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

https://doi.org/10.1017/9781107286184.002 Published online by Cambridge University Press

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 $\boldsymbol{x}$ 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

https://doi.org/10.1017/9781107286184.002 Published online by Cambridge University Press

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

https://doi.org/10.1017/9781107286184.002 Published online by Cambridge University Press

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.

Chutzpah, No Proselytizing

Our task is twofold: not only must we analyze statistical methods; we must also scrutinize the jousting on various sides of the debates. Our meta-level standpoint will let us rise above much of the cacophony; but the excursion will involve a dose of chutzpah that is out of the ordinary in professional discussions. You will need to critically evaluate the texts and the teams of critics, including brilliant leaders, high priests, maybe even royalty. Are they asking the most unbiased questions in examining methods, or are they like admen touting their brand, dragging out howlers to make their favorite method look good? (I am not sparing any of the statistical tribes here.) There are those who are earnest but brainwashed, or are stuck holding banners from an earlier battle now over; some are wedded to what they've learned, to what's in fashion, to what pays the rent.

Some are so jaundiced about the abuses of statistics as to wonder at my admittedly herculean task. I have a considerable degree of sympathy with them. But, I do not sympathize with those who ask: "why bother to clarify statistical concepts if they are invariably misinterpreted?" and then proceed to misinterpret them. Anyone is free to dismiss statistical notions as irrelevant to them, but then why set out a shingle as a "statistical reformer"? You may even be shilling for one of the proffered reforms, thinking it the road to restoring credibility, when it will do nothing of the kind.

You might say, since rival statistical methods turn on issues of philosophy and on rival conceptions of scientific learning, that it's impossible to say anything "true" about them. You just did. It's precisely these interpretative and philosophical issues that I plan to discuss. Understanding the issues is different from settling them, but it's of value nonetheless. Although statistical disagreements involve philosophy, statistical practitioners and not philosophers are the ones leading today's discussions of foundations. Is it possible to pursue our task in a way that will be seen as neither too philosophical nor not philosophical enough? Too statistical or not statistically sophisticated enough? Probably not, I expect grievances from both sides.

Finally, I will not be proselytizing for a given statistical school, so you can relax. Frankly, they all have shortcomings, insofar as one can even glean a clear statement of a given statistical "school." What we have is more like a jumble with tribal members often speaking right past each other. View the severity requirement as a heuristic tool for telling what's true about statistical controversies. Whether you resist some of the ports of call we arrive at is unimportant; it suffices that visiting them provides a key to unlock current mysteries that are leaving many consumers and students of statistics in the dark about a crucial portion of science.

1.2 Probabilism, Performance, and Probativeness

> I shall be concerned with the foundations of the subject. But in case it should be thought that this means I am not here strongly concerned with practical applications, let me say right away that confusion about the foundations of the subject is responsible, in my opinion, for much of the misuse of the statistics that one meets in fields of application such as medicine, psychology, sociology, economics, and so forth. (George Barnard 1985, p. 2)

While statistical science (as with other sciences) generally goes about its business without attending to its own foundations, implicit in every statistical methodology are core ideas that direct its principles, methods, and interpretations. I will call this its *statistical philosophy*. To tell what's true about statistical inference, understanding the associated philosophy (or philosophies) is essential. Discussions of statistical foundations tend to focus on how to interpret probability, and much less on the overarching question of how probability ought to be used in inference. Assumptions about the latter lurk implicitly behind debates, but rarely get the limelight. If we put the spotlight on them, we see that there are two main philosophies about the roles of probability in statistical inference: We may dub them *performance* (in the long run) and *probabilism*.

The performance philosophy sees the key function of statistical method as controlling the relative frequency of erroneous inferences in the long run of applications. For example, a frequentist statistical test, in its naked form, can be seen as a rule: whenever your outcome exceeds some value (say, $\boldsymbol{X} > \boldsymbol{x}^*$), reject a hypothesis H_0 and infer H_1. The value of the rule, according to its performance-oriented defenders, is that it can ensure that, regardless of which hypothesis is true, there is both a low probability of erroneously rejecting H_0 (rejecting H_0 when it is true) as well as erroneously accepting H_0 (failing to reject H_0 when it is false).

The second philosophy, probabilism, views probability as a way to assign degrees of belief, support, or plausibility to hypotheses. Many keep to a comparative report, for example that H_0 is more believable than is H_1 given data $\boldsymbol{x}$; others strive to say H_0 is less believable given data $\boldsymbol{x}$ than before, and offer a quantitative report of the difference.

What happened to the goal of scrutinizing BENT science by the severity criterion? Neither "probabilism" nor "performance" directly captures that demand. To take these goals at face value, it's easy to see why they come up short. Potti and Nevins' strong belief in the reliability of their prediction model for cancer therapy scarcely made up for the shoddy testing. Neither is good long-run performance a sufficient condition. Most obviously, there may be no

https://doi.org/10.1017/9781107286184.002 Published online by Cambridge University Press

long-run repetitions, and our interest in science is often just the particular statistical inference before us. Crude long-run requirements may be met by silly methods. Most importantly, good performance alone fails to get at *why* methods work when they do; namely – I claim – to let us assess and control the stringency of tests. This is the key to answering a burning question that has caused major headaches in statistical foundations: why should a low relative frequency of error matter to the appraisal of the inference at hand? It is not probabilism or performance we seek to quantify, but *probativeness*.

I do not mean to disparage the long-run performance goal – there are plenty of tasks in inquiry where performance is absolutely key. Examples are screening in high-throughput data analysis, and methods for deciding which of tens of millions of collisions in high-energy physics to capture and analyze. New applications of machine learning may lead some to say that only low rates of prediction or classification errors matter. Even with prediction, "black-box" modeling, and non-probabilistic inquiries, there is concern with solving a problem. We want to know if a good job has been done in the case at hand.

Severity (Strong): Argument from Coincidence

The weakest version of the severity requirement (Section 1.1), in the sense of easiest to justify, is negative, warning us when BENT data are at hand, and a surprising amount of mileage may be had from that negative principle alone. It is when we recognize how poorly certain claims are warranted that we get ideas for improved inquiries. In fact, if you wish to stop at the negative requirement, you can still go pretty far along with me. I also advocate the positive counterpart:

> *Severity (strong): We have evidence for a claim C just to the extent it survives a stringent scrutiny.* If *C* passes a test that was highly capable of finding flaws or discrepancies from *C*, and yet none or few are found, then the passing result, $\boldsymbol{x}$, is evidence for *C*.

One way this can be achieved is by an *argument from coincidence*. The most vivid cases occur outside formal statistics.

Some of my strongest examples tend to revolve around my weight. Before leaving the USA for the UK, I record my weight on two scales at home, one digital, one not, and the big medical scale at my doctor's office. Suppose they are well calibrated and nearly identical in their readings, and they also all pick up on the extra 3 pounds when I'm weighed carrying three copies of my 1-pound book, *Error and the Growth of Experimental Knowledge* (EGEK). Returning from the UK, to my astonishment, not one but all three scales

https://doi.org/10.1017/9781107286184.002 Published online by Cambridge University Press

show anywhere from a 4–5 pound gain. There's no difference when I place the three books on the scales, so I must conclude, unfortunately, that I've gained around 4 pounds. Even for me, that's a lot. I've surely falsified the supposition that I lost weight! From this informal example, we may make two rather obvious points that will serve for less obvious cases. First, there's the idea I call lift-off.

> *Lift-off: An overall inference can be more reliable and precise than its premises individually.*

Each scale, by itself, has some possibility of error, and limited precision. But the fact that all of them have me at an over 4-pound gain, while none show any difference in the weights of EGEK, pretty well seals it. Were one scale off balance, it would be discovered by another, and would show up in the weighing of books. They cannot all be systematically misleading just when it comes to objects of unknown weight, can they? Rejecting a conspiracy of the scales, I conclude I've gained weight, at least 4 pounds. We may call this an *argument from coincidence*, and by its means we can attain lift-off. Lift-off runs directly counter to a seemingly obvious claim of drag-down.

> *Drag-down: An overall inference is only as reliable/precise as is its weakest premise.*

The drag-down assumption is common among empiricist philosophers: As they like to say, "It's turtles all the way down." Sometimes our inferences do stand as a kind of tower built on linked stones – if even one stone fails they all come tumbling down. Call that a *linked* argument.

Our most prized scientific inferences would be in a very bad way if piling on assumptions invariably leads to weakened conclusions. Fortunately we also can build what may be called *convergent* arguments, where lift-off is attained. This seemingly banal point suffices to combat some of the most well entrenched skepticisms in philosophy of science. And statistics happens to be the science par excellence for demonstrating lift-off!

Now consider what justifies my weight conclusion, based, as we are supposing it is, on a strong argument from coincidence. No one would say: "I can be assured that by following such a procedure, in the long run I would rarely report weight gains erroneously, but I can tell nothing from these readings about my weight now." To justify my conclusion by long-run performance would be absurd. Instead we say that the procedure had enormous capacity to reveal if any of the scales were wrong, and from this I argue about the source of the readings: *H*: I've gained weight. Simple as that. It would be a preposterous coincidence if none of

the scales registered even slight weight shifts when weighing objects of known weight, and yet were systematically misleading when applied to my weight. You see where I'm going with this. This is the key – granted with a homely example – that can fill a very important gap in frequentist foundations: Just because an account is touted as having a long-run rationale, it does not mean it lacks a short run rationale, or even one relevant for the particular case at hand.

Nor is it merely the improbability of all the results were H false; it is rather like denying an evil demon has read my mind just in the cases where I do not know the weight of an object, and deliberately deceived me. The argument to "weight gain" is an example of an argument from coincidence to the absence of an error, what I call:

> *Arguing from Error*: There is evidence an error is absent to the extent that a procedure with a very high capability of signaling the error, if and only if it is present, nevertheless detects no error.

I am using "signaling" and "detecting" synonymously: It is important to keep in mind that we don't know if the test output is correct, only that it gives a signal or alert, like sounding a bell. Methods that enable strong arguments to the absence (or presence) of an error I call *strong error probes*. Our ability to develop strong arguments from coincidence, I will argue, is the basis for solving the "problem of induction."

Glaring Demonstrations of Deception

Intelligence is indicated by a capacity for deliberate deviousness. Such deviousness becomes self-conscious in inquiry: An example is the use of a placebo to find out what it would be like if the drug has no effect. What impressed me the most in my first statistics class was the demonstration of how apparently impressive results are readily produced when nothing's going on, i.e., "by chance alone." Once you see how it is done, and done easily, there is no going back. The toy hypotheses used in statistical testing are nearly always overly simple as scientific hypotheses. But when it comes to framing rather blatant deceptions, they are just the ticket!

When Fisher offered Muriel Bristol-Roach a cup of tea back in the 1920s, she refused it because he had put the milk in first. What difference could it make? Her husband and Fisher thought it would be fun to put her to the test (1935a). Say she doesn't claim to get it right all the time but does claim that she has some genuine discerning ability. Suppose Fisher subjects her to 16 trials and she gets 9 of them right. Should I be impressed or not? By a simple experiment of randomly assigning milk first/tea first Fisher sought to answer

this stringently. But don't be fooled: a great deal of work goes into controlling biases and confounders before the experimental design can work. The main point just now is this: so long as lacking ability is sufficiently like the canonical "coin tossing" (Bernoulli) model (with the probability of success at each trial of 0.5), we can learn from the test procedure. In the Bernoulli model, we record success or failure, assume a fixed probability of success θ on each trial, and that trials are independent. If the probability of getting even more successes than she got, merely by guessing, is fairly high, there's little indication of special tasting ability. The probability of at least 9 of 16 successes, even if $\theta = 0.5$, is 0.4. To abbreviate, Pr(at least 9 of 16 successes; H_0: $\theta = 0.5$) = 0.4. This is the *P*-value of the observed difference; an unimpressive 0.4. You'd expect as many or even more "successes" 40% of the time merely by guessing. It's also the *significance level attained* by the result. (I often use *P*-value as it's shorter.) Muriel Bristol-Roach pledges that if her performance may be regarded as scarcely better than guessing, then she hasn't shown her ability. Typically, a small value such as 0.05, 0.025, or 0.01 is required.

Such artificial and simplistic statistical hypotheses play valuable roles at stages of inquiry where what is needed are blatant standards of "nothing's going on." There is no presumption of a metaphysical chance agency, just that there is expected variability – otherwise one test would suffice – and that probability models from games of chance can be used to distinguish genuine from spurious effects. Although the goal of inquiry is to find things out, the hypotheses erected to this end are generally approximations and may be deliberately false. To present statistical hypotheses as identical to substantive scientific claims is to mischaracterize them. We want to tell what's true about statistical inference. Among the most notable of these truths is:

> *P*-values can be readily invalidated due to how the data (or hypotheses!) are generated or selected for testing.

If you fool around with the results afterwards, reporting only successful guesses, your report will be invalid. You may claim it's very difficult to get such an impressive result due to chance, when in fact it's very easy to do so, with selective reporting. Another way to put this: your *computed P*-value is small, but the *actual P*-value is high! Concern with spurious findings, while an ancient problem, is considered sufficiently serious to have motivated the American Statistical Association to issue a guide on how not to interpret *P*-values (Wasserstein and Lazar 2016); hereafter, ASA 2016 Guide. It may seem that if a statistical account is free to ignore such fooling around then the problem disappears! It doesn't.

https://doi.org/10.1017/9781107286184.002 Published online by Cambridge University Press

Incidentally, Bristol-Roach got all the cases correct, and thereby taught her husband a lesson about putting her claims to the test.

Peirce

The philosopher and astronomer C. S. Peirce, writing in the late nineteenth century, is acknowledged to have anticipated many modern statistical ideas (including randomization and confidence intervals). Peirce describes how "so accomplished a reasoner" as Dr. Playfair deceives himself by a technique we know all too well – scouring the data for impressive regularities (2.738). Looking at the specific gravities of three forms of carbon, Playfair seeks and discovers a formula that holds for all of them (each is a root of the atomic weight of carbon, which is 12). Can this regularity be expected to hold in general for metalloids? It turns out that half of the cases required Playfair to modify the formula after the fact. If one limits the successful instances to ones where the formula was predesignated, and not altered later on, only half satisfy Playfair's formula. Peirce asks, how often would such good agreement be found due to chance? Again, should we be impressed?

Peirce introduces a mechanism to arbitrarily pair the specific gravity of a set of elements with the atomic weight of another. By design, such agreements could only be due to the chance pairing. Lo and behold, Peirce finds about the same number of cases that satisfy Playfair's formula. "It thus appears that there is no more frequent agreement with Playfair's proposed law than what is due to chance" (2.738).

At first Peirce's demonstration seems strange. He introduces an accidental pairing just to simulate the ease of obtaining so many agreements in an entirely imaginary situation. Yet that suffices to show Playfair's evidence is BENT. The popular inductive accounts of his time, Peirce argues, do not prohibit adjusting the formula to fit the data, and, because of that, they would persist in Playfair's error. The same debate occurs today, as when Anil Potti (of the Duke scandal) dismissed the whistleblower Perez thus: "we likely disagree with what constitutes validation" (Nevins and Potti 2015). Erasing genomic data that failed to fit his predictive model was justified, Potti claimed, by the fact that other data points fit (Perez 2015)! Peirce's strategy, as that of Coombes et al., is to introduce a blatant standard to put the method through its paces, without bogus agreements. If the agreement is no better than bogus agreement, we deny there is evidence for a genuine regularity or valid prediction. Playfair's formula may be true, or probably true, but Peirce's little demonstration is enough to show his method did a lousy job of testing it.

Texas Marksman

Take an even simpler and more blatant argument of deception. It is my favorite: the Texas Marksman. A Texan wants to demonstrate his shooting prowess. He shoots all his bullets any old way into the side of a barn and then paints a bull's-eye in spots where the bullet holes are clustered. This fails utterly to severely test his marksmanship ability. When some visitors come to town and notice the incredible number of bull's-eyes, they ask to meet this marksman and are introduced to a little kid. How'd you do so well, they ask? Easy, I just drew the bull's-eye around the most tightly clustered shots. There is impressive "agreement" with shooting ability, he might even compute how improbably so many bull's-eyes would occur by chance. Yet his ability to shoot was not tested in the least by this little exercise. There's a real effect all right, but it's not caused by his marksmanship! It serves as a potent analogy for a cluster of formal statistical fallacies from data-dependent findings of "exceptional" patterns.

The term "apophenia" refers to a tendency to zero in on an apparent regularity or cluster within a vast sea of data and claim a genuine regularity. One of our fundamental problems (and skills) is that we're apopheniacs. Some investment funds, none that we actually know, are alleged to produce several portfolios by random selection of stocks and send out only the one that did best. Call it the Pickrite method. They want you to infer that it would be a preposterous coincidence to get so great a portfolio if the Pickrite method were like guessing. So their methods are genuinely wonderful, or so you are to infer. If this had been their only portfolio, the probability of doing so well by luck is low. But the probability of at least one of many portfolios doing so well (even if each is generated by chance) is high, if not guaranteed.

Let's review the rogues' gallery of glaring arguments from deception. The lady tasting tea showed how a statistical model of "no effect" could be used to amplify our ordinary capacities to discern if something really unusual is going on. The *P*-value is the probability of at least as high a success rate as observed, assuming the test or null hypothesis, the probability of success is 0.5. Since even more successes than she got is fairly frequent through guessing alone (the *P*-value is moderate), there's poor evidence of a genuine ability. The Playfair and Texas sharpshooter examples, while quasi-formal or informal, demonstrate how to invalidate reports of significant effects. They show how gambits of post-data adjustments or selection can render a method highly capable of spewing out impressive looking fits even when it's just random noise.

> We appeal to the same statistical reasoning to show the problematic cases as to show genuine arguments from coincidence.

So am I proposing that a key role for statistical inference is to identify ways to spot egregious deceptions (BENT cases) and create strong arguments from coincidence? Yes, I am.

Spurious P-values and Auditing

In many cases you read about you'd be right to suspect that someone has gone circling shots on the side of a barn. Confronted with the statistical news flash of the day, your first question is: Are the results due to selective reporting, cherry picking, or any number of other similar ruses? This is a central part of what we'll call *auditing* a significance level.

A key point too rarely appreciated: Statistical facts about *P*-values themselves demonstrate how data finagling can yield spurious significance. This is true for all error probabilities. That's what a self-correcting inference account should do. Ben Goldacre, in *Bad Pharma* (2012), sums it up this way: the gambits give researchers an abundance of chances to find something when the tools assume you have had just one chance. Scouring different subgroups and otherwise "trying and trying again" are classic ways to blow up the actual probability of obtaining an impressive, but spurious, finding – and that remains so even if you ditch *P*-values and never compute them. FDA rules are designed to outlaw such gambits. To spot the cheating or questionable research practices (QRPs) responsible for a finding may not be easy. New research tools are being developed to detect them. Unsurprisingly, *P*-value analysis is relied on to discern spurious *P*-values (e.g., by lack of replication, or, in analyzing a group of tests, finding too many *P*-values in a given range). Ultimately, a qualitative severity scrutiny is necessary to get beyond merely raising doubts to falsifying purported findings.

Association Is Not Causation: Hormone Replacement Therapy (HRT)

Replicable results from high-quality research are sound, except for the sin that replicability fails to uncover: systematic bias.[2] Gaps between what is actually producing the statistical effect and what is inferred open the door by which biases creep in. Stand-in or proxy variables in statistical models may have little to do with the phenomenon of interest.

[2] This is the traditional use of "bias" as a systematic error. Ioannidis (2005) alludes to biasing as behaviors that result in a reported significance level differing from the value it actually has or ought to have (e.g., post-data endpoints, selective reporting). I will call those biasing selection effects.

So strong was the consensus-based medical judgment that hormone replacement therapy helps prevent heart disease that many doctors deemed it "unethical to ask women to accept the possibility that they might be randomized to a placebo" (The National Women's Health Network (NWHN) 2002, p. 180). Post-menopausal women who wanted to retain the attractions of being "Feminine Forever," as in the title of an influential tract (Wilson 1971), were routinely given HRT. Nevertheless, when a large randomized controlled trial (RCT) was finally done, it revealed statistically significant increased risks of heart disease, breast cancer, and other diseases that HRT was to have helped prevent. The observational studies on HRT, despite reproducibly showing a benefit, had little capacity to unearth biases due to "the healthy women's syndrome." There were confounding factors separately correlated with the beneficial outcomes enjoyed by women given HRT: they were healthier, better educated, and less obese than women not taking HRT. (That certain subgroups are now thought to benefit is a separate matter.)

Big Data scientists are discovering there may be something in the data collection that results in the bias being "hard-wired" into the data, and therefore even into successful replications. So replication is not enough. Beyond biased data, there's the worry that lab experiments may be only loosely connected to research claims. Experimental economics, for instance, is replete with replicable effects that economist Robert Sugden calls "exhibits." "An exhibit is an experimental design which reliably induces a surprising regularity" with at best an informal hypothesis as to its underlying cause (Sugden 2005, p. 291). Competing interpretations remain. (In our museum travels, "exhibit" will be used in the ordinary way.) In analyzing a test's capability to control erroneous interpretations, we must consider the porousness at multiple steps from data, to statistical inference, to substantive claims.

Souvenir A: Postcard to Send

The gift shop has a postcard listing the four slogans from the start of this Tour. Much of today's handwringing about statistical inference is unified by a call to block these fallacies. In some realms, trafficking in too-easy claims for evidence, if not criminal offenses, are "bad statistics"; in others, notably some social sciences, they are accepted cavalierly – much to the despair of panels on research integrity. We are more sophisticated than ever about the ways researchers can repress unwanted, and magnify wanted, results. Fraud-busting is everywhere, and the most important grain of truth is this: all the fraud-

busting is based on error statistical reasoning (if only on the meta-level). The minimal requirement to avoid BENT isn't met. It's hard to see how one can grant the criticisms while denying the critical logic.

We should oust mechanical, recipe-like uses of statistical methods that have long been lampooned, and are doubtless made easier by Big Data mining. They should be supplemented with tools to report magnitudes of effects that have and have not been warranted with severity. But simple significance tests have their uses, and shouldn't be ousted simply because some people are liable to violate Fisher's warning and report isolated results. They should be seen as a part of a conglomeration of error statistical tools for distinguishing genuine and spurious effects. They offer assets that are essential to our task: they have the means by which to register formally the fallacies in the postcard list. The failed statistical assumptions, the selection effects from trying and trying again, all alter a test's error-probing capacities. This sets off important alarm bells, and we want to hear them. Don't throw out the error-control baby with the bad statistics bathwater.

The slogans about lying with statistics? View them, not as a litany of embarrassments, but as announcing what any responsible method must register, if not control or avoid. Criticisms of statistical tests, where valid, boil down to problems with the critical alert function. Far from the high capacity to warn, "Curb your enthusiasm!" as correct uses of tests do, there are practices that make sending out spurious enthusiasm as easy as pie. This is a failure for sure, but don't trade them in for methods that cannot detect failure at all. If you're shopping for a statistical account, or appraising a statistical reform, your number one question should be: does it embody trigger warnings of spurious effects? Of bias? Of cherry picking and multiple tries? If the response is: "No problem; if you use our method, those practices require no change in statistical assessment!" all I can say is, if it sounds too good to be true, you might wish to hold off buying it.

We shouldn't be hamstrung by the limitations of any formal methodology. Background considerations, usually absent from typical frequentist expositions, must be made more explicit; taboos and conventions that encourage "mindless statistics" (Gigerenzer 2004) eradicated. The severity demand is what we naturally insist on as consumers. We want methods that are highly capable of finding flaws just when they're present, and we specify worst case scenarios. With the data in hand, we custom tailor our assessments depending on how severely (or inseverely) claims hold up. Here's an informal statement of the severity requirements (weak and strong):

> *Severity Requirement (weak):* If data $\boldsymbol{x}$ agree with a claim C but the method was practically incapable of finding flaws with C even if they exist, then $\boldsymbol{x}$ is poor evidence for C.
>
> *Severity (strong):* If C passes a test that was highly capable of finding flaws or discrepancies from C, and yet none or few are found, then the passing result, $\boldsymbol{x}$, is an indication of, or evidence for, C.

You might aver that we are too weak to fight off the lures of retaining the status quo – the carrots are too enticing, given that the sticks aren't usually too painful. I've heard some people say that evoking traditional mantras for promoting reliability, now that science has become so crooked, only makes things worse. Really? Yes there is gaming, but if we are not to become utter skeptics of good science, we should understand how the protections can work. In either case, I'd rather have rules to hold the "experts" accountable than live in a lawless wild west. I, for one, would be skeptical of entering clinical trials based on some of the methods now standard. There will always be cheaters, but give me an account that has eyes with which to spot them, and the means by which to hold cheaters accountable. That is, in brief, my basic statistical philosophy. The stakes couldn't be higher in today's world. Feynman said to take on an "extra type of integrity" that is not merely the avoidance of lying but striving "to check how you're maybe wrong." I couldn't agree more. But we laywomen are still going to have to proceed with a cattle prod.

1.3 The Current State of Play in Statistical Foundations: A View From a Hot-Air Balloon

> How can a discipline, central to science and to critical thinking, have two methodologies, two logics, two approaches that frequently give substantively different answers to the same problems? ... Is complacency in the face of contradiction acceptable for a central discipline of science? (Donald Fraser 2011, p. 329)
>
> We [statisticians] are not blameless ... we have not made a concerted professional effort to provide the scientific world with a unified testing methodology. (J. Berger 2003, p. 4)

From the aerial perspective of a hot-air balloon, we may see contemporary statistics as a place of happy multiplicity: the wealth of computational ability allows for the application of countless methods, with little handwringing about foundations. Doesn't this show we may have reached "the end of statistical foundations"? One might have thought so. Yet, descending close to a marshy wetland, and especially scratching a bit below the surface, reveals unease on all

https://doi.org/10.1017/9781107286184.002 Published online by Cambridge University Press

sides. The false dilemma between probabilism and long-run performance lets us get a handle on it. In fact, the Bayesian versus frequentist dispute arises as a dispute between probabilism and performance. This gets to my second reason for why the time is right to jump back into these debates: the "statistics wars" present new twists and turns. Rival tribes are more likely to live closer and in mixed neighborhoods since around the turn of the century. Yet, to the beginning student, it can appear as a jungle.

Statistics Debates: Bayesian versus Frequentist

> These days there is less distance between Bayesians and frequentists, especially with the rise of objective [default] Bayesianism, and we may even be heading toward a coalition government. (Efron 2013, p. 145)

A central way to formally capture probabilism is by means of the formula for conditional probability, where $\Pr(\boldsymbol{x}) > 0$:

$$\Pr(H|\boldsymbol{x}) = \frac{\Pr(H \text{ and } \boldsymbol{x})}{\Pr(\boldsymbol{x})}.$$

Since $\Pr(H \text{ and } \boldsymbol{x}) = \Pr(\boldsymbol{x}|H)\Pr(H)$ and $\Pr(\boldsymbol{x}) = \Pr(\boldsymbol{x}|H)\Pr(H) + \Pr(\boldsymbol{x}|{\sim}H)\Pr({\sim}H)$, we get:

$$\Pr(H|\boldsymbol{x}) = \frac{\Pr(\boldsymbol{x}|H)\Pr(H)}{\Pr(\boldsymbol{x}|H)\Pr(H) + \Pr(\boldsymbol{x}|{\sim}H)\Pr({\sim}H)},$$

where ~H is the denial of H. It would be cashed out in terms of all rivals to H within a frame of reference. Some call it Bayes' Rule or inverse probability. Leaving probability uninterpreted for now, if the data are very improbable given H, then our probability in H after seeing $\boldsymbol{x}$, the *posterior* probability $\Pr(H|\boldsymbol{x})$, may be lower than the probability in H prior to $\boldsymbol{x}$, the *prior* probability $\Pr(H)$. Bayes' Theorem is just a theorem stemming from the definition of conditional probability; it is only when statistical inference is thought to be encompassed by it that it becomes a statistical philosophy. Using Bayes' Theorem doesn't make you a Bayesian.

Larry Wasserman, a statistician and master of brevity, boils it down to a contrast of goals. According to him (2012b):

The Goal of Frequentist Inference: Construct procedure with frequentist guarantees [i.e., low error rates].

The Goal of Bayesian Inference: Quantify and manipulate your degrees of beliefs. In other words, Bayesian inference is the Analysis of Beliefs.

https://doi.org/10.1017/9781107286184.002 Published online by Cambridge University Press

At times he suggests we use B(H) for belief and F(H) for frequencies. The distinctions in goals are too crude, but they give a feel for what is often regarded as the Bayesian-frequentist controversy. However, they present us with the false dilemma (performance or probabilism) I've said we need to get beyond.

Today's Bayesian–frequentist debates clearly differ from those of some years ago. In fact, many of the same discussants, who only a decade ago were arguing for the irreconcilability of frequentist *P*-values and Bayesian measures, are now smoking the peace pipe, calling for ways to unify and marry the two. I want to show you what really drew me back into the Bayesian–frequentist debates sometime around 2000. If you lean over the edge of the gondola, you can hear some Bayesian family feuds starting around then or a bit after. Principles that had long been part of the Bayesian hard core are being questioned or even abandoned by members of the Bayesian family. Suddenly sparks are flying, mostly kept shrouded within Bayesian walls, but nothing can long be kept secret even there. Spontaneous combustion looms. Hard core subjectivists are accusing the increasingly popular "objective (non-subjective)" and "reference" Bayesians of practicing in bad faith; the new frequentist–Bayesian unificationists are taking pains to show they are not subjective; and some are calling the new Bayesian kids on the block "pseudo Bayesian." Then there are the Bayesians camping somewhere in the middle (or perhaps out in left field) who, though they still use the Bayesian umbrella, are flatly denying the very idea that Bayesian updating fits anything they actually do in statistics. Obeisance to Bayesian reasoning remains, but on some kind of a priori philosophical grounds. Let's start with the unifications.

While subjective Bayesianism offers an algorithm for coherently updating prior degrees of belief in possible hypotheses $H_1, H_2, \ldots, H_n$, these unifications fall under the umbrella of non-subjective Bayesian paradigms. Here the prior probabilities in hypotheses are not taken to express degrees of belief but are given by various formal assignments, ideally to have minimal impact on the posterior probability. I will call such Bayesian priors *default*. Advocates of unifications are keen to show that (i) default Bayesian methods have good performance in a long series of repetitions – so probabilism may yield performance; or alternatively, (ii) frequentist quantities are similar to Bayesian ones (at least in certain cases) – so performance may yield probabilist numbers. Why is this not bliss? Why are so many from all sides dissatisfied?

True blue subjective Bayesians are understandably unhappy with non-subjective priors. Rather than quantify prior beliefs, non-subjective priors are viewed as primitives or conventions for obtaining posterior probabilities. Take Jay Kadane (2008):

The growth in use and popularity of Bayesian methods has stunned many of us who were involved in exploring their implications decades ago. The result . . . is that there are users of these methods who do not understand the *philosophical basis of the methods they are using*, and hence may misinterpret or badly use the results . . . No doubt helping people to use Bayesian methods more appropriately is an important task of our time. (p. 457, emphasis added)

I have some sympathy here: Many modern Bayesians aren't aware of the traditional philosophy behind the methods they're buying into. Yet there is not just one philosophical basis for a given set of methods. This takes us to one of the most dramatic shifts in contemporary statistical foundations. It had long been assumed that only subjective or personalistic Bayesianism had a shot at providing genuine philosophical foundations, but you'll notice that groups holding this position, while they still dot the landscape in 2018, have been gradually shrinking. Some Bayesians have come to question whether the widespread use of methods under the Bayesian umbrella, however useful, indicates support for subjective Bayesianism as a foundation.

Marriages of Convenience?

The current frequentist–Bayesian unifications are often marriages of convenience; statisticians rationalize them less on philosophical than on practical grounds. For one thing, some are concerned that methodological conflicts are bad for the profession. For another, frequentist tribes, contrary to expectation, have not disappeared. Ensuring that accounts can control their error probabilities remains a desideratum that scientists are unwilling to forgo. Frequentists have an incentive to marry as well. Lacking a suitable epistemic interpretation of error probabilities – significance levels, power, and confidence levels – frequentists are constantly put on the defensive. Jim Berger (2003) proposes a construal of significance tests on which the tribes of Fisher, Jeffreys, and Neyman could agree, yet none of the chiefs of those tribes concur (Mayo 2003b). The success stories are based on agreements on numbers that are not obviously true to any of the three philosophies. Beneath the surface – while it's not often said in polite company – the most serious disputes live on. I plan to lay them bare.

If it's assumed an evidential assessment of hypothesis H should take the form of a posterior probability of H – a form of probabilism – then P-values and confidence levels are applicable only through misinterpretation and mistranslation. Resigned to live with P-values, some are keen to show that construing them as posterior probabilities is not so bad (e.g., Greenland and Poole 2013). Others focus on long-run error control, but cede territory

wherein probability captures the epistemological ground of statistical inference. Why assume significance levels and confidence levels lack an authentic epistemological function? I say they do: to secure and evaluate how well probed and how severely tested claims are.

Eclecticism and Ecumenism

If you look carefully between dense forest trees, you can distinguish unification country from lands of eclecticism (Cox 1978) and ecumenism (Box 1983), where tools first constructed by rival tribes are separate, and more or less equal (for different aims). Current-day eclecticisms have a long history – the dabbling in tools from competing statistical tribes has not been thought to pose serious challenges. For example, frequentist methods have long been employed to check or calibrate Bayesian methods (e.g., Box 1983); you might test your statistical model using a simple significance test, say, and then proceed to Bayesian updating. Others suggest scrutinizing a posterior probability or a likelihood ratio from an error probability standpoint. What this boils down to will depend on the notion of probability used. If a procedure frequently gives high probability for *claim C* even if *C* is false, severe testers deny convincing evidence has been provided, and never mind about the meaning of probability.

One argument is that throwing different methods at a problem is all to the good, that it increases the chances that at least one will get it right. This may be so, provided one understands how to interpret competing answers. Using multiple methods is valuable when a shortcoming of one is rescued by a strength in another. For example, when randomized studies are used to expose the failure to replicate observational studies, there is a presumption that the former is capable of discerning problems with the latter. But what happens if one procedure fosters a goal that is not recognized or is even opposed by another? Members of rival tribes are free to sneak ammunition from a rival's arsenal – but what if at the same time they denounce the rival method as useless or ineffective?

Decoupling. On the horizon is the idea that statistical methods may be decoupled from the philosophies in which they are traditionally couched. In an attempted meeting of the minds (Bayesian and error statistical), Andrew Gelman and Cosma Shalizi (2013) claim that "implicit in the best Bayesian practice is a stance that has much in common with the error-statistical approach of Mayo" (p. 10). In particular, Bayesian model checking, they say, uses statistics to satisfy Popperian criteria for *severe tests*. The idea of error statistical foundations for Bayesian tools is not as preposterous as it may seem. The concept of severe testing is sufficiently general to apply to any of the methods now in use.

On the face of it, any inference, whether to the adequacy of a model or to a posterior probability, can be said to be warranted just to the extent that it has withstood severe testing. Where this will land us is still futuristic.

Why Our Journey?

> We have all, or nearly all, moved past these old [Bayesian-frequentist] debates, yet our textbook explanations have not caught up with the eclecticism of statistical practice. (Kass 2011, p. 1)

When Kass proffers "a philosophy that matches contemporary attitudes," he finds resistance to his big tent. Being hesitant to reopen wounds from old battles does not heal them. Distilling them in inoffensive terms just leads to the marshy swamp. Textbooks can't "catch-up" by soft-peddling competing statistical accounts. They show up in the current problems of scientific integrity, irreproducibility, questionable research practices, and in the swirl of methodological reforms and guidelines that spin their way down from journals and reports.

From an elevated altitude we see how it occurs. Once high-profile failures of replication spread to biomedicine, and other "hard" sciences, the problem took on a new seriousness. Where does the new scrutiny look? By and large, it collects from the earlier social science "significance test controversy" and the traditional philosophies coupled to Bayesian and frequentist accounts, along with the newer Bayesian–frequentist unifications we just surveyed. This jungle has never been disentangled. No wonder leading reforms and semi-popular guidebooks contain misleading views about all these tools. No wonder we see the same fallacies that earlier reforms were designed to avoid, and even brand new ones. Let me be clear, I'm not speaking about flat-out howlers such as interpreting a *P*-value as a posterior probability. By and large, they are more subtle; you'll want to reach your own position on them. It's not a matter of switching your tribe, but excavating the roots of tribal warfare. To tell what's true about them. I don't mean understand them at the socio-psychological levels, although there's a good story there (and I'll leak some of the juicy parts during our travels).

How can we make progress when it is difficult even to tell what is true about the different methods of statistics? We must start afresh, taking responsibility to offer a new standpoint from which to interpret the cluster of tools around which there has been so much controversy. Only then can we alter and extend their limits. I admit that the statistical philosophy that girds our explorations is not out there ready-made; if it was, there would be no need for our holiday cruise. While there are plenty of giant shoulders on which we stand, we won't

be restricted by the pronouncements of any of the high and low priests, as sagacious as many of their words have been. In fact, we'll brazenly question some of their most entrenched mantras. Grab on to the gondola, our balloon's about to land.

In Tour II, I'll give you a glimpse of the core behind statistics battles, with a firm promise to retrace the steps more slowly in later trips.