Excursion 1 How to Tell What's True about Statistical Inference

Itinerary

Tour I Beyond Probabilism and Performance

> I'm talking about a specific, extra type of integrity that is [beyond] not lying, but bending over backwards to show how you're maybe wrong, that you ought to have when acting as a scientist. (Feynman 1974/1985, p. 387)

It is easy to lie with statistics. Or so the cliché goes. It is also very difficult to uncover these lies without statistical methods – at least of the right kind. Self-correcting statistical methods are needed, and, with minimal technical fanfare, that's what I aim to illuminate. Since Darrell Huff wrote *How to Lie with Statistics* in 1954, ways of lying with statistics are so well worn as to have emerged in reverberating slogans:

- Association is not causation.
- Statistical significance is not substantive significance.
- No evidence of risk is not evidence of no risk.
- If you torture the data enough, they will confess.

Exposés of fallacies and foibles ranging from professional manuals and task forces to more popularized debunking treatises are legion. New evidence has piled up showing lack of replication and all manner of selection and publication biases. Even expanded "evidence-based" practices, whose very rationale is to emulate experimental controls, are not immune from allegations of illicit cherry picking, significance seeking, *P*-hacking, and assorted modes of extraordinary rendition of data. Attempts to restore credibility have gone far beyond the cottage industries of just a few years ago, to entirely new research programs: statistical fraud-busting, statistical forensics, technical activism, and widespread reproducibility studies. There are proposed methodological reforms – many are generally welcome (preregistration of experiments, transparency about data collection, discouraging mechanical uses of statistics), some are quite radical. If we are to appraise these evidence policy reforms, a much better grasp of some central statistical problems is needed.

Getting Philosophical

Are philosophies about science, evidence, and inference relevant here? Because the problems involve questions about uncertain evidence, probabilistic models, science, and pseudoscience – all of which are intertwined with technical

statistical concepts and presuppositions – they certainly ought to be. Even in an open-access world in which we have become increasingly fearless about taking on scientific complexities, a certain trepidation and groupthink take over when it comes to philosophically tinged notions such as inductive reasoning, objectivity, rationality, and science versus pseudoscience. The general area of philosophy that deals with knowledge, evidence, inference, and rationality is called *epistemology*. The epistemological standpoints of leaders, be they philosophers or scientists, are too readily taken as canon by others. We want to understand what's true about some of the popular memes: "All models are false," "Everything is equally subjective and objective," "*P*-values exaggerate evidence," and "[M]ost published research findings are false" (Ioannidis 2005) – at least if you publish a single statistically significant result after data finagling. (Do people do that? Shame on them.) Yet R. A. Fisher, founder of modern statistical tests, denied that an isolated statistically significant result counts.

> [W]e need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result. (Fisher 1935b/1947, p. 14)

Satisfying this requirement depends on the proper use of background knowledge and deliberate design and modeling.

This opening excursion will launch us into the main themes we will encounter. You mustn't suppose, by its title, that I will be talking about how to tell the truth using statistics. Although I expect to make some progress there, my goal is to tell what's true about statistical methods themselves! There are so many misrepresentations of those methods that telling what is true about them is no mean feat. It may be thought that the basic statistical concepts are well understood. But I show that this is simply not true.

Nor can you just open a statistical text or advice manual for the goal at hand. The issues run deeper. Here's where I come in. Having long had one foot in philosophy of science and the other in foundations of statistics, I will zero in on the central philosophical issues that lie below the surface of today's raging debates. "Getting philosophical" is not about articulating rarified concepts divorced from statistical practice. It is to provide tools to avoid obfuscating the terms and issues being bandied about. Readers should be empowered to understand the core presuppositions on which rival positions are based – and on which they depend.

Do I hear a protest? "There is nothing philosophical about our criticism of statistical significance tests" (someone might say). The problem is that a small *P*-value is invariably, and erroneously, interpreted as giving a small probability

to the null hypothesis." Really? *P*-values are not intended to be used this way; presupposing they ought to be so interpreted grows out of a specific conception of the role of probability in statistical inference. *That conception is philosophical.* Methods characterized through the lens of over-simple epistemological orthodoxies are methods misapplied and mischaracterized. This may lead one to lie, however unwittingly, about the nature and goals of statistical inference, when what we want is to tell what's true about them.

1.1 Severity Requirement: Bad Evidence, No Test (BENT)

Fisher observed long ago, "[t]he political principle that anything can be proved by statistics arises from the practice of presenting only a selected subset of the data available" (Fisher 1955, p. 75). If you report results selectively, it becomes easy to prejudge hypotheses: yes, the data may accord amazingly well with a hypothesis *H*, but such a method is practically guaranteed to issue so good a fit even if *H* is false and not warranted by the evidence. If it is predetermined that a way will be found to either obtain or interpret data as evidence for *H*, then data are not being taken seriously in appraising *H*. *H* is essentially immune to having its flaws uncovered by the data. *H* might be said to have "passed" the test, but it is a test that lacks stringency or severity. Everyone understands that this is bad evidence, or no test at all. I call this the *severity requirement*. In its weakest form it supplies a *minimal requirement* for evidence:

> *Severity Requirement (weak): One does not have evidence for a claim if nothing has been done to rule out ways the claim may be false.* If data ***x*** agree with a claim *C* but the method used is practically guaranteed to find such agreement, and had little or no capability of finding flaws with *C* even if they exist, then we have bad evidence, no test (BENT).

The "practically guaranteed" acknowledges that even if the method had some slim chance of producing a disagreement when *C* is false, we still regard the evidence as lousy. Little if anything has been done to rule out erroneous construals of data. We'll need many different ways to state this minimal principle of evidence, depending on context.

A Scandal Involving Personalized Medicine

A recent scandal offers an example. Over 100 patients signed up for the chance to participate in the Duke University (2007–10) clinical trials that promised a custom-tailored cancer treatment. A cutting-edge prediction model

developed by Anil Potti and Joseph Nevins purported to predict your response to one or another chemotherapy based on large data sets correlating properties of various tumors and positive responses to different regimens (Potti et al. 2006). Gross errors and data manipulation eventually forced the trials to be halted. It was revealed in 2014 that a whistleblower – a student – had expressed concerns that

> . . . in developing the model, only those samples which fit the model best in cross validation were included. Over half of the original samples were removed. . . . This was an incredibly biased approach. (Perez 2015)

In order to avoid the overly rosy predictions that ensue from a model built to fit the data (called the training set), a portion of the data (called the test set) is to be held out to "cross validate" the model. If any unwelcome test data are simply excluded, the technique has obviously not done its job. Unsurprisingly, when researchers at a different cancer center, Baggerly and Coombes, set out to avail themselves of this prediction model, they were badly disappointed: "When we apply the same methods but maintain the separation of training and test sets, predictions are poor" (Coombes et al. 2007, p. 1277). Predicting which treatment would work was no better than chance.

You might be surprised to learn that Potti dismissed their failed replication on grounds that they didn't use his method (Potti and Nevins 2007)! But his technique had little or no ability to reveal the unreliability of the model, and thus failed utterly as a cross check. By contrast, Baggerly and Coombes' approach informed about what it *would be like* to apply the model to brand new patients – the intended function of the cross validation. Medical journals were reluctant to publish Baggerly and Coombes' failed replications and report of critical flaws. (It eventually appeared in a statistics journal, *Annals of Applied Statistics* 2009, thanks to editor Brad Efron.) The clinical trials – yes on patients – were only shut down when it was discovered Potti had exaggerated his honors in his CV! The bottom line is, tactics that stand in the way of discovering weak spots, whether for prediction or explanation, create obstacles to the severity requirement; it would be puzzling if accounts of statistical inference failed to place this requirement, or something akin to it, right at the center – or even worse, permitted loopholes to enable such moves. Wouldn't it?

Do We Always Want to Find Things Out?

The severity requirement gives a minimal principle based on the fact that highly insevere tests yield bad evidence, no tests (BENT). We can all agree on this much, I think. We will explore how much mileage we can get from it. It applies at a number of junctures in collecting and modeling data, in linking

data to statistical inference, and to substantive questions and claims. This will be our linchpin for understanding what's true about statistical inference. In addition to our minimal principle for evidence, one more thing is needed, at least during the time we are engaged in this project: *the goal of finding things out*.

The desire to find things out is an obvious goal; yet most of the time it is not what drives us. We typically may be uninterested in, if not quite resistant to, finding flaws or incongruencies with ideas we like. Often it is entirely proper to gather information to make your case, and ignore anything that fails to support it. Only if you really desire to find out something, or to challenge so-and-so's ("trust me") assurances, will you be prepared to stick your (or their) neck out to conduct a genuine "conjecture and refutation" exercise. Because you want to learn, you will be prepared to risk the possibility that the conjecture is found flawed.

We hear that "motivated reasoning has interacted with tribalism and new media technologies since the 1990s in unfortunate ways" (Haidt and Iyer 2016). Not only do we see things through the tunnel of our tribe, social media and web searches enable us to live in the echo chamber of our tribe more than ever. We might think we're trying to find things out but we're not. Since craving truth is rare (unless your life depends on it) and the "perverse incentives" of publishing novel results so shiny, the wise will invite methods that make uncovering errors and biases as quick and painless as possible. Methods of inference that fail to satisfy the minimal severity requirement fail us in an essential way.

With the rise of Big Data, data analytics, machine learning, and bioinformatics, statistics has been undergoing a good deal of introspection. Exciting results are often being turned out by researchers without a traditional statistics background; biostatistician Jeff Leek (2016) explains: "There is a structural reason for this: data was sparse when they were trained and there wasn't any reason for them to learn statistics." The problem goes beyond turf battles. It's discovering that many data analytic applications are missing key ingredients of statistical thinking. Brown and Kass (2009) crystalize its essence. "Statistical thinking uses probabilistic descriptions of variability in (1) inductive reasoning and (2) analysis of procedures for data collection, prediction, and scientific inference" (p. 107). A word on each.

(1) Types of statistical inference are too varied to neatly encompass. Typically we employ data to learn something about the process or mechanism producing the data. The claims inferred are not specific events, but statistical generalizations, parameters in theories and models, causal claims, and general predictions. Statistical inference goes beyond the data – by definition that

makes it an *inductive* inference. The risk of error is to be expected. There is no need to be reckless. The secret is controlling and learning from error. Ideally we take precautions in advance: *pre-data*, we devise methods that make it hard for claims to pass muster unless they are approximately true or adequately solve our problem. With data in hand, *post-data*, we scrutinize what, if anything, can be inferred.

What's the essence of analyzing procedures in (2)? Brown and Kass don't specifically say, but the gist can be gleaned from what vexes them; namely, ad hoc data analytic algorithms where researchers "have done nothing to indicate that it performs well" (p. 107). Minimally, statistical thinking means never ignoring the fact that there are alternative methods: Why is this one a good tool for the job? Statistical thinking requires stepping back and examining a method's capabilities, whether it's designing or choosing a method, or scrutinizing the results.

A Philosophical Excursion

Taking the severity principle then, along with the aim that we desire to find things out without being obstructed in this goal, let's set sail on a philosophical excursion to illuminate statistical inference. Envision yourself embarking on a special interest cruise featuring "exceptional itineraries to popular destinations worldwide as well as unique routes" (Smithsonian Journeys). What our cruise lacks in glamour will be more than made up for in our ability to travel back in time to hear what Fisher, Neyman, Pearson, Popper, Savage, and many others were saying and thinking, and then zoom forward to current debates. There will be exhibits, a blend of statistics, philosophy, and history, and even a bit of theater. Our standpoint will be pragmatic in this sense: my interest is not in some ideal form of knowledge or rational agency, no omniscience or God's-eye view – although we'll start and end surveying the landscape from a hot-air balloon. I'm interested in the problem of how we get the kind of knowledge we do manage to obtain – and how we can get more of it. Statistical methods should not be seen as tools for what philosophers call "rational reconstruction" of a piece of reasoning. Rather, they are forward-looking tools to find something out faster and more efficiently, and to discriminate how good or poor a job others have done.

The job of the philosopher is to clarify but also to provoke reflection and scrutiny precisely in those areas that go unchallenged in ordinary practice. My focus will be on the issues having the most influence, and being most liable to obfuscation. Fortunately, that doesn't require an abundance of technicalities, but you can opt out of any daytrip that appears too technical: an idea not

caught in one place should be illuminated in another. Our philosophical excursion may well land us in positions that are provocative to all existing sides of the debate about probability and statistics in scientific inquiry.

Methodology and Meta-methodology

We are studying statistical methods from various schools. What shall we call methods for doing so? Borrowing a term from philosophy of science, we may call it our meta-methodology – it's one level removed.[1] To put my cards on the table: A severity scrutiny is going to be a key method of our meta-methodology. It is fairly obvious that we want to scrutinize how capable a statistical method is at detecting and avoiding erroneous interpretations of data. So when it comes to the role of probability as a pedagogical tool for our purposes, severity – its assessment and control – will be at the center. The term "severity" is Popper's, though he never adequately defined it. It's not part of any statistical methodology as of yet. Viewing statistical inference as severe testing lets us stand one level removed from existing accounts, where the air is a bit clearer.

Our intuitive, minimal, requirement for evidence connects readily to formal statistics. The probabilities that a statistical method lands in erroneous interpretations of data are often called its *error probabilities*. So an account that revolves around control of error probabilities I call an *error statistical account*. But "error probability" has been used in different ways. Most familiar are those in relation to hypotheses tests (Type I and II errors), significance levels, confidence levels, and power – all of which we will explore in detail. It has occasionally been used in relation to the proportion of false hypotheses among those now in circulation, which is different. For now it suffices to say that none of the formal notions directly give severity assessments. There isn't even a statistical school or tribe that has explicitly endorsed this goal. I find this perplexing. That will not preclude our immersion into the mindset of a futuristic tribe whose members use error probabilities for assessing severity; it's just the ticket for our task: understanding and getting beyond the statistics wars. We may call this tribe the *severe testers*.

We can keep to testing language. See it as part of the meta-language we use to talk about formal statistical methods, where the latter include estimation, exploration, prediction, and data analysis. I will use the term "hypothesis," or just "claim," for any conjecture we wish to entertain; it need not be one set out in advance of data. Even predesignating hypotheses, by the way, doesn't

[1] This contrasts with the use of "metaresearch" to describe work on methodological reforms by non-philosophers. This is not to say they don't tread on philosophical territory often: they do.

preclude bias: that view is a holdover from a crude empiricism that assumes data are unproblematically "given," rather than selected and interpreted. Conversely, using the same data to arrive at and test a claim can, in some cases, be accomplished with stringency.

As we embark on statistical foundations, we must avoid blurring formal terms such as probability and likelihood with their ordinary English meanings. Actually, "probability" comes from the Latin *probare*, meaning to try, test, or prove. "Proof" in "The proof is in the pudding" refers to how you put something to the test. You must show or demonstrate, not just believe strongly. Ironically, using probability this way would bring it very close to the idea of measuring well-testedness (or how well shown). But it's not our current, informal English sense of probability, as varied as that can be. To see this, consider "improbable." Calling a claim improbable, in ordinary English, can mean a host of things: I bet it's not so; all things considered, given what I know, it's implausible; and other things besides. Describing a claim as *poorly tested* generally means something quite different: little has been done to probe whether the claim holds or not, the method used was highly unreliable, or things of that nature. In short, our informal notion of poorly tested comes rather close to the lack of severity in statistics. There's a difference between finding H poorly tested by data $\boldsymbol{x}$, and finding $\boldsymbol{x}$ renders H improbable – in any of the many senses the latter takes on. The existence of a Higgs particle was thought to be probable if not necessary before it was regarded as well tested around 2012. Physicists had to show or demonstrate its existence for it to be well tested. It follows that you are free to pursue our testing goal without implying there are no other statistical goals. One other thing on language: I will have to retain the terms currently used in exploring them. That doesn't mean I'm in favor of them; in fact, I will jettison some of them by the end of the journey.

To sum up this first tour so far, statistical inference uses data to reach claims about aspects of processes and mechanisms producing them, accompanied by an assessment of the properties of the inference methods: their capabilities to control and alert us to erroneous interpretations. We need to report if the method has satisfied the most minimal requirement for solving such a problem. Has anything been tested with a modicum of severity, or not? The severe tester also requires reporting of what has been poorly probed, and highlights the need to "bend over backwards," as Feynman puts it, to admit where weaknesses lie. In formal statistical testing, the crude dichotomy of "pass/fail" or "significant or not" will scarcely do. We must determine the magnitudes (and directions) of any statistical discrepancies warranted, and the limits to any

substantive claims you may be entitled to infer from the statistical ones. Using just our minimal principle of evidence, and a sturdy pair of shoes, join me on a tour of statistical inference, back to the leading museums of statistics, and forward to current offshoots and statistical tribes.

Why We Must Get Beyond the Statistics Wars

Some readers may be surprised to learn that the field of statistics, arid and staid as it seems, has a fascinating and colorful history of philosophical debate, marked by unusual heights of passion, personality, and controversy for at least a century. Others know them all too well and regard supporting any one side largely as proselytizing. I've heard some refer to statistical debates as "theological." I do not want to rehash the "statistics wars" that have raged in every decade, although the significance test controversy is still hotly debated among practitioners, and even though each generation fights these wars anew – with task forces set up to stem reflexive, recipe-like statistics that have long been deplored.

The time is ripe for a fair-minded engagement in the debates about statistical foundations; more than that, it is becoming of pressing importance. Not only because

> (i) these issues are increasingly being brought to bear on some very public controversies;

nor because

> (ii) the "statistics wars" have presented new twists and turns that cry out for fresh analysis

– as important as those facets are – but because what is at stake is a critical standpoint that we may be in danger of losing. Without it, we forfeit the ability to communicate with, and hold accountable, the "experts," the agencies, the quants, and all those data handlers increasingly exerting power over our lives. Understanding the nature and basis of statistical inference must not be considered as all about mathematical details; it is at the heart of what it means to reason scientifically and with integrity about any field whatever. Robert Kass (2011) puts it this way:

> We care about our philosophy of statistics, first and foremost, because statistical inference sheds light on an important part of human existence, inductive reasoning, and we want to understand it. (p. 19)

Isolating out a particular conception of statistical inference as severe testing is a way of telling what's true about the statistics wars, and getting beyond them.