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 $\boldsymbol{x}$ 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

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.

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

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

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.

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

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.

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.

Tour II Error Probing Tools versus Logics of Evidence

1.4 The Law of Likelihood and Error Statistics

If you want to understand what's true about statistical inference, you should begin with what has long been a holy grail – to use probability to arrive at a type of logic of evidential support – and in the first instance you should look not at full-blown Bayesian probabilism, but at comparative accounts that sidestep prior probabilities in hypotheses. An intuitively plausible logic of comparative support was given by the philosopher Ian Hacking (1965) – the Law of Likelihood. Fortunately, the Museum of Statistics is organized by theme, and the Law of Likelihood and the related Likelihood Principle is a big one.

> *Law of Likelihood (LL):* Data $\boldsymbol{x}$ are better evidence for hypothesis H_1 than for H_0 if $\boldsymbol{x}$ is more probable under H_1 than under H_0: $\Pr(\boldsymbol{x}; H_1) > \Pr(\boldsymbol{x}; H_0)$, that is, the *likelihood ratio (LR)* of H_1 over H_0 exceeds 1.

H_0 and H_1 are statistical hypotheses that assign probabilities to values of the random variable $\boldsymbol{X}$. A fixed value of $\boldsymbol{X}$ is written $\boldsymbol{x}_0$, but we often want to generalize about this value, in which case, following others, I use $\boldsymbol{x}$. The *likelihood of the hypothesis H*, given data $\boldsymbol{x}$, is the probability of observing $\boldsymbol{x}$, under the assumption that H is true or adequate in some sense. Typically, the ratio of the likelihood of H_1 over H_0 also supplies the quantitative measure of comparative support. Note, when $\boldsymbol{X}$ is continuous, the probability is assigned over a small interval around $\boldsymbol{X}$, to avoid probability 0.

Does the Law of Likelihood Obey the Minimal Requirement for Severity?

Likelihoods are vital to all statistical accounts, but they are often misunderstood because the data are fixed and the hypothesis varies. Likelihoods of hypotheses should not be confused with their probabilities. Two ways to see this. First, suppose you discover all of the stocks in Pickrite's promotional letter went up in value ($\boldsymbol{x}$) – all winners. A hypothesis H to explain this is that their method always succeeds in picking winners. H *entails* $\boldsymbol{x}$, so the likelihood of H given $\boldsymbol{x}$ is 1. Yet we wouldn't say H is therefore highly probable, especially without reason to put to rest that they culled the winners post hoc. For a second

way, at any time, the same phenomenon may be perfectly predicted or explained by two rival theories; so both theories are equally likely on the data, even though they cannot both be true.

Suppose Bristol-Roach, in our Bernoulli tea tasting example, got two correct guesses followed by one failure. The observed data can be represented as $\boldsymbol{x}_0 = \langle 1,1,0 \rangle$. Let the hypotheses be different values for θ, the probability of success on each independent trial. The likelihood of the hypothesis $H_0 : \theta = 0.5$, given $\boldsymbol{x}_0$, which we may write as Lik(0.5), equals (1/2)(1/2)(1/2) = 1/8. Strictly speaking, we should write Lik(θ;$\boldsymbol{x}_0$), because it's always computed given data $\boldsymbol{x}_0$; I will do so later on. The likelihood of the hypothesis $\theta = 0.2$ is Lik(0.2) = (0.2)(0.2)(0.8) = 0.032. In general, the likelihood in the case of Bernoulli independent and identically distributed trials takes the form: $\text{Lik}(\theta) = \theta^s(1-\theta)^f$, $0 < \theta < 1$, where s is the number of successes and f the number of failures. Infinitely many values for θ between 0 and 1 yield positive likelihoods; clearly then, likelihoods do not sum to 1, or any number in particular. Likelihoods do not obey the probability calculus.

The Law of Likelihood (LL) will immediately be seen to fail our minimal severity requirement – at least if it is taken as an account of inference. Why? There is no onus on the Likelihoodist to predesignate the rival hypotheses – you are free to search, hunt, and post-designate a more likely, or even maximally likely, rival to a test hypothesis H_0.

Consider the hypothesis that $\theta = 1$ on trials one and two and 0 on trial three. That makes the probability of $\boldsymbol{x}$ maximal. For another example, hypothesize that the observed pattern would always recur in three-trials of the experiment (I. J. Good said in his cryptoanalysis work these were called "kinkera"). Hunting for an impressive fit, or trying and trying again, one is sure to find a rival hypothesis H_1 much better "supported" than H_0 even when H_0 is true. As George Barnard puts it, "there *always* is such a rival hypothesis, viz. that things just had to turn out the way they actually did" (1972, p. 129).

Note that for any outcome of n Bernoulli trials, the likelihood of $H_0 : \theta = 0.5$ is $(0.5)^n$, so is quite small. The likelihood ratio (LR) of a best-supported alternative compared to H_0 would be quite high. Since one could always erect such an alternative,

(*) Pr(LR in favor of H_1 over H_0; H_0) = maximal.

Thus the LL permits BENT evidence. The severity for H_1 is minimal, though the particular H_1 is not formulated until the data are in hand. I call such maximally fitting, but minimally severely tested, hypotheses *Gellerized*, since Uri Geller was apt to erect a way to explain his results in ESP trials. Our Texas sharpshooter is analogous because he can always draw a circle around a cluster of bullet holes, or around each single hole. One needn't go to such an extreme

rival, but it suffices to show that the LL does not control the probability of erroneous interpretations.

What do we do to compute (*)? We look beyond the specific observed data to the behavior of the general rule or method, here the LL. The output is always a comparison of likelihoods. We observe one outcome, but we can consider that for any outcome, unless it makes H_0 maximally likely, we can find an H_1 that is more likely. This lets us compute the relevant properties of the method: its inability to block erroneous interpretations of data. As always, a severity assessment is one level removed: you give me the rule, and I consider its latitude for erroneous outputs. We're actually looking at the probability distribution of the rule, over outcomes in the sample space. This distribution is called a *sampling distribution*. It's not a very apt term, but nothing has arisen to replace it. For those who embrace the LL, once the data are given, it's irrelevant what other outcomes could have been observed but were not. Likelihoodists say that such considerations make sense only if the concern is the performance of a rule over repetitions, but not for inference from the data. Likelihoodists hold to "the irrelevance of the sample space" (once the data are given). This is the key contrast between accounts based on error probabilities (error statistical accounts) and logics of statistical inference.

Hacking "There is No Such Thing as a Logic of Statistical Inference"

Hacking's (1965) book was so ahead of its time that by the time philosophers of science started to get serious about philosophy of statistics, he had already broken the law he had earlier advanced. Hacking (1972, 1980) admits to having been caught up in the "logicist" mindset wherein we assume a logical relationship exists between any data and hypothesis; and even denies (1980, p. 145) there is any such thing.

In his review of A. F. Edwards' (1972) book *Likelihood*, Hacking (1972) gives his main reasons for rejecting the LL:

> We capture enemy tanks at random and note the serial numbers on their engines. We know the serial numbers start at 0001. We capture a tank number 2176. How many did the enemy make? On the likelihood analysis, the best-supported guess is: 2176. Now one can defend this remarkable result by saying that it does not follow that we should estimate the actual number as 2176 only that comparing individual numbers, 2176 is better supported than any larger figure. My worry is deeper. Let us compare the relative likelihood of the two hypotheses, 2176 and 3000. Now pass to a situation where we are measuring, say, widths of a grating in which error has a normal distribution with known variance; we can devise data and a pair of hypotheses about the mean which will have the same log-likelihood ratio. I have no inclination to say that the relative support in the

tank case is 'exactly the same as' that in the normal distribution case, even though the likelihood ratios are the same. (pp. 136–7)

Likelihoodists will insist that the law may be upheld by appropriately invoking background information, and by drawing distinctions between evidence, belief, and action.

Royall's Road to Statistical Evidence

Statistician Richard Royall, a longtime leader of Likelihoodist tribes, has had a deep impact on current statistical foundations. His views are directly tied to recent statistical reforms – even if those reformers go Bayesian rather than stopping, like Royall, with comparative likelihoods. He provides what many consider a neat proposal for settling disagreements about statistical philosophy. He distinguishes three questions: belief, action, and evidence:

1. What do I believe, now that I have this observation?
2. What should I do, now that I have this observation?
3. How should I interpret this observation as evidence regarding $[H_0]$ versus $[H_1]$? (Royall 1997, p. 4)

Can we line up these three goals to my probabilism, performance, and probativeness (Section 1.2)? No. Probativeness gets no pigeonhole. According to Royall, what to believe is captured by Bayesian posteriors, how to act is captured by a frequentist performance (in some cases he will add costs). What's his answer to the evidence question? The Law of Likelihood.

Let's use one of Royall's first examples, appealing to Bernoulli distributions again – independent, dichotomous trials, "success" or "failure":

> Medical researchers are interested in the success probability, θ, associated with a new treatment. They are particularly interested in how θ relates to the old treatment's success probability, believed to be about 0.2. They have reason to hope that θ is considerably greater, perhaps 0.8 or even greater. (Royall 1997, p. 19)

There is a set of possible outcomes, a sample space, S, and a set of possible parameter values, a parameter space Ω. He considers two hypotheses:

$$\theta = 0.2 \text{ and } \theta = 0.8.$$

These are *simple* or *point* hypotheses. To illustrate take a miniature example with only $n = 4$ trials where each can be a "success" $\{X = 1\}$ or a "failure" $\{X = 0\}$. A possible result might be $\boldsymbol{x}_0 = \langle 1,1,0,1 \rangle$. Since $\Pr(X = 1) = \theta$ and $\Pr(X = 0) = (1 - \theta)$, the probability of $\boldsymbol{x}_0$ is $(\theta)(\theta)(1 - \theta)(\theta)$. Given independent trials, they multiply. Under the two hypotheses, given $\langle 1,1,0,1 \rangle$, the likelihoods are

$$\text{Lik}(H_0) = (0.2)(0.2)(0.8)(0.2) = 0.0064,$$

$$\text{Lik}(H_1) = (0.8)(0.8)(0.2)(0.8) = 0.1024.$$

A hypothesis that would make the data most probable would be that $\theta = 1$, on the three trials that yield successes, and 0 where it yields failure.

We typically denigrate "just so" stories, purposely erected to fit the data, as "unlikely." Yet they are *most* likely in the technical sense! So in hearing likelihood used formally, you must continually keep this swap of meanings in mind. (We call them Gellerized only if they pass with minimal severity.) If θ is to be constant on each trial, as in the Bernoulli model, the maximum likely hypothesis equates θ with the relative frequency of success, 0.75. [Exercise for reader: find Lik (0.75)]

Exhibit (i): Law of Likelihood Compared to a Significance Test. Here Royall contrasts his handling of the medical example to the standard significance test:

> A standard statistical analysis of their observations would use a *Bernoulli*(θ) statistical model and test the composite hypotheses H_0: $\theta \leq 0.2$ versus H_1: $\theta > 0.2$. That analysis would show that H_0 can be rejected in favor of H_1 at any significance level greater than 0.003, a result that is conventionally taken to mean that the observations are very strong evidence supporting H_1 over H_0. (Royall 1997, p. 19; substituting H_0 and H_1 for H_1 and H_2.)

So the significance tester looks at the composite hypotheses H_0: $\theta \leq 0.2$ vs. H_1: $\theta > 0.2$, rather than his point hypotheses $\theta = 0.2$ and $\theta = 0.8$. Here, she would look at how much larger the mean success rate is in the sample $(X_1 + X_2 + \ldots X_{17})/17$, which we abbreviate as $\overline{x} = 9/17 = 0.53$, compared to what is expected under H_0, put in standard deviation units. Using Royall's numbers, the observed success rate is

$$\overline{x} = 9/17 = .53;$$

$$\sigma = \sqrt{[\theta\,(1-\theta)]},\ \text{which, under the null, is } \sqrt{[0.2\ (0.8)]} = 0.4.$$

The *test statistic* $\text{d}(\boldsymbol{X})$ is $\sqrt{17}(\overline{X} - 0.2)/\sigma$; it gets larger and larger the more the data deviate from what is expected under H_0 – as is sensible for a good test statistic. Its value is

$$\text{d}(\boldsymbol{x}_0) = \sqrt{17}\ (0.53 - 0.2)/\ 0.4 \simeq 3.3.$$

The significance level associated with $\text{d}(\boldsymbol{x}_0)$ is

$$\Pr(\text{d}(\boldsymbol{X}) \geq \text{d}(\boldsymbol{x}); H_0) \simeq 0.003.$$

This is read, "the probability d($\boldsymbol{X}$) would be at least as large as the particular value d($\boldsymbol{x}_0$), under the supposition that H_0 adequately describes the data generation procedure" (see Souvenir C). It's not strictly a conditional probability – a subtle point that won't detain us here. We continue to follow Royall's treatment, though we'd want to distinguish the mere *indication* of an isolated significant result from strong *evidence*. We'd also have to audit for model assumptions and selection effects, but we assume these check out; after all, Royall's likelihood account also depends on the model holding.

We'd argue along the following lines: were H_0 a reasonable description of the process, then with very high probability you would not be able to regularly produce d($\boldsymbol{x}$) values as large as this:

$$\Pr(\mathrm{d}(\boldsymbol{X}) < \mathrm{d}(\boldsymbol{x}); H_0) \simeq 0.997.$$

So if you manage to get such a large difference, I may infer that $\boldsymbol{x}$ *indicates* a genuine effect. Let's go back to Royall's contrast, because he's very unhappy with this.

Why Does the LL Reject Composite Hypotheses?

Royall tells us that his account is unable to handle composite hypotheses, even this one (for which there is a uniformly most powerful [UMP] test over all points in H_0). He does not conclude that his test comes up short. He and other Likelihoodists maintain that any genuine test or "rule of rejection" should be restricted to comparing the likelihood of H versus some point alternative H' *relative to* fixed data $\boldsymbol{x}$ (Royall 1997, pp. 19–20). It is a virtue. No wonder the Likelihoodist disagrees with the significance tester. In their view, a simple significance test is not a "real" testing account because it is not a comparative appraisal. Elliott Sober, a well-known philosopher of science, echoes Royall: "The fact that significance tests don't contrast the null with alternatives suffices to show that they do not provide a good rule for rejection" (Sober 2008, p. 56). Now, Royall's significance test *has* an alternative H_1: $\theta > 0.2$! It's just not a point alternative but is compound or composite (including all values greater than 0.2). The form of inference, admittedly, is not of the comparative ("evidence favoring") variety. In this discussion, H_0 and H_1 replace his H_1 and H_2.

What untoward consequences occur if we consider composite hypotheses (according to the Likelihoodist)? The problem is that even though the likelihood of $\theta = 0.2$ is small, there are values within alternative H_1: $\theta > 0.2$ that are even less likely on the data $\overline{x} = 0.53$. For instance consider $\theta = 0.9$.

> [B]ecause H_0 contains some simple hypotheses that are better supported than some hypotheses in H_1 (e.g., $\theta = 0.2$ is better supported than $\theta = 0.9$ by a likelihood ratio of

LR = $(0.2/0.9)^9(0.8/0.1)^8 = 22.2$), the law of likelihood does not allow the characterization of these observations as strong evidence for H_1 over H_0. (Royall 1997, p. 20)

For Royall, rejecting H_0: $\theta \leq 0.2$ and inferring H_1: $\theta > 0.2$ is to assert *every* parameter point within H_1 is more likely than every point in H_0. That seems an idiosyncratic meaning to attach to "infer evidence of $\theta > 0.2$"; but it explains this particular battle. It still doesn't explain the alleged problem for the significance tester who just takes it to mean what it says:

To reject H_0: $\theta \leq 0.2$ is to infer *some* positive discrepancy from 0.2.

We readily agree with Royall that there's a problem with taking a rejection of H_0: $\theta \leq 0.2$, with $\overline{x} = 0.53$, as evidence of a discrepancy as large as $\theta = 0.9$. It's terrible evidence even that θ is as large as 0.7 or 0.8. Here's how a tester articulates this terrible evidence.

Consider the test rule: infer evidence of a discrepancy from 0.2 as large as 0.9, based on observing $\overline{x} = 0.53$. The data differ from 0.2 in the direction of H_1, but to take that difference as indicating an underlying $\theta > 0.9$ would be wrong with probability ~1. Since the standard error of the mean, $\sigma_{\overline{x}}$, is 0.1, alternative 0.9 is more than $3\sigma_{\overline{x}}$ greater than 0.53. ($\sigma_{\overline{x}} = \sigma/\sqrt{n}$) The inference gets low severity.

We'll be touring significance tests and confidence bounds in detail later. We're trying now to extract some core contrasts between error statistical methods and logics of evidence such as the LL. According to the LL, so long as there is a point within H_1 that is less likely given $\boldsymbol{x}$ than is H_0, the data are "evidence *in favor* of the null hypothesis, not evidence *against* it" (Sober 2008, pp. 55–6). He should add "as compared to" some less likely alternative. We never infer a statistical hypothesis according to the LL, but rather a likelihood ratio of two hypotheses, neither of which might be likely. The significance tester and the comparativist hold very different images of statistical inference.

Can an account restricted to comparisons answer the questions: is $\boldsymbol{x}$ good evidence for H? Or is it a case of bad evidence, no test? Royall says no. He declares that all attempts to say whether $\boldsymbol{x}$ is good evidence for H, or even if $\boldsymbol{x}$ is better evidence for H than is $\boldsymbol{y}$, are utterly futile. Similarly, "What *does* the [LL] say when one hypothesis attaches the same probability to two different observations? It says absolutely nothing . . . [it] applies when two different hypotheses attach probabilities to the same observation" (Royall 2004, p. 148). That cuts short important tasks of inferential scrutiny. Since model checking concerns the adequacy of a single model, the Likelihoodist either forgoes such checks or must go beyond the paradigm.

Still, if the model can be taken as adequate, and the Likelihoodist gives a sufficiently long list of comparisons, the differences between us don't seem so marked. Take Royall:

> One statement that we can make is that the observations are only weak evidence in favor of $\theta = 0.8$ versus $\theta = 0.2$ (LR = 4) . . . and at least moderately strong evidence for $\theta = 0.5$ over any value $\theta > 0.8$ (LR) > 22). (1977, p. 20)

Nonetheless, we'd want to ask: what do these numbers mean? Is 22 a lot? Is 4 small? We're back to Hacking's attempt to compare tank cars with widths of a grating. How do we calibrate them? Neyman and Pearson's answer, we'll see, is to look at the probability of so large a likelihood ratio, under various hypotheses, as in (*).

LRs and Posteriors. Royall is loath to add prior probabilities to the assessment of the import of the evidence. This, he says, allows the LR to be "a precise and objective numerical measure of the strength of evidence" in comparing hypotheses (2004, p. 123). At the same time, Royall argues, the LL "constitutes the essential core of the Bayesian account of evidence . . . the Bayesian who rejects the [LL] undermines his own position" (ibid., p. 146). The LR, after all, is the factor by which the ratio of posterior probabilities is changed by the data. Consider just two hypotheses, switching from the ";" in the significance test to conditional probability "|":[1]

$$\Pr(H_0|\boldsymbol{x}) = \frac{\Pr(\boldsymbol{x}|H_0)\,\Pr(H_0)}{\Pr(\boldsymbol{x}|H_0)\,\Pr(H_0) \;+\; \Pr(\boldsymbol{x}|H_1)\,\Pr(H_1)}.$$

Likewise:

$$\Pr(H_1|\boldsymbol{x}) = \frac{\Pr(\boldsymbol{x}|H_1)\,\Pr(H_1)}{\Pr(\boldsymbol{x}|H_1)\,\Pr(H_1) \;+\; \Pr(\boldsymbol{x}|H_0)\,\Pr(H_0)}.$$

The denominators equal $\Pr(\boldsymbol{x})$, so they cancel in the LR:

$$\frac{\Pr(H_1|\boldsymbol{x})}{\Pr(H_0|\boldsymbol{x})} = \frac{\Pr(\boldsymbol{x}|\,H_1)\Pr(H_1)}{\Pr(\boldsymbol{x}|\,H_0)\Pr(H_0)}.$$

All of this assumes the likelihoods and the model are deemed adequate.

[1] Divide the numerator and the denominator by $\Pr(\boldsymbol{x}|H_0)\Pr(H_0)$. Then

$$\Pr(H_0|\boldsymbol{x}) = \frac{1}{1 + \frac{\Pr(\boldsymbol{x}|H_1)\Pr(H_1)}{\Pr(\boldsymbol{x}|H_0)\Pr(H_0)}}$$

Data Dredging: Royall Bites the Bullet

Return now to our most serious problem: The Law of Likelihood permits finding evidence in favor of a hypothesis deliberately arrived at using the data, even in the extreme case that it is Gellerized. Allan Birnbaum, who had started out as a Likelihoodist, concludes, "the likelihood concept cannot be construed so as to allow useful appraisal, and thereby possible control, of probabilities of erroneous interpretations" (Birnbaum 1969, p. 128). But Royall has a clever response. Royall thinks control of error probabilities arises only in answering his second question about action, not evidence. He is prepared to bite the bullet. He himself gives the example of a "trick deck." You've shuffled a deck of ordinary-looking playing cards; you turn over the top card and find an ace of diamonds:

> According to the law of likelihood, the hypothesis that the deck consists of 52 aces of diamonds (H_1) is better supported than the hypothesis that the deck is normal (H_N) [by the factor 52] . . . Some find this disturbing. (Royall 1997, pp. 13–14)

Royall does not. He admits:

> . . . it seems unfair; no matter what card is drawn, the law implies that the corresponding trick-deck hypothesis (52 cards just like the one drawn) is better supported than the normal-deck hypothesis. Thus even if the deck is normal we will always claim to have found strong evidence that it is not. (ibid.)

What he is admitting then is, given any card:

$$\Pr(\text{LR favors trick deck hypothesis; normal deck}) = 1.$$

Even though different trick deck hypotheses would be formed for different outcomes, we may compute the sampling distribution (*). The severity for "trick deck" would be 0. It need not be this extreme to have BENT results, but you get the idea.

What's Royall's way out? At the level of a report on comparative likelihoods, Royall argues, there's no need for a way out. To Royall, it only shows a confusion between evidence and belief.[2] If you're not convinced the deck has 52 aces of diamonds rather than being a normal deck "it does not mean that the observation is not strong evidence in favor of H_1 versus H_N" where H_N is a normal deck (ibid., p. 14). It just wasn't strong enough to overcome your prior beliefs. If you regard the maximally likely alternative as unpalatable, you should have given it a suitably low prior degree of probability. The more likely hypothesis is still favored on grounds of evidence, but your posterior belief

[2] He notes that the comparative evidence for a trick versus a normal deck is not evidence against a normal deck alone (pp. 14–15).

may be low. Don't confuse evidence with belief! For the question of evidence, your beliefs have nothing to do with it, according to Royall's Likelihoodist.

What if we grant the Likelihoodist this position? What do we do to tackle the essential challenge to the credibility of statistical inference today, when it's all about Texas Marksmen, hunters, snoopers, and cherry pickers? These moves, which play havoc with a test's ability to control erroneous interpretations, do not alter the evidence at all, say Likelihoodists. The fairest reading of Royall's position might be this: the data indicate only the various LRs. If they are the same, it matters not whether hypotheses arose through data dredging – at least, so long as you are in the category of "what the data say." As soon as you're troubled, you slip into the category of belief. What if we're troubled by the ease of exaggerating findings when you're allowed to rummage around? What if we wish to clobber the Texas sharpshooter method, never mind my beliefs in the particular claims they infer. You might aver, we should never be considering trick deck hypotheses, but this is the example Royall gives, and he is a, if not the, leading Likelihoodist.

To him, appealing to error probabilities is relevant only pre-data, which wouldn't trouble the severe tester so much if Likelihoodists didn't regard them as relevant only for a performance goal, not inference. Given that frequentists have silently assented to the performance use of error probabilities, it's perhaps not surprising that others accept this. The problem with cherry picking is not about long runs, it's that a poor job has been done in the case at hand. The severity requirement reflects this intuition. By contrast, Likelihoodists hold that likelihood ratios, and unadjusted *P*-values, still convey what the data say, even with claims arrived at through data dredging. It's true you can explore, arrive at *H*, then test *H* on other data; but isn't the reason there's a need to test on new data that your assessment will otherwise fail to convey how well tested *H* is?

Downsides to the "Appeal to Beliefs" Solution to Inseverity

What's wrong with Royall's appeal to prior beliefs to withhold support to a "just so" hypothesis? It may get you out of a jam in some cases. Here's why the severe tester objects. First, she insists on distinguishing the *evidential* warrant for one and the same hypothesis *H* in two cases: one where it was constructed post hoc, cherry picked, and so on, a second where it was predesignated. A cherry-picked hypothesis *H* could well be believable, but we'd still want to distinguish the evidential credit *H* deserves in the two cases. Appealing to priors can't help, since here there's one and the same *H*.

Perhaps someone wants to argue that the mode of testing alters the degree of belief in *H*, but this would be non-standard (violating the Likelihood Principle to be discussed shortly). Philosopher Roger Rosenkrantz puts it thus: The LL entails the irrelevance "of whether the theory was formulated in advance or suggested by the observations themselves" (Rosenkrantz 1977, p. 121). For Rosenkrantz, a default Bayesian last I checked, this irrelevance of predesignation is altogether proper. By contrast, he admits, "Orthodox (non-Bayesian) statisticians have found this to be strong medicine indeed!" (ibid.). Many might say instead that it is bad medicine. Take, for instance, something called the CONSORT, the Consolidated Standards of Reporting Trials from RCTs in medicine:

> Selective reporting of outcomes is widely regarded as misleading. It undermines the validity of findings, particularly when driven by statistical significance or the direction of the effect [4], and has memorably been described in the New England Journal of Medicine as "Data Torturing" [5]. (COMpare Team 2015)

This gets to a second problem with relying on beliefs to block data-dredged hypotheses. Post-data explanations, even if it took a bit of data torture, are often incredibly convincing, and you don't have to be a sleaze to really believe them. Goldacre (2016) expresses shock that medical journals continue to report outcomes that were altered post-data – he calls this *outcome-switching*. Worse, he finds, some journals defend the practice because they are convinced that their very good judgment entitles them to determine when to treat post-designated hypotheses as if they were predesignated. Unlike the LL, the CONSORT and many other best practice guides view these concerns as an essential part of reporting what the data say. Now you might say this is just semantics, as long as, in the end, they report that outcome-switching occurred. Maybe so, provided the report mentions why it would be misleading to hide the information. At least people have stopped referring to frequentist statistics as "Orthodox."

There is a third reason to be unhappy with supposing the only way to block evidence for "just so" stories is by the *deus ex machina* of a low prior degree of belief: it misidentifies what the problem really *is*. The influence of the biased selection is not on the believability of *H* but rather on the capability of the test to have unearthed errors. The error probing capability of the testing procedure is being diminished. If you engage in cherry picking, you are not "sincerely trying," as Popper puts it, to find flaws with claims, but instead you are finding evidence in favor of a well-fitting hypothesis that you deliberately construct – barred only if your intuitions say it's unbelievable. The job that was supposed to be accomplished by an account of statistics now has to be performed by *you*. Yet you are the one most likely to follow your preconceived opinions, biases, and pet

theories. If an account of statistical inference or evidence doesn't supply self-critical tools, it comes up short in an *essential* way. So says the severe tester.

Souvenir B: Likelihood versus Error Statistical

Like pamphlets from competing political parties, the gift shop from this tour proffers pamphlets from these two perspectives.

To the Likelihoodist, points in favor of the LL are:

- The LR offers "a precise and objective numerical measure of the strength of statistical evidence" for one hypotheses over another; it is a frequentist account and does not use prior probabilities (Royall 2004, p. 123).
- The LR is fundamentally related to Bayesian inference: the LR is the factor by which the ratio of posterior probabilities is changed by the data.
- A Likelihoodist account does not consider outcomes other than the one observed, unlike *P*-values, and Type I and II errors. (Irrelevance of the sample space.)
- Fishing for maximally fitting hypotheses and other gambits that alter error probabilities do not affect the assessment of evidence; they may be blocked by moving to the "belief" category.

To the error statistician, problems with the LL include:

- LRs do not convey the same evidential appraisal in different contexts.
- The LL denies it makes sense to speak of how well or poorly tested a single hypothesis is on evidence, essential for model checking; it is inapplicable to composite hypothesis tests.
- A Likelihoodist account does not consider outcomes other than the one observed, unlike *P*-values, and Type I and II errors. (Irrelevance of the sample space.)
- Fishing for maximally fitting hypotheses and other gambits that alter error probabilities do not affect the assessment of evidence; they may be blocked by moving to the "belief" category.

Notice, the last two points are identical for both. What's a selling point for a Likelihoodist is a problem for an error statistician.

1.5 Trying and Trying Again: The Likelihood Principle

> The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing when data collection stops are irrelevant to data interpretation. (Edwards, Lindman, and Savage 1963, p. 193)

Several well-known gambits make it altogether easy to find evidence in support of favored claims, even when they are unwarranted. A responsible statistical inference report requires information about whether the method used is capable of controlling such erroneous interpretations of data or not. Now we see that adopting a statistical inference account is also to buy into principles for processing data, hence criteria for "what the data say," hence grounds for charging an inference as illegitimate, questionable, or even outright cheating. The best way to survey the landscape of statistical debates is to hone in on some pivotal points of controversy – saving caveats and nuances for later on.

Consider for example the gambit of "trying and trying again" to achieve statistical significance, stopping the experiment only when reaching a nominally significant result. Kosher, or not? Suppose somebody reports data showing a statistically significant effect, say at the 0.05 level. Would it matter to your appraisal of the evidence if you found out that each time they failed to find significance, they went on to collect more data, until finally they did? A rule for when to stop sampling is called a *stopping rule*.

The question is generally put by considering a random sample $\boldsymbol{X}$ that is Normally distributed with mean μ and standard deviation $\sigma = 1$, and we are testing the hypotheses:

H_0: $\mu = 0$ against H_1: $\mu \neq 0$.

This is a two-sided test: a discrepancy in either direction is sought. (The details of testing are in Excursions 3 and thereafter.) To ensure a significance level of 0.05, H_0 is rejected whenever the sample mean differs from 0 by more than $1.96\sigma/\sqrt{n}$, and, since $\sigma = 1$, the rule is: Declare $\boldsymbol{x}$ is statistically significant at the 0.05 level whenever $|\overline{X}| > 1.96/\sqrt{n}$. However, instead of fixing the sample size in advance, n is determined by the optional stopping rule:

Optional stopping rule: keep sampling until $|\overline{X}| \geq (1.96/\sqrt{n})$.

Equivalently, since the test statistic $\mathrm{d}(\boldsymbol{X}) = (\overline{X} - 0)/\sqrt{n}$:

Keep sampling until $|\mathrm{d}(\boldsymbol{X})| \geq 1.96$.

Our question was: would it be relevant to your evaluation of the evidence if you learned she'd planned to keep running trials until reaching 1.96? Having failed to rack up a 1.96 difference after, say, 10 trials, she goes on to 20, and failing yet again, she goes to 30 and on and on until finally, say, on trial 169 she gets a 1.96 difference. Then she stops and declares the statistical significance is ~0.05.

This is an example of what's called a *proper stopping rule*: the probability it will stop in a finite number of trials is 1, regardless of the true value of μ. Thus, in one of the most seminal papers in statistical foundations, by Ward Edwards,

Harold Lindman, and Leonard (Jimmie) Savage (E, L, & S) tell us, "if an experimenter uses this procedure, then with probability 1 he will eventually reject any sharp null hypothesis, even though it be true" (1963, p. 239). Understandably, they observe, the significance tester frowns on optional stopping, or at least requires the auditing of the *P*-value to require an adjustment. Had *n* been fixed, the significance level would be 0.05, but with optional stopping it increases.

Imagine instead if an account advertised itself as ignoring stopping rules. What if an account declared:

> In general, suppose that you collect data of any kind whatsoever – not necessarily Bernoullian, nor identically distributed, nor independent of each other. . . – stopping only when the data thus far collected satisfy some criterion of a sort that is sure to be satisfied sooner or later, then the import of the sequence of *n* data actually observed will be exactly the same as it would be had you planned to take exactly *n* observations in the first place. (ibid., pp. 238–9)

I've been teasing you, because these same authors who warn that to ignore stopping rules is to guarantee rejecting the null hypothesis even if it's true are the individuals who tout the irrelevance of stopping rules in the above citation – E, L, & S. They call it the *Stopping Rule Principle*. Are they contradicting themselves?

No. It is just that what looks to be, and indeed is, cheating from the significance testing perspective is not cheating from these authors' Bayesian perspective. "[F]requentist test results actually depend not only on what $\boldsymbol{x}$ was observed, but on how the experiment was stopped" (Carlin and Louis 2008, p. 8). Yes, but shouldn't they? Take a look at Table 1.1: by the time one reaches 50 trials, the probability of attaining a nominally significant 0.05 result is not 0.05 but 0.32. The actual or overall significance level is the probability of finding a 0.05 nominally significant result at some stopping point *or other*, up to the point it stops. The actual significance level accumulates.

Well-known statistical critics from psychology, Joseph Simmons, Leif Nelson, and Uri Simonsohn, place at the top of their list of requirements the need to block flexible stopping: "Researchers often decide when to stop data collection on the basis of interim data analysis . . . many believe this practice exerts no more than a trivial influence on false-positive rates" (Simmons et al. 2011, p. 1361). "Contradicting this intuition" they show the probability of erroneous rejections balloons. "A researcher who starts with 10 observations per condition and then tests for significance after every new . . . observation finds a significant effect 22% of the time" erroneously (ibid., p. 1362). Yet the followers of the Stopping Rule Principle deny it makes a difference to evidence. On their account, it *doesn't*. It's easy to see why there's disagreement.

Table 1.1. The effect of repeated significance tests (the "try and try again" method)

Number of trials n	Probability of rejecting H_0 with a result nominally significant at the 0.05 level at or before n trials, given H_0 is true
1	0.05
2	0.083
10	0.193
20	0.238
30	0.280
40	0.303
50	0.320
60	0.334
80	0.357
100	0.375
200	0.425
500	0.487
750	0.512
1000	0.531
Infinity	1.000

The Likelihood Principle

By what magic can such considerations disappear? One way to see the vanishing act is to hold, with Royall, that "what the data have to say" is encompassed in likelihood ratios. This is the gist of a very important principle of evidence, the *Likelihood Principle* (LP). Bayesian inference requires likelihoods plus prior probabilities in hypotheses; but the LP has long been regarded as a crucial part of their foundation: to violate it is to be *incoherent* Bayesianly. Disagreement about the LP is a pivot point around which much philosophical debate between frequentists and Bayesians has turned. Here is a statement of the LP:

> According to Bayes's Theorem, $\Pr(\boldsymbol{x}|\mu)$ ... constitutes the entire evidence of the experiment, that is it tells all that the experiment has to tell. More fully and more precisely, if $\boldsymbol{y}$ is the datum of some other experiment, and if it happens that $\Pr(\boldsymbol{x}|\mu)$ and $\Pr(\boldsymbol{y}|\mu)$ are proportional functions of μ (that is constant multiples of each other), then each of the two data $\boldsymbol{x}$ and $\boldsymbol{y}$ have exactly the same thing to say about the value of μ ... (Savage 1962, p. 17; replace λ with μ)

Some go further and claim that if $\boldsymbol{x}$ and $\boldsymbol{y}$ give the same likelihood, "they should give the same inference, analysis, conclusion, decision, action or anything else" (Pratt et al. 1995, p. 542). Does the LP entail the LL? No. Bayesians, for

example, generally hold to the LP, but would insist on priors that go beyond the LL. Even the converse may be denied (according to Hacking) but this is not of concern to us.

Weak Repeated Sampling Principle. For sampling theorists (my error statisticians), by contrast, this example "taken in the context of examining consistency with $\theta = 0$, is enough to refute the strong likelihood principle" (Cox 1978, p. 54), since, with probability 1, it will stop with a "nominally" significant result even though $\theta = 0$. It contradicts what Cox and Hinkley call "the weak repeated sampling principle" (Cox and Hinkley 1974, p. 51). "[W]e should not follow procedures which for some possible parameter values would give, in hypothetical repetitions, misleading conclusions most of the time" (ibid., pp. 45–6).

For Cox and Hinkley, to report a 1.96 standard deviation difference from optional stopping just the same as if the sample size had been fixed, is to discard relevant information for inferring inconsistency with the null, while "according to any approach that is in accord with the strong likelihood principle, the fact that this particular stopping rule has been used is irrelevant" (ibid., p. 51). What they call the "strong" likelihood principle will just be called the LP here. (A weaker form boils down to sufficiency, see Excursion 3.)

Exhibit (ii): How Stopping Rules Drop Out. Our question remains: by what magic can such considerations disappear? Formally, the answer is straightforward. Consider two versions of the above experiment: In the first, 1.96 is reached via fixed sample size (n = 169); in the second, by means of optional stopping that ended at 169. While $\mathrm{d}(\boldsymbol{x}) = \mathrm{d}(\boldsymbol{y})$, because of the stopping rule, the likelihood of $\boldsymbol{y}$ differs from that of $\boldsymbol{x}$ by a constant k, that is,

$$\Pr(\boldsymbol{x}|H_i) = k\Pr(\boldsymbol{y}|H_i) \text{ for constant } k.$$

Given that likelihoods enter as ratios, such proportional likelihoods are often said to be the "same." Now suppose inference is by Bayes' Theorem. Since likelihoods enter as ratios, the constant k drops out. This is easily shown. I follow E, L, & S; p. 237.

For simplicity, suppose the possible hypotheses are exhausted by two, H_0 and H_1, neither with probability of 0.

To show $\Pr(H_0|\boldsymbol{y}) = \Pr(H_0|\boldsymbol{x})$:

(1) We are given the proportionality of likelihoods, for an arbitrary value of k:

$$\Pr(\boldsymbol{y}|H_0) = k\Pr(\boldsymbol{x}|H_0),$$

$$\Pr(\boldsymbol{y}|H_1) = k\Pr(\boldsymbol{x}|H_1).$$

(2) By definition:

$$\Pr(H_0|\boldsymbol{y}) = \frac{\Pr(\boldsymbol{y}|H_0)\Pr(H_0)}{\Pr(\boldsymbol{y})}.$$

The denominator $\Pr(\boldsymbol{y}) = \Pr(\boldsymbol{y}|H_0)\ \Pr(H_0) + \Pr(\boldsymbol{y}|H_1)\ \Pr(H_1)$.

Now substitute for each term in (2) the proportionality claims in (1). That is, replace $\Pr(\boldsymbol{y}|H_0)$ with $k\Pr(\boldsymbol{x}|H_0)$ and $\Pr(\boldsymbol{y}|H_1)$ with $k\Pr(\boldsymbol{x}|H_1)$.

(3) The result is

$$\Pr(H_0|\boldsymbol{y}) = \frac{k\Pr(\boldsymbol{x}|H_0)\ \Pr(H_0)}{k\Pr(\boldsymbol{x})} = \Pr(H_0|\boldsymbol{x}).$$

The posterior probabilities are the same whether the 1.96 result emerged from optional stopping, $\boldsymbol{Y}$, or fixed sample size, $\boldsymbol{X}$.

This essentially derives the LP from inference by Bayes' Theorem, and shows the equivalence for the particular case of interest, optional stopping. As always, when showing a Bayesian computation I use the conditional probability "|" rather than the ";" of the frequentist.[3]

The 1959 Savage Forum: What Counts as Cheating?

My colleague, well-known Bayesian I. J. Good, would state it as a "paradox":

> [I]f a Fisherian is prepared to use optional stopping (which usually he is not) he can be sure of rejecting a true null hypothesis provided that he is prepared to go on sampling for a long time. The way I usually express this 'paradox' is that a Fisherian [but not a Bayesian] can cheat by pretending he has a plane to catch like a gambler who leaves the table when he is ahead. (Good 1983, p. 135)

The lesson about who is allowed to cheat depends on your statistical philosophy. Error statisticians require that the overall and not the "computed" significance level be reported. To them, cheating would be to report the significance level you got after trying and trying again in just *the same way* as if the test had a fixed sample size (Mayo 1996, p. 351). Viewing statistical methods as tools for severe tests, rather than as probabilistic logics of evidence, makes a deep difference to the tools we seek. Already we find ourselves thrust into some of the knottiest and most intriguing foundational issues.

This is Jimmie Savage's message at a 1959 forum deemed sufficiently important to occupy a large gallery of the Museum of Statistics (hereafter "The Savage Forum" (Savage 1962)). Attendees include Armitage, Barnard,

[3] $\Pr(\boldsymbol{x}) = \Pr(\boldsymbol{x}\ \&\ H_0) + \Pr(\boldsymbol{x}\ \&\ H_1)$, where H_0 and H_1 are exhaustive.

Bartlett, Cox, Good, Jenkins, Lindley, Pearson, Rubin, and Smith. Savage announces to this eminent group of statisticians that if adjustments in significance levels are required for optional stopping, which they are, then the fault must be with significance levels. Not all agreed. Needling Savage on this issue, was Peter Armitage:

> I feel that if a man deliberately stopped an investigation when he had departed sufficiently far from his particular hypothesis, then 'Thou shalt be misled if thou dost not know that.' If so, prior probability methods seem to appear in a less attractive light than frequency methods where one can take into account the method of sampling. (Armitage 1962, p. 72)

Armitage, an expert in sequential trials in medicine, is fully in favor of them, but he thinks stopping rules should be reflected in overall inferences. He goes further:

> [Savage] remarked that, using conventional significance tests, if you go on long enough you can be sure of achieving any level of significance; does not the same sort of result happen with Bayesian methods? (ibid., p. 72)

He has in mind using a type of uniform prior probability for μ, wherein the posterior for the null hypothesis matches the significance level. (We return to this in Excursion 6. For $\sigma = 1$, its distribution is Normal($\bar{x}$, $1/n$).)

Not all cases of trying and trying again injure error probabilities. Think of trying and trying again until you find a key that fits a lock. When you stop, there's no probability of being wrong. (We return to this in Excursion 4.)

Savage's Sleight of Hand

Responding to Armitage, Savage engages in a bit of sleight of hand. Moving from the problematic example to one of two predesignated point hypotheses, H_0: $\mu = \mu_0$, and H_1: $\mu = \mu_1$, he shows that the error probabilities are controlled in that case. In particular, the probability of obtaining a result that makes H_1 r times more likely than H_0 is less than $1/r$: $\Pr(\text{LR} > r; H_0) < 1/r$. But, that wasn't Armitage's example; nor does Savage return to it. Now, it is open to Likelihoodists to resist being saddled "with ideas that are alien to them" (Sober 2008, p. 77). Since the Likelihoodist keeps to this type of comparative appraisal, they can set bounds to the probabilities of error. However, the bounds are no longer impressively small as we add hypotheses, even if they are predesignated[4] (Mayo and Kruse 2001).

[4] A general result, stated in Kerridge (1963, p. 1109), is that with k simple hypotheses, where H_0 is true and $H_1, \ldots, H_{k-1}$ are false, and equal priors, *"the frequency with which, at the termination of sampling the posterior probability of the true hypothesis is* p *or less cannot exceed* $(k-1)p/(1-p)$." Such bounds depend on having countably additive probability, while the uniform prior in Armitage's example imposes finite additivity.

Something more revealing is going on when the Likelihoodist sets pre-data bounds. Why the sudden concern with showing the rule for comparative evidence would very improbably find evidence in favor of the wrong hypothesis? This is an error probability. So it appears they also care about error probabilities – at least before-trial – or they are noting, for those of us who do, that they also have error control in the simple case of predesignated point hypotheses. The severe tester asks: If you want to retain these pre-data safeguards, why allow them to be spoiled by data-dependent hypotheses and stopping rules?

Some have said: the evidence is the same, but you take into account things like stopping rules and data-dependent selections *afterwards*. When making an inference, this *is* afterwards, and we need an epistemological rationale to pick up on their influences *now*. Perhaps knowing someone uses optional stopping warrants a high belief he's trying to deceive you, leading to a high enough prior belief in the null. Maybe so, but this is to let priors reflect methods in a non-standard way. Besides, Savage (1961, p. 583) claimed optional stopping "is no sin," so why should it impute deception? So far as I know, subjective Bayesians have resisted the idea that rules for stopping alter the prior. Couldn't you pack the concern in some background *B*? You could, but you would need another account to justify doing so, thereby only pushing back the issue. I've discussed an assortment of attempts elsewhere: Mayo (1996), Mayo and Kruse (2001), Mayo (2014b). Others have too, discussed here and elsewhere; please see our online sources (preface).

Arguments from Intentions: All in Your Head?

A funny thing happened at the Savage Forum: George Barnard announces he no longer holds the LP for the two-sided test under discussion, only for the predesignated point alternatives. Savage is shocked to hear it:

> I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resist an idea so patently right. (Savage 1962, p. 76)

The argument Barnard gave him was that the plan for when to stop was a matter of the researchers' intentions, all wrapped up in their heads. While Savage denies he was ever sold on the argument from intentions, it's a main complaint you will hear about taking account, not just of stopping rules, but of error probabilities in general. Take the subjective Bayesian philosophers Howson and Urbach (1993):

A significance test inference, therefore, depends not only on the outcome that a trial produced, but also on the outcomes that it could have produced but did not. And the latter are determined by certain private intentions of the experimenters, embodying their stopping rule. It seems to us that this fact precludes a significance test delivering any kind of judgment about empirical support. (p. 212)

The truth is, whether they're hidden or not turns on your methodology being able to pick up on them. So the deeper question is: *ought* your account pick up on them?

The answer isn't a matter of mathematics, it depends on your goals and perspective – yes on your philosophy of statistics. Ask yourself: What features lead you to worry about cherry picking, and selective reporting? Why do the CONSORT and myriad other best practice manuals care? Looking just at the data and hypotheses – as a "logic" of evidence would – you will not see the machinations. Nevertheless, these machinations influence the capabilities of the tools. Much of the handwringing about irreproducibility is the result of wearing blinders as to the construction and selection of both hypotheses and data. In one sense, all test specifications are determined by a researcher's intentions; that doesn't make them private or invisible to us. They're visible to accounts with antennae to pick up on them!

You might try to deflect the criticism of stopping rules by pointing out that some stopping rules do alter priors. Armitage wasn't ignoring that, nor are we. These are called informative stopping rules, and examples are rather contrived. For instance, "a man who wanted to know how frequently lions watered at a certain pool was chased away by lions" (E, L, & S 1963, p. 239). They add, "we would not give a facetious example had we been able to think of a serious one." In any event, this is irrelevant for the Armitage example, which is non-informative.

Error Probabilities Violate the LP

> [I]t seems very strange that a frequentist could not analyze a given set of data, such as $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ if the stopping rule is not given . . . [D]ata should be able to speak for itself. (Berger and Wolpert 1988, p. 78)

Inference by Bayes' Theorem satisfies this intuition, which sounds appealing; but for our severe tester, data no more speak for themselves in the case of stopping rules than with cherry picking, hunting for significance, and the like. We may grant to the Bayesian that

> [The] irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design that had been lost by classical emphasis on significance levels (in the sense of Neyman and Pearson). (E, L, & S 1963, p. 239)

The question is whether this latitude is desirable. If you are keen to use statistical methods critically, as our severe tester, you'll be suspicious of a simplicity and freedom to mislead.

Admittedly, this should have been more clearly spelled out by Neyman and Pearson. They rightly note:

> In order to fix a limit between 'small' and 'large' values of [the likelihood ratio] we must know how often such values appear when we deal with a true hypothesis. (Pearson and Neyman 1930, p. 106)

That's true, but putting it in terms of the desire "to control the error involved in rejecting a true hypothesis" it is easy to dismiss it as an affliction of a frequentist concerned only with long-run performance. Bayesians and Likelihoodists are free of this affliction. Pearson and Neyman should have said: ignoring the information as to how readily true hypotheses are rejected, we cannot determine if there really is evidence of inconsistency with them.

Our minimal requirement for evidence insists that data only provide genuine or reliable evidence for H if H survives a severe test – a test H would probably have failed if false. Here the hypothesis H of interest is the non-null of Armitage's example: the existence of a genuine effect. A warranted inference to H depends on the test's ability to find H false when it is, i.e., when the null hypothesis is true. The severity conception of tests provides the link between a test's error probabilities and what's required for a warranted inference.

The error probability computations in significance levels, confidence levels, power, all depend on violating the LP! Aside from a concern with "intentions," you will find two other terms used in describing the use of error probabilities: a concern with (i) outcomes other than the one observed, or (ii) the sample space. Recall Souvenir B, where Royall, who obeys the LP, speaks of "the irrelevance of the sample space" once the data are in hand. It's not so obvious what's meant. To explain, consider Jay Kadane: "Significance testing violates the Likelihood Principle, which states that, having observed the data, inference must rely only on what happened, and not on what might have happened but did not" (Kadane 2011, p. 439). According to Kadane, the probability statement: $\Pr(|d(\boldsymbol{X})| > 1.96) = 0.05$ "is a statement about $d(\boldsymbol{X})$ before it is observed. After it is observed, the event $\{d(\boldsymbol{X}) > 1.96\}$ either

happened or did not happen and hence has probability either one or zero" (ibid.).

Knowing d($\boldsymbol{x}$) = 1.96, Kadane is saying there's no more uncertainty about it. But would he really give it probability 1? That's generally thought to invite the problem of "known (or old) evidence" made famous by Clark Glymour (1980). If the probability of the data $\boldsymbol{x}$ is 1, Glymour argues, then Pr($\boldsymbol{x}|H$) also is 1, but then Pr($H|\boldsymbol{x}$) = Pr(H)Pr($\boldsymbol{x}|H$)/Pr($\boldsymbol{x}$) = Pr(H), so there is no boost in probability given $\boldsymbol{x}$. So does that mean known data don't supply evidence? Surely not. Subjective Bayesians try different solutions: either they abstract to a context prior to knowing $\boldsymbol{x}$, or view the known data as an instance of a general type, in relation to a sample space of outcomes. Put this to one side for now in order to continue the discussion.[5]

Kadane is emphasizing that Bayesian inference is *conditional* on the particular outcome. So once $\boldsymbol{x}$ is known and fixed, other possible outcomes that could have occurred but didn't are irrelevant. Recall finding that Pickrite's procedure was to build k different portfolios and report just the one that did best. It's as if Kadane is asking: "Why are you considering other portfolios that you might have been sent but were not, to reason from the one that you got?" Your answer is: "Because that's how I figure out whether your boast about Pickrite is warranted." With the "search through k portfolios" procedure, the possible outcomes are the success rates of the k different attempted portfolios, each with its own null hypothesis. The actual or "audited" P-value is rather high, so the severity for H: Pickrite has a reliable strategy, is low ($1 - p$). For the holder of the LP to say that, once $\boldsymbol{x}$ is known, we're not allowed to consider the other chances they gave themselves to find an impressive portfolio, is to put the kibosh on a crucial way to scrutinize the testing process.

Interestingly, nowadays, non-subjective or default Bayesians concede they "have to live with some violations of the likelihood and stopping rule principles" (Ghosh, Delampady, and Samanta 2010, p. 148) since their prior probability distributions are influenced by the sampling distribution. Is it because ignoring stopping rules can wreak havoc with the well-testedness of inferences? If that is their aim, too, then that is very welcome. Stay tuned.

[5] Colin Howson, a long-time subjective Bayesian, has recently switched to being a non-subjective Bayesian at least in part because of the known evidence problem (Howson 2017, p. 670).

Souvenir C: A Severe Tester's Translation Guide

Just as in ordinary museum shops, our souvenir literature often probes treasures that you didn't get to visit at all. Here's an example of that, and you'll need it going forward. There's a confusion about what's being done when the significance tester considers the set of all of the outcomes leading to a d($\boldsymbol{x}$) greater than or equal to 1.96, i.e., {$\boldsymbol{x}$: d($\boldsymbol{x}$) ≥ 1.96}, or just d($\boldsymbol{x}$) ≥ 1.96. This is generally viewed as throwing away the particular $\boldsymbol{x}$, and lumping all these outcomes together. What's really happening, according to the severe tester, is quite different. What's actually being signified is that we are interested in the method, not just the particular outcome. Those who embrace the LP make it very plain that data-dependent selections and stopping rules drop out. To get them to drop in, we signal an interest in what the test procedure *would have* yielded. This is a counterfactual and is altogether essential in expressing the properties of the method, in particular, the probability it would have yielded some nominally significant outcome *or other*.

When you see Pr(d($\boldsymbol{X}$) ≥ d($\boldsymbol{x_0}$); H_0), or Pr(d($\boldsymbol{X}$) ≥ d($\boldsymbol{x_0}$); H_1), for any particular alternative of interest, insert:

"the test procedure would have yielded"

just before the d($\boldsymbol{X}$). In other words, this expression, with its inequality, is a signal of interest in, and an abbreviation for, the error probabilities associated with a test.

Applying the Severity Translation. In Exhibit (i), Royall described a significance test with a Bernoulli(θ) model, testing H_0: $\theta \leq 0.2$ vs. H_1: $\theta > 0.2$. We blocked an inference from observed difference d($\boldsymbol{x}$) = 3.3 to $\theta = 0.8$ as follows. (Recall that $\overline{x} = 0.53$ and d($\boldsymbol{x_0}$) $\simeq$ 3.3.)

We computed Pr(d($\boldsymbol{X}$) > 3.3; $\theta = 0.8$) $\simeq$1.

We translate it as Pr(The test would yield d($\boldsymbol{X}$) > 3.3; $\theta = 0.8$) $\simeq$1.

We then reason as follows:

Statistical inference: If $\theta = 0.8$, then the method would virtually always give a difference larger than what we observed. Therefore, the data indicate $\theta < 0.8$.

(This follows for rejecting H_0 in general.) When we ask: "How often would your test have found such a significant effect even if H_0 is approximately true?" we are asking about the properties of the experiment that *did* happen.

The counterfactual "would have" refers to how the procedure would behave in general, not just with these data, but with other possible data sets in the sample space.

Exhibit (iii). Analogous situations to the optional stopping example occur even without optional stopping, as with selecting a data-dependent, maximally likely, alternative. Here's an example from Cox and Hinkley (1974, 2.4.1, pp. 51–2), attributed to Allan Birnbaum (1969).

A single observation is made on $\boldsymbol{X}$, which can take values 1,2, . . .,100. "There are 101 possible distributions conveniently indexed by a parameter θ taking values 0, 1,..., 100" (ibid.). We are not told what θ is, but there are 101 possible point hypotheses about the value of θ: from 0 to 100. If $\boldsymbol{X}$ is observed to be $\boldsymbol{r}$, written $\boldsymbol{X} = \boldsymbol{r}$ ($\boldsymbol{r} \neq 0$), then the most likely hypothesis is $\theta = \boldsymbol{r}$: in fact, $\Pr(\boldsymbol{X} = \boldsymbol{r}; \theta = \boldsymbol{r}) = 1$. By contrast, $\Pr(\boldsymbol{X} = \boldsymbol{r}; \theta = 0) = 0.01$. Whatever value $\boldsymbol{r}$ that is observed, hypothesis $\theta = \boldsymbol{r}$ is 100 times as likely as is $\theta = 0$. Say you observe $\boldsymbol{X} = 50$, then H: $\theta = 50$ is 100 times as likely as is $\theta = 0$. So "even if in fact $\theta = 0$, we are certain to find evidence apparently pointing strongly against $\theta = 0$, if we allow comparisons of likelihoods chosen in the light of the data" (Cox and Hinkley 1974, p. 52). This does not happen if the test is restricted to two preselected values. In fact, if $\theta = 0$ the probability of a ratio of 100 in favor of the false hypothesis is 0.01.[6]

Allan Birnbaum gets the prize for inventing chestnuts that deeply challenge both those who do, and those who do not, hold the Likelihood Principle!

Souvenir D: Why We Are So New

What's Old? You will hear critics say that the reason to overturn frequentist, sampling theory methods – all of which fall under our error statistical umbrella – is that, well, they've been around a long, long time. First, they are scarcely stuck in a time warp. They have developed with, and have often been the source of, the latest in modeling, resampling, simulation, Big Data, and machine learning techniques. Second, all the methods have roots in long-ago ideas. Do you know what is really up-to-the-minute in this time of massive, computer algorithmic methods and "trust me" science? A new vigilance about retaining hard-won error control techniques. Some thought that, with enough data, experimental design

[6] From Cox and Hinkley 1974, p. 51. The likelihood function corresponds to the normal distribution of $\overline{X}$ around μ with SE $\sigma/\sqrt{n}$. The likelihood at $\mu = 0$ is $\exp(-0.5k^2)$ times that at $\mu = \bar{x}$. One can choose k to make the ratio small. "That is, even if in fact $\mu = 0$, there always appears to be strong evidence against $\mu = 0$, at least if we allow comparison of the likelihood at $\mu = 0$ against any value of μ and hence in particular against the value of μ giving maximum likelihood". However, if we confine ourselves to comparing the likelihood at $\mu = 0$ with that at some fixed $\mu = \mu'$, this difficulty does not arise.

could be ignored, so we have a decade of wasted microarray experiments. To view outcomes other than what you observed as irrelevant to what $\boldsymbol{x}_0$ says is also at odds with cures for irreproducible results. When it comes to cutting-edge fraud-busting, the ancient techniques (e.g., of Fisher) are called in, refurbished with simulation.

What's really old and past its prime is the idea of a logic of inductive inference. Yet core discussions of statistical foundations today revolve around a small cluster of (very old) arguments based on that vision. Tour II took us to the crux of those arguments. Logics of induction focus on the relationships between given data and hypotheses – so outcomes other than the one observed drop out. This is captured in the Likelihood Principle (LP). According to the LP, trying and trying again makes no difference to the probabilist: it is what someone intended to do, locked up in their heads.

> It is interesting that frequentist analyses often need to be adjusted to account for these 'looks at the data,'. . . That Bayesian analysis claims no need to adjust for this 'look elsewhere' effect – called the *stopping rule principle* – has long been a controversial and difficult issue. . . (J. Berger 2008, p. 15)

The irrelevance of optional stopping is an asset for holders of the LP. For the task of criticizing and debunking, this puts us in a straightjacket. The warring sides talk past each other. We need a new perspective on the role of probability in statistical inference that will illuminate, and let us get beyond, this battle.

New Role of Probability for Assessing What's Learned. A passage to locate our approach within current thinking is from Reid and Cox (2015):

> Statistical theory continues to focus on the interplay between the roles of probability as representing physical haphazard variability . . . and as encapsulating in some way, directly or indirectly, aspects of the uncertainty of knowledge, often referred to as epistemic. (p. 294)

> We may avoid the need for a different version of probability by appeal to a notion of calibration, as measured by the behavior of a procedure under hypothetical repetition. That is, we study assessing uncertainty, as with other measuring devices, by assessing the performance of proposed methods under hypothetical repetition. Within this scheme of repetition, probability is defined as a hypothetical frequency. (p. 295)

This is an ingenious idea. Our meta-level appraisal of methods proceeds this way too, but with one important difference. A key question for us is the proper epistemic role for probability. It is standardly taken as providing a probabilism, as an assignment of degree of actual or rational belief in a claim, absolute or comparative. We reject this. We proffer an alternative theory: a severity assessment. An account of what is warranted and unwarranted to infer – a normative epistemology – is not a matter of using probability to assign rational beliefs, but to control and assess how well probed claims are.

If we keep the presumption that the epistemic role of probability is a degree of belief of some sort, then we can "avoid the need for a different version of probability" by supposing that good/poor performance of a method warrants high/low belief in the method's output. Clearly, poor performance is a problem, but I say a more nuanced construal is called for. The idea that partial or imperfect knowledge is all about degrees of belief is handed down by philosophers. Let's be philosophical enough to challenge it.

New Name? An error statistician assesses inference by means of the error probabilities of the method by which the inference is reached. As these stem from the sampling distribution, the conglomeration of such methods is often called "sampling theory." However, sampling theory, like classical statistics, Fisherian, Neyman–Pearsonian, or frequentism are too much associated with hardline or mish-mashed views. Our job is to clarify them, but in a new way. Where it's apt for taking up discussions, we'll use "frequentist" interchangeably with "error statistician." However, frequentist error statisticians tend to embrace the long-run performance role of probability that I find too restrictive for science. In an attempt to remedy this, Birnbaum put forward the "confidence concept" (Conf), which he called the "one rock in a shifting scene" in statistical thinking and practice. This "one rock," he says, takes from the Neyman–Pearson (N-P) approach "techniques for systematically appraising and bounding the probabilities (under respective hypotheses) of seriously misleading interpretations of data" (Birnbaum 1970, p.1033). Extending his notion to a composite alternative:

> Conf: An adequate concept of statistical evidence should find strong evidence against H_0 (for $\sim H_0$) with small probability α when H_0 is true, and with much larger probability $(1 - \beta)$ when H_0 is false, increasing as discrepancies from H_0 increase.

This is an entirely right-headed pre-data performance requirement, but I agree with Birnbaum that it requires a reinterpretation for evidence post-data (Birnbaum 1977). Despite hints and examples, no such evidential interpretation has been given. The switch that I'm hinting at as to what's required for an evidential or epistemological assessment is key. Whether one uses a frequentist or a propensity interpretation of error probabilities (as Birnbaum did) is not essential. *What we want is an error statistical approach that controls and assesses a test's stringency or severity*. That's not much of a label. For short, we call someone who embraces such an approach a severe tester. For now I will just venture that a severity scrutiny illuminates all statistical approaches currently on offer.