Exactly! This is our minimal requirement for evidence: If it's so easy to find agreement with a pet theory or claim, such agreement is bad evidence, no test, BENT. To scrutinize the scientific credentials of an inquiry is to determine if there was a serious attempt to detect and report errors and biasing selection effects. We'll meet Stapel again when we reach the temporary installation on the upper level: The Replication Revolution in Psychology.

The issue of demarcation (point (1)) is closely related to Popper's conjecture and refutation (point (2)). While he regards a degree of dogmatism to be necessary before giving theories up too readily, the trial and error methodology "gives us a chance to survive the elimination of an inadequate hypothesis –

when a more dogmatic attitude would eliminate it by eliminating us" (Popper 1962, p. 52). Despite giving lip service to testing and falsification, many popular accounts of statistical inference do not embody falsification – even of a statistical sort.

Nearly everyone, however, now accepts point (3), that observations are not just "given" – knocking out a crucial pillar on which naïve empiricism stood. To the question: What came first, hypothesis or observation? Popper answers, another hypothesis, only lower down or more local. Do we get an infinite regress? No, because we may go back to increasingly primitive theories and even, Popper thinks, to an inborn propensity to search for and find regularities (ibid., p. 47). I've read about studies appearing to show that babies are aware of what is statistically unusual. In one, babies were shown a box with a large majority of red versus white balls (Xu and Garcia 2008, Gopnik 2009). When a succession of white balls are drawn, one after another, with the contents of the box covered with a screen, the babies looked longer than when the more common red balls were drawn. I don't find this far-fetched. Anyone familiar with preschool computer games knows how far toddlers can get in solving problems without a single word, just by trial and error.

Greater Content, Greater Severity. The position people are most likely to take a pass on is (4), his view of the role of probability. Yet Popper's central intuition is correct: if we wanted highly probable claims, scientists would stick to low-level observables and not seek generalizations, much less theories with high explanatory content. In this day of fascination with Big Data's ability to predict what book I'll buy next, a healthy Popperian reminder is due: humans also want to understand and to explain. We want bold "improbable" theories. I'm a little puzzled when I hear leading machine learners praise Popper, a realist, while proclaiming themselves fervid instrumentalists. That is, they hold the view that theories, rather than aiming at truth, are just instruments for organizing and predicting observable facts. It follows from the success of machine learning, Vladimir Cherkassky avers, that "realism is not possible." This is very quick philosophy! ". . . [I]n machine learning we are given a set of [random] data samples, and the goal is to select the best model (function, hypothesis) from a given set of possible models" (Cherkassky 2012). Fine, but is the background knowledge required for this setup itself reducible to a prediction–classification problem? I say no, as would Popper. Even if Cherkassky's problem is relatively theory free, it wouldn't follow this is true for all of science. Some of the most impressive "deep learning" results in AI have been criticized for lacking the ability to generalize beyond observed "training" samples, or to solve open-ended problems (Gary Marcus 2018).

A valuable idea to take from Popper is that probability 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, even leaving murky cases in the middle, and how to advance knowledge of detectable, while strictly unobservable, effects.

The choices for probability that we are commonly offered are stark: "in here" (beliefs ascertained by introspection) or "out there" (frequencies in long runs, or chance mechanisms). This is the "epistemology" versus "variability" shoehorn we reject (Souvenir D). To qualify the method by which H was tested, frequentist performance is necessary, but it's not sufficient. The assessment must be relevant to ensuring that claims have been put to severe tests. You can talk of a test having a type of *propensity* or capability to have discerned flaws, as Popper did at times. 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. 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. It is the fuller, unifying, theory developed in the course of solving interconnected problems that enables severe tests.

Methodological probability is not to quantify my beliefs, but neither is it about a world I came across without considerable effort to beat nature into line. Let alone is it about a world-in-itself which, by definition, can't be accessed by us. Deliberate effort and ingenuity are what allow me to ensure I shall come up against a brick wall, and be forced to reorient myself, at least with reasonable probability, when I test a flawed conjecture. The capabilities of my tools to uncover mistaken claims (its error probabilities) are real properties of the tools. Still, they are *my* tools, specially and imaginatively designed. If people say they've made so many judgment calls in building the inferential apparatus that what's learned cannot be objective, I suggest they go back and work some more at their experimental design, or develop better theories.

Falsification Is Rarely Deductive. It is rare for any interesting scientific hypotheses to be logically falsifiable. This might seem surprising given all the applause heaped on falsifiability. For a scientific hypothesis H to be deductively falsified, it would be required that some observable result taken together with H yields a logical contradiction (A & ~A). But the only theories that

deductively prohibit observations are of the sort one mainly finds in philosophy books: All swans are white is falsified by a single non-white swan. There are some statistical claims and contexts, I will argue, where it's possible to achieve or come close to deductive falsification: claims such as, these data are independent and identically distributed (IID). Going beyond a mere denial to replacing them requires more work.

However, interesting claims about mechanisms and causal generalizations require numerous assumptions (substantive and statistical) and are rarely open to deductive falsification. How then can good science be all about falsifiability? The answer is that we can erect reliable rules for falsifying claims with severity. We corroborate their denials. If your statistical account denies we can reliably falsify interesting theories, it is irrelevant to real-world knowledge. Let me draw your attention to an exhibit on a strange disease, kuru, and how it falsified a fundamental dogma of biology.

Exhibit (v): Kuru. Kuru (which means "shaking") was widespread among the Fore people of New Guinea in the 1960s. In around 3–6 months, Kuru victims go from having difficulty walking, to outbursts of laughter, to inability to swallow and death. Kuru, and (what we now know to be) related diseases, e.g., mad cow, Creutzfeldt–Jakob, and scrapie, are "spongiform" diseases, causing brains to appear spongy. Kuru clustered in families, in particular among Fore women and their children, or elderly parents. They began to suspect transmission was through mortuary cannibalism. Consuming the brains of loved ones, a way of honoring the dead, was also a main source of meat permitted to women. Some say men got first dibs on the muscle; others deny men partook in these funerary practices. What we know is that ending these cannibalistic practices all but eradicated the disease. No one expected at the time that understanding kuru's cause would falsify an established theory that only viruses and bacteria could be infectious. This "central dogma of biology" says:

H: All infectious agents have nucleic acid.

Any infectious agent free of nucleic acid would be *anomalous* for *H* – meaning it goes against what *H* claims. A separate step is required to decide when *H*'s anomalies should count as falsifying *H*. There needn't be a cut-off so much as a standpoint as to when continuing to defend *H* becomes bad science. Prion researchers weren't looking to test the central dogma of biology, but to understand kuru and related diseases. The anomaly erupted only because kuru appeared to be transmitted by a protein alone, by changing a normal protein shape into an abnormal fold. Stanley Prusiner called the infectious protein a prion – for which he received much grief. He thought, at first, he'd made

a mistake "and was puzzled when the data kept telling me that our preparations contained protein but not nucleic acid" (Prusiner 1997). The anomalous results would not go away and, eventually, were demonstrated via experimental transmission to animals. The discovery of prions led to a "revolution" in molecular biology, and Prusiner received a Nobel prize in 1997. It is *logically* possible that nucleic acid is somehow involved. But continuing to block the falsification of H (i.e., block the "protein only" hypothesis) precludes learning more about prion diseases, which now include Alzheimer's. (See Mayo 2014a.)

Insofar as we falsify general scientific claims, we are all methodological falsificationists. Some people say, "I know my models are false, so I'm done with the job of falsifying before I even begin." Really? That's not falsifying. Let's look at your method: always infer that H is false, fails to solve its intended problem. Then you're bound to infer this even when this is erroneous. Your method fails the minimal severity requirement.

Do Probabilists Falsify? It isn't obvious a probabilist desires to falsify, rather than supply a probability measure indicating disconfirmation, the opposite of a B-boost (a B-bust?), or a low posterior. Members of some probabilist tribes propose that Popper is subsumed under a Bayesian account by taking a low value of $\Pr(\boldsymbol{x}|H)$ to falsify H. That could not work. Individual outcomes described in detail will easily have very small probabilities under H without being genuine anomalies for H. To the severe tester, this as an attempt to distract from the inability of probabilists to falsify, insofar as they remain probabilists. What about comparative accounts (Likelihoodists or Bayes factor accounts), which I also place under probabilism? Reporting that one hypothesis is more likely than the other is not to falsify anything. Royall is clear that it's wrong to even take the comparative report as evidence against one of the two hypotheses: they are not exhaustive. (Nothing turns on whether you prefer to put Likelihoodism under its own category.) Must all such accounts abandon the ability to falsify? No, they can *indirectly* falsify hypotheses by adding a methodological falsification rule. A natural candidate is to falsify H if its posterior probability is sufficiently low (or, perhaps, sufficiently disconfirmed). Of course, they'd need to justify the rule, ensuring it wasn't often mistaken.

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). Early on, Popper (1959) bases his statistical falsifying rules on Fisher, though citations are rare. Even where a scientific hypothesis is thought to be deterministic, inaccuracies and knowledge gaps involve error-laden predictions; so our methodological rules typically involve inferring a statistical hypothesis. Popper calls it a *falsifying hypothesis*. It's a hypothesis inferred in order to falsify some other claim. A first step is often to infer an anomaly is real, by falsifying a "due to chance" hypothesis.

The recognition that we need methodological rules to warrant falsification led Popperian Imre Lakatos to dub Popper's philosophy "methodological falsificationism" (Lakatos 1970, p. 106). If you look at this footnote, where Lakatos often buried gems, you read about "the philosophical basis of some of the most interesting developments in modern statistics. The Neyman–Pearson approach rests completely on methodological falsificationism" (ibid., p. 109, note 6). Still, neither he nor Popper made explicit use of N-P tests. Statistical hypotheses are the perfect tool for "falsifying hypotheses." However, this means you can't be a falsificationist and remain a strict deductivist. When statisticians (e.g., Gelman 2011) claim they are deductivists like Popper, I take it they mean they favor a testing account like Popper, rather than inductively building up probabilities. The falsifying hypotheses that are integral for Popper also necessitate an evidence-transcending (inductive) statistical inference.

This is hugely problematic for Popper because being a strict Popperian means never having to justify a claim as true or a method as reliable. After all, this was part of Popper's escape from induction. The problem is this: 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. This leads us to Pierre Duhem, physicist and philosopher of science.

Duhemian Problems of Falsification

Consider the simplest form of deductive falsification: If H entails observation O, and we observe $\sim O$, then we infer $\sim H$. To infer $\sim H$ is to infer H is false, or there is some discrepancy in what H claims about the phenomenon in

question. As with any argument, in order to *detach* its conclusion (without which there is no *inference*), the premises must be true or approximately true. But O is derived only with the help of various additional claims. In statistical contexts, we may group these under two umbrellas: auxiliary factors linking substantive and statistical claims, $A_1 \,\&\, \cdots \,\&\, A_n$, and assumptions of the statistical model $E_1 \,\&\, \cdots \,\&\, E_k$. You are to imagine a great big long conjunction of factors, in the following argument:

1. If $H \,\&\, (A_1 \,\&\, \cdots \,\&\, A_n) \,\&\, (E_1 \,\&\, \cdots \,\&\, E_k)$, then O.
2. $\sim O$.
3. Therefore, either $\sim H$ or $\sim A_1$ or . . . or $\sim A_n$ or $\sim E_1$ or . . . or $\sim E_k$.

This is an instance of deductively valid *modus tollens*. The catchall $\sim H$ itself is an exhaustive list of alternatives. This is too ugly for words. Philosophers, ever appealing to logic, often take this as the entity facing scientists who are left to fight their way through a great big disjunction: either H or one (or more) of the assumptions used in deriving observation claim O is to blame for anomaly $\sim O$.

When we are faced with an anomaly for H, Duhem argues, "The only thing the experiment teaches us is . . . there is at least one error; but where this error lies is just what it does not tell us" (Duhem 1954, p. 185). *Duhem's problem* is the problem of pinpointing what is warranted to blame for an observed anomaly with a claim H.

Bayesian philosophers deal with Duhem's problem by assigning each of the elements used to derive a prediction a prior probability. Whether H itself, or one of the A_i or E_k, is blamed is a matter of their posterior probabilities. Even if a failed prediction lowers the probability of hypothesis H, its posterior probability may still remain high, while the probability in A_{16}, say, drops down. The trouble is that one is free to tinker around with these assignments so that an auxiliary is blamed, and a main hypothesis H retained, or the other way around. Duhem's problem is what's really responsible for the anomaly (Mayo 1997a) – what's *warranted* to blame. On the other hand, the Bayesian approach is an excellent way to formally reconstruct Duhem's position. In his view, different researchers may choose to restore consistency according to their beliefs or to what Duhem called good sense, "bon sens." Popper was allergic to such a thing.

How can Popper, if he is really a deductivist, solve Duhem in order to falsify? At best he'd subject each of the conjuncts to as stringent a test as possible, and falsify accordingly. This still leaves, Popper admits, a disjunction of non-falsified hypotheses (he thought infinitely many)! Popperian philosophers of science advise you to choose a suitable overall package of hypotheses, assumptions, auxiliaries, on a set of criteria: simplicity, explanatory power, unification and so

on. There's no agreement on which, nor how to define them. On this view, you can't really solve Duhem, you accept or "prefer" (as Popper said) the large-scale research program or paradigm as a whole. It's intended to be an advance over *bon sens* in blocking certain types of tinkering (see Section 2.4). There's a remark in the Popper museum display I only recently came across:

> [W]e can be reasonably successful in attributing our refutations to definite portions of the theoretical maze. (For we *are* reasonably successful in this – a fact which must remain inexplicable for one who adopts Duhem's and Quine's view on the matter.) (1962, p. 243)

That doesn't mean he supplied an account for such attributions. He should have, but did not. There is a tendency to suppose Duhem's problem, like demarcation and induction, is insoluble and that it's taboo to claim to solve it. Our journey breaks with these taboos.

We should reject these formulations of Duhem's problem, starting with the great big conjunction in the antecedent of the conditional. It is vintage "rational reconstruction" of science, a very linear but straight-jacketed way to view the problem. Falsifying the central dogma of biology (infection requires nucleic acid) involved no series of conjunctions from *H* down to observations, but moving from *the bottom up*, as it were. The first clues that no nucleic acids were involved came from the fact that prions are not eradicated with techniques known to kill viruses and bacteria (e.g., UV irradiation, boiling, hospital disinfectants, hydrogen peroxide, and much else). If it were a mistake to regard prions as having no nucleic acid, then at least one of these known agents would have eradicated it. Further, prions are deactivated with substances known to kill proteins. Post-positive philosophers of science, many of them, are right to say philosophers need to pay more attention to experiments (a trend I call the New Experimentalism), but this must be combined with an account of statistical inference.

Frequentist statistics "allows interesting parts of a complicated problem to be broken off and solved separately" (Efron 1986, p. 4). We invent methods that take account of the effect of as many unknowns as possible, perhaps randomizing the rest. I never had to affirm that each and every one of my scales worked in my weighing example, the strong argument from coincidence lets me rule out, with severity, the possibility that accidental errors were producing precisely the same artifact in each case. Duhem famously compared the physicist to the doctor, as opposed to the watchmaker who can pull things apart. But the doctor may determine what it would be like if such and such were operative and *distinguish* the effects of different sources. The effect of violating an assumption of a constant mean looks very different from

a changing variance; despite all the causes of a sore throat, strep tests are quite reliable. Good research should at least be able to embark on inquiries to solve their Duhemian problems.

Popper Comes Up Short. Popper's account rests on severe tests, tests that would probably have falsified a claim if false, but he cannot warrant saying any such thing. High corroboration, Popper freely admits, is at most a report on past successes with little warrant for future reliability.

> Although Popper's work is full of exhortations to put hypotheses through the wringer, to make them "suffer in our stead in the struggle for the survival of the fittest" (Popper 1962, p. 52), the tests Popper sets out are white-glove affairs of logical analysis . . . it is little wonder that they seem to tell us only that there is an error somewhere and that they are silent about its source. We have to become shrewd inquisitors of errors, interact with them, simulate them (with models and computers), amplify them: we have to learn to make them talk. (Mayo 1996, p. 4)

Even to falsify non-trivial claims – as Popper grants – requires grounds for inferring a reliable effect. Singular observation statements will not do. We need "lift-off." Popper never saw how to solve the problem of "drag down" wherein empirical claims are only as reliable as the data involved in reaching them (Excursion 1). We cannot just pick up his or any other past account. Yet there's no reason to be hamstrung by the limits of the logical positivist or empiricist era. Scattered measurements are not of much use, but with adequate data massaging and averaging we can estimate a quantity of interest far more accurately than individual measurements. Recall Fisher's "it should never be true" in Exhibit (iii), Section 2.1. Fisher and Neyman–Pearson were ahead of Popper here (as was Peirce). When Popper wrote me "I regret not studying statistics," my thought was "not as much as I do."

Souvenir E: An Array of Questions, Problems, Models

> It is a fundamental contribution of modern mathematical statistics to have recognized the explicit need of a model in analyzing the significance of experimental data. (Suppes 1969, p. 33)

Our framework cannot abide by oversimplifications of accounts that blur statistical hypotheses and research claims, that ignore assumptions of data or limit the entry of background information to any one portal or any one form. So what do we do if we're trying to set out the problems of statistical inference? I appeal to a general account (Mayo 1996) that builds on Patrick Suppes' (1969) idea of a hierarchy of models between models of data, experiment, and theory. Trying to cash out a full-blown picture of inquiry that purports to represent all

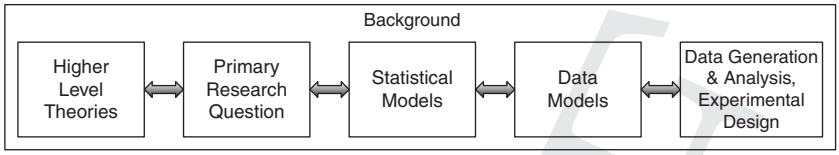

Figure 2.1 Array of questions, problems, models.

contexts of inquiry is a fool's errand. Or so I discovered after many years of trying. If one is not to land in a Rube Goldberg mess of arrows and boxes, only to discover it's not pertinent to every inquiry, it's best to settle for pigeonholes roomy enough to organize the interconnected pieces of a given inquiry as in Figure 2.1.

Loosely, there's an inferential move from the data model to the primary claim or question via the statistical test or inference model. Secondary questions include a variety of inferences involved in generating and probing conjectured answers to the primary question. A sample: How might we break down a problem into one or more local questions that can be probed with reasonable severity? How should we generate and model raw data, put them in canonical form, and check their assumptions? Remember, we are using "tests" to encompass probing any claim, including estimates. It's standard to distinguish "confirmatory" and "exploratory" contexts, but each is still an inferential or learning problem, although criteria for judging the solutions differ. In explorations, we may simply wish to infer that a model is worth developing further, that another is wildly off target.

Souvenir F: Getting Free of Popperian Constraints on Language

Popper allows that anyone who wants to define induction as the procedure of corroborating by severe testing is free to do so; and I do. Free of the bogeyman that induction must take the form of a probabilism, let's get rid of some linguistic peculiarities inherited by current-day Popperians (critical rationalists). They say things such as: it is *warranted* to infer (prefer or believe) H (because H has passed a severe test), but there is no *justification* for H (because "justifying" H would mean H was true or highly probable). In our language, if H passes a severe test, you can say it is warranted, corroborated, justified – along with whatever qualification is appropriate. I tend to use "warranted." The Popperian "hypothesis H is corroborated by data $\boldsymbol{x}$" is such a tidy abbreviation of "H has passed a severe test with $\boldsymbol{x}$" that we may use the two interchangeably. I've already co-opted Popper's description of science as *problem solving*. A hypothesis can be seen as a potential solution to

a problem (Laudan 1978). For example, the theory of protein folding purports to solve the problem of how pathological prions are transmitted. The problem might be to explain, to predict, to unify, to suggest new problems, etc. When we severely probe, it's not for falsity per se, but to investigate if a problem has been adequately solved by a model, method, or theory.

In rejecting probabilism, there is nothing to stop us from speaking of believing in *H*. It's not the direct output of a statistical inference. A post-statistical inference might be to believe a severely tested claim; disbelieve a falsified one. There are many different grounds for believing something. We may be tenacious in our beliefs in the face of given evidence; they may have other grounds, or be prudential. By the same token, talk of deciding to conclude, infer, prefer, or act can be fully epistemic in the sense of assessing evidence, warrant, and well-testedness. Popper, like Neyman and Pearson, employs such language because it allows talking about inference distinct from assigning probabilities to hypotheses. Failing to recognize this has created unnecessary combat.

Live Exhibit (vi): Revisiting Popper's Demarcation of Science. Here's an experiment: try shifting what Popper says about theories to a related claim about inquiries to find something out. To see what I have in mind, let's listen to an exchange between two fellow travelers over coffee at Statbucks.

Traveler 1: If mere logical falsifiability suffices for a theory to be scientific, then, we can't properly oust astrology from the scientific pantheon. Plenty of nutty theories have been falsified, so by definition they're scientific. Moreover, scientists aren't always looking to subject well-corroborated theories to "grave risk" of falsification.

Traveler 2: I've been thinking about this. On your first point, Popper confuses things by making it sound as if he's asking: *When is a theory unscientific?* What he is actually asking or should be asking is: *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 cavalier methods for their testing. Adhering to a falsified theory no matter what is poor science. Some areas have so much noise and/or flexibility that they can't or won't distinguish warranted from unwarranted explanations of failed predictions. Rivals may find flaws in one another's inquiry or model, but the criticism is not constrained by what's actually responsible. This is another way inquiries can become unscientific.[1]

[1] For example, astronomy, but not astrology, can reliably solve its Duhemian puzzles. Chapter 2, Mayo (1996), following my reading of Kuhn (1970) on "normal science."

She continues:

> On your second point, it's true that Popper talked of wanting to subject theories to grave risk of falsification. I suggest that it's really our *inquiries* into, or tests of, the theories that we want to subject to grave risk. The onus is on interpreters of data to show how they are countering the charge of a poorly run test. I admit this is a modification of Popper. One could reframe the entire demarcation problem as one of the characters of an inquiry or test.

She makes a good point. In addition to blocking inferences that fail the minimal requirement for severity:

> *A scientific inquiry or test: 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. Without solving, not merely embarking on, some Duhemian problems there are no interesting falsifications. The ability or inability to pin down the source of failed replications – a familiar occupation these days – speaks to the scientific credentials of an inquiry. At any given time, even in good sciences there are anomalies whose sources haven't been traced – unsolved Duhemian problems – generally at "higher" levels of the theory-data array. Embarking on solving these is the impetus for new conjectures. Checking test assumptions is part of working through the Duhemian maze. 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. Some physicists worry that highly theoretical realms can't be expected to be constrained by empirical data. Theoretical constraints are also important. We'll flesh out these ideas in future tours.

2.4 Novelty and Severity

> When you have put a lot of ideas together to make an elaborate theory, you want to make sure, when explaining what it fits, that those things it fits are not just the things that gave you the idea for the theory; but that the finished theory makes something else come out right, in addition. (Feynman 1974, p. 385)

This "something else that must come out right" is often called a "novel" predictive success. Whether or not novel predictive success is required is a very old battle that parallels debates between frequentists and inductive logicians, in both statistics and philosophy of science, for example, between Mill and Peirce